Atti del XV Convegno Annuale **AIUCD**

Digitale e *Public Engagement*

Pratiche e prospettive nelle Digital Humanities

Cagliari 3-4-5 giugno 2026

a cura di

Cristina **Marras** | Andrea **Pergola** | Giampaolo **Salice**

UNICA
UNIVERSITÀ DEGLI STUDI DI CAGLIARI

DH.UNICA.IT
Centro interdipartimentale per l'Umanistica Digitale

ASSOCIAZIONE per l'INFORMATICA UMANISTICA e la CULTURA DIGITALE

ISBN 9791298618817

Copyright ©2026 AIUCD
Associazione per l'Informatica Umanistica e la Cultura Digitale

Il presente volume e tutti i contributi sono rilasciati sotto licenza Creative Commons Attribution ShareAlike 4.0 International license (CC-BY-SA 4.0). Ogni altro diritto rimane in capo ai singoli autori.

This volume and all contributions are released under the Creative Commons Attribution Share-Alike 4.0 International license (CC-BY-SA 4.0). All other rights retained by the legal owners.

A cura di: Cristina Marras; Andrea Pergola; Giampaolo Salice (2026). *Digitale e* Public Engagement*: pratiche e prospettive nelle* Digital Humanities, Atti del XV Convegno Annuale AIUCD, Cagliari 3-5 giugno 2026, Università degli Studi di Cagliari.

Ultimo accesso agli URL in data 13 maggio 2026.
Last URL access May, 13 2026.

Si prega di notificare all'editore ogni omissione o errore si riscontri: segreteria [at] aiucd.org
Please notify the publisher of any omissions or errors found: segreteria [at] aiucd.org

Il programma del Convegno AIUCD 2026 è disponibile online:
https://www.aiucd2026.unica.it/companion/
The AIUCD 2026 Conference Program is available online:
https://www.aiucd2026.unica.it/companion/

I contributi pubblicati nel presente volume hanno ottenuto il parere favorevole da parte di valutatori esperti della materia, attraverso un processo di revisione anonima mediante *double-blind peer review*, effettuata dai membri del Comitato di Programma sotto la supervisione del Comitato Scientifico di AIUCD 2026.
All the papers published in this volume have received favourable reviews by experts in the field of DH, through an anonymous double-blind peer review, carried out by the members of the Programme Committee under the supervision of the Scientific Committee of AIUCD 2026.

Gli Atti del Convegno AIUCD 2026 sono pubblicati come raccolta dei contributi forniti direttamente dagli autori e dalle autrici. I file sono stati raccolti e assemblati senza interventi redazionali significativi da parte dei curatori.

The proceedings of the AIUCD 2026 Conference are published as a collection of contributions provided directly by the authors. The files have been collected and compiled without significant editorial intervention by the editors.

Il logo di AIUCD 2026 è opera di Raffaele Argiolas
The AIUCD 2026 logo was designed by Raffaele Argiolas
La copertina è stata realizzata da Giampaolo Salice
The cover was created by Giampaolo Salice

GENERAL CHAIRS

Giampaolo Salice | Università degli Studi di Cagliari | DH UNICA
Cristina Marras | CNR-ILIESI | AIUCD

COMITATO ORGANIZZATORE / ORGANIZING COMMITTEE

Il Comitato Organizzatore è composto da Giampaolo Salice (Università degli Studi di Cagliari), Cristina Marras (CNR-ILIESI), Raffaele Argiolas, Alessandro Capra, Giommaria Carboni, Andrea Pergola, Eleonora Todde (DH UNICA Università degli Studi di Cagliari), Christian D'Agata (Università di Catania), Francesca Frontini (CNR-ILC), Simone Rebora (Università di Verona), Marco Rospocher (Università di Verona), Laura Stochino (ISSASCO), Matteo Tatti e Alice Nozza (Associazione Itzokor).

***Iscrizioni** / Conference Registration*
Eleonora **Todde** – DH UniCa – Università degli Studi di Cagliari

Atti / Conference Proceedings
Andrea **Pergola** – DH UniCa – Università degli Studi di Cagliari

***Comunicazione** / Communication*
Raffaele **Argiolas** – DH UniCa – Università degli Studi di Cagliari

***Promozione** / Promotion*
Giommaria **Carboni** – DH UniCa – Università degli Studi di Cagliari

***Tecnologo dell'Evento** / Event Technologist*
Alessandro **Capra** – DH UniCa

Organizzazione locale

Dipartimento di Lettere, Lingue e Beni Culturali, Università degli Studi di Cagliari, Ufficio Amministrativo: Serena Serra, Giulia Cadoni, Claudia Serri

COMITATO SCIENTIFICO / SCIENTIFIC COMMITTEE

Il Comitato Scientifico del convegno AIUCD 2026 è composto dai Chairs di Programma e dai Chairs delle cinque aree tematiche.

Program Chairs

Giampaolo Salice | Università degli Studi di Cagliari | DH UNICA
Cristina Marras | CNR-ILIESI | AIUCD

DH e co-costruzione del sapere con le comunità: sfide, metodi, strumenti / *DH and co-construction of knowledge with communities: challenges, methods, tools*

Arianna Ciula | King's Digital Lab, London
Greta Franzini | EURAC, Bolzano

Archivi ed edizioni: descrizioni aumentate, accessibilità e sistemi informativi / *Archives and editions: augmented descriptions, accessibility, and information systems*

Federico Valacchi | Università di Macerata
Paolo Monella | Università Kore Enna | AIUCD

Testualità digitali: prospettive, sviluppi e sperimentazioni / *Digital textualities: perspectives, developments, and experimentations*

Simone Rebora | Università di Verona | AIUCD
Alessandro Adamou | Biblioteca Hertziana

Rappresentazione di Dati e Conoscenza / *Data and Knowledge Representation*

Francesca Tomasi | Alma Mater Studiorum – Università di Bologna
Angelo Mario Del Grosso | CNR-ILC

Memorie, storia e patrimoni culturali digitali / *Tangible and Intangible Digital Heritage*

Michela Tardella | CNR-ILIESI
Enrica Salvatori | Università di Pisa

COMITATO DI PROGRAMMA / PROGRAM COMMITTEE

Stefano Allegrezza (Università di Macerata), Laura Antonietti (Université de Versailles Saint Quentin en Yvelines - Université Paris-Saclay), Alessio Antonini (Open University), Liborio P. Barbarino (Università di Catania), Nicola Barbuti (Università di Bari "Aldo Moro"), Stefano Bazzaco (Università di Verona), Andrea Bellandi (CNR-ILC), Giulia Benotto (CNR-ILC), Monica Berti (Leipzig University), Mario A. Bochicchio (Università di Bari "Aldo Moro"), Andrea Bolioli (Ricercatore indipendente), Marco Bombieri (Università di Verona), Flavia Bruni (Università di Chieti-Pescara), Marina Buzzoni (Università Ca' Foscari di Venezia), Alberto Campagnolo (KU Leuven), Vittore Casarosa (CNR-ILC), Raffaele Cioffi (Università di Napoli Federico II), Vincenzo Colaprice (Università di Torino), Francesca Congiu (Università degli Studi di Cagliari), Giuseppe Consolo (Università degli studi di Napoli, Federico II), Christian D'Agata (Università di Catania), Giulia D'Agostino (TU Darmstadt), Elisa D'Argenio (HUN-REN Hungarian Research Centre for Linguistics), Davide Dainese (Alma Mater Studiorum Università di Bologna), Stefano Dall'Aglio (Università Ca' Foscari Venezia), Marilena Daquino (Alma Mater Studiorum Università di Bologna), Mauro De Bari (Università di Bari "Aldo Moro"), Angelo M. Del Grosso (CNR-ILC), Matteo Di Cristofaro (UniMore), Francesca Di Donato (CNR-ILC), Giorgia Di Marcantonio (Università di Macerata), Giorgio Maria Di Nunzio (Università di Padova), Roberto Evangelista (CNR-ISPF), Stefano Ferilli (Università di Bari "Aldo Moro"), Franz Fischer (Università Ca' Foscari di Venezia), Greta H. Franzini (EURAC), Francesca Frontini (CNR-ILC), Daniele Fusi (Università Ca' Foscari di Venezia), Mariangela Giglio (Alma Mater Studiorum Università di Bologna), Tiago Luis Gil (University of Brasilia), Michela Giordano (Università degli Studi di Cagliari), Luca Giovannini (University of Potsdam), Milena Giuffrida (Università di Catania), Giovanna Granata (Università degli Studi di Cagliari), Edmondo Grassi (Università Telematica San Raffaele Roma), Miryam Grasso (Università di Catania), Piergiovanna Grossi (Università di Verona), Marina M. Guglielmi (Università degli Studi di Cagliari), Alessandro Iannella (Università degli Studi di Milano), Sabrina Iorio (Università Napoli Federico II), Michele Lacriola (Università di Siena), Alessandro Laruffa (ISSTOR), Anna Mambelli (UniMore), Francesco Mambrini (Università Cattolica del Sacro Cuore), Francesco Mameli (Università degli Studi di Cagliari), Lorenzo Mancini (CNR-ILIESI), Anna Maria Marras (Università di Torino), Cristina Marras (CNR-ILIESI), Pietro Mazzarisi (Università di Trieste), Barbara McGillivray (King's College London), Federico V. Meschini (Università della Tuscia), Alessio Miaschi (CNR-ILC), Giulia Miglietta (Università del Salento), Paolo Monella (Università Kore di Enna), Rossana Morriello (Università degli Studi di Firenze), Giulia Murgia (Università degli Studi di Cagliari), Enrico Natale (infoclio.ch), Sebastiana Nocco (CNR-ISEM), Serge Noiret (AIPH), Giuseppe Palazzolo (Università di Catania), Mafalda Papini (CNR-ILC), Enrico Pasini (Università di Torino), Giulia Pedonese (CNR-ILC), Paola Peratello (Università Ca' Foscari Venezia), Andrea Pergola (Università degli Studi di Cagliari), Ginevra Peruginelli (CNR-IGSG) University of Groningen, Federico Pianzola (University of Groningen), Fabio C. Pinna (Università degli Studi di Cagliari), Igor Pizzirusso (AIPH), Tiziana Pontillo (Università degli Studi di Cagliari), Valeria Quochi (CNR-ILC), Lisa Reggiani (CNR-ILIESI), Giulia Renda (Alma Mater Studiorum Università di Bologna), Pietro Restaneo (CNR-ILIESI), Dario Rodighiero (University of Groningen), Roberto Rosselli Del Turco (Università di Torino), Federica Rovelli (Università di Pavia), Giampaolo Salice (Università degli Studi di Cagliari), Enrica Salvatori (Università di Pisa), Emilio M. Sanfilippo (CNR-ISTC), Eva Sassolini (CNR-ILC), Manfredi Scanagatta (UniMore), Andrea Schimmenti (Alma Mater Studiorum Università di Bologna), Flavia Sciolette (CNR-ILC), Alessia Scognamiglio (CNR-ISPF), Luigi Serra (CNR-ISEM), Pietro Sichera (CNR-ILIESI), Daria Spampinato (CNR-ISTC), Giulia Speranza (Università di Napoli "L'Orientale"), Rachele Sprugnoli (Università di Parma), Francesco V. Stella (Università di Siena), Timothy Tambassi (Università Ca' Foscari di Venezia), Mirko Tavosanis (Università di Pisa), Francesca Tomasi (Alma Mater Studiorum Università di Bologna), Simona Turbanti (Università di Milano), Nicoletta Usai (Università degli Studi di Cagliari), Marco Venuti (Università di Catania), Gennaro Vessio (Università di Bari "Aldo Moro", Gabriele Vezzani (Università di Verona), Simone Zenzaro (CNR-ILC).

Enti organizzatori / Organisers

Il XV convegno annuale dell'Associazione per l'Informatica Umanistica e la Cultura Digitale (AIUCD) è organizzata dal DH UNICA - Centro Interdipartimentale per l'Umanistica Digitale e dal Dipartimento di Lettere, Lingue e Beni Culturali dell'Università di Cagliari in collaborazione con AIUCD.

Sommario

PREMESSA I-IV

Cristina Marras, Andrea Pergola, Giampaolo Salice

[1/A] INTELLIGENZA ARTIFICIALE GENERATIVA E RIFLESSIONE TEORICA

Una Agenda sull'AI generativa per Digital Humanities 6

Fabio Ciotti

L'errore dell'IA come euristica per una deliberazione assistita 14

Valerio De Luce; Tiberio Uricchio

Verso una memoria artificiale? Il progetto Johannes tra questioni tecnologiche, archeologiche e morali 21

Stefano Bertoldi

AI confini della realtà: il pensiero umano nell'epoca della sua riproducibilità generativa 31

Daniele Silvi

[1/B] CORPORA DIGITALI: REQUISITI E SOLUZIONI GESTIONALI

Verba volant, emails manent: una proposta di metodologia per l'acquisizione e la preservazione degli archivi di posta elettronica di personaggi illustri 38

Alberto Abis, Stefano Allegrezza, Fabrizio Stupino

Sfide alla sostenibilità dei progetti di digitalizzazione: i casi studio degli archivi manicomiali calabresi 45

Emanuela Nicole Donato, Maria Natalia Federico, Grazia Serratore

A technical solution for a digital cultural heritage corpus: Notes on collaboration, requirements and sustainability 53

Arianna Ciula, Miguel Vieira, Geoffroy Noel, Neil Jakeman, Zihao Lu, Tiffany Ong, Simona Stoyanova, Jonathan Prag, Alessia Coccato, Ryan Heuser

[1/C] CO-PRODUZIONE DI ARCHIVI E COLLEZIONI DI COMUNITÀ

L'Archivio Digitale di Vigne Nuove. Costruire un archivio di comunità, tra teoria e pratica 62

Manfredi Scanagatta

Participatory Digital Archives as a Method for Co-constructing Cultural Memory: The Bunjevke Women's Platform 68

Sandra Iršević PhD

Il ritorno dei partigiani sardi. Una collezione digitale co-prodotta 73

Walter Falgio, Laura Stochino, Matteo Quarantiello

Dalla materialità delle fonti al dato condiviso: metodi ibridi per un archivio digitale partecipativo 78

Tiziana Pasciuto

[2/A] LARGE LANGUAGE MODELS: APPLICAZIONI UMANISTICHE

Scholarly Opinion Mining using LLMs and Knowledge Graphs: the case of the Van den vos Reynaerde 87
Andrea Schimmenti, Joris van Zundert, Fabio Vitali, Marieke van Erp

Un LLM per le Scienze Umane e Sociali? ReSearch_SSH come esperimento europeo 95
Adam Faci, Alessio Miaschi, Anne Combe, Pascal Cuxac[4], Francesca Frontini, Nicolas Larrousse, Stéphane Pouyllau

L'AI a servizio della Musicologia: costruzione di thesauri musicologici con il supporto dei Large Language Models 102
Manuela Grillo, Paolo Bonora

Per una formazione del filologo computazionale del futuro: Domain Specific Language e Large Language Model 107
Christian D'Agata, Angelo Mario Del Grosso, Federico Boschetti

[2/B] OPEN SCIENCE, SOSTENIBILITÀ E PRATICHE APERTE

Are Digital Humanities really committed to open? An exploratory study on the availability of methodological workflows and open peer review practices 116
Silvio Peroni

Monitorare la transizione: l'Osservatorio sulle pratiche di Open Science nelle SSH come strumento di Public Engagement 122
Pamela Barletta, Marta Caradonna, Nicola Giampietro, Alice Orrù, Federico Silvestri

Nel 25° anniversario di Wikipedia: l'ecosistema Wikimedia e le Digital Humanities tra pratiche collaborative e sostenibilità della conoscenza 130
Piergiovanna Grossi

Una pausa e tre domande. La crisi della memoria nell'era dell'accesso 137
Enrica Salvatori, Federico Valacchi

[2/C] MODELLI DIGITALI PER GESTIRE E CONDIVIDERE FONTI MULTIDISCIPLINARI

Digital Repatriation in Practice: Sharing Ethnographic Photography with Communities of Origin at the Náprstek Museum 142
Daniela Šnapková

Una storia da ascoltare. Descrivere, mappare, connettere: un modello per la valorizzazione degli archivi musicali 147
Martina Gremignai, Luca Andrea Ludovico

LETSdigit: Archivio digitale della letteratura a Trieste 154
Alina Jill Simeone, Cristina Fenu, Marina Buzzoni, Roberto Rosselli Del Turco

La mostra *Coerenza In Coerenza* (1984) oltre la fotogrammetria: ricostruzione digitale tramite reti neurali 161
Filippo Yahia Masri, Maria Carmeliti

[3/A] AI E NLP PER TESTI E RECUPERO DELL'INFORMAZIONE

Creating a new gender-fair Italian morphological classifier based on UmBERTo fine-tuning 168
Irene Caiazzo, Giovanna Maria Dimitri

Assessing LLM Capabilities in Cultural Heritage: Recent Trends and Gaps in Benchmarks 175

Remo Grillo , Gianmarco Spinaci, Lukas Klic, Giovanni Colavizza

L'intelligenza artificiale per il potenziamento dell'information retrieval nei discovery tool delle biblioteche 183
Ilaria Belvedere, Simona Turbanti

Sinergie ludonarrative: l'evoluzione del racconto videoludico tra divulgazione, tecnica ed intelligenza artificiale 191
Giulia Angelini, Tommaso Mazzoli

[3/B] DH PER STORIA, DIRITTO E ARTE

Archivi ibridi di persona e preservazione forense dei materiali nativi digitali 198
Mariangela Giglio, Adele Gorini

Digital Humanities e studi neogreci: il progetto della Biblioteca Digitale dell'Osservatorio "Mario Vitti" 204
Georgios Katsantonis

artresearch.net: A Research Infrastructure for Exploring Interpretive Plurality in Art-Historical Photo Archives 211
Marilena Daquino, Francesca Mambelli, Alessandro Adamou, Pietro Maria Liuzzo, Stefanie Schneider, Spyros Koulouris, Artem Kozlov, Rafael Brundo Uriarte

Legislative Changes Concerning the Reproduction of Images of Italian Cultural Heritage Assets between 2023 and 2025 217
Marco Ciurcina, Piergiovanna Grossi

[3/C] ANALISI COMPUTAZIONALE DEL TESTO LETTERARIO

Flaiano postcoloniale? Un approccio computazionale alla *vexata quaestio* di *Tempo di uccidere* 226
Silvia Lilli, Daniel Raffini

Verso una rappresentazione del macrotesto: filologia d'autore e analisi quantitativa nei racconti di Fausta Cialente 234
Emmanuela Carbè

Analisi fluidodinamica dei campi semantici in testi letterari: il Numero di Reynolds come modello di turbolenza testuale 241
Pietro Sichera, Carla Perrone, Antonio Sichera

Dentro l'Officina del Testo: Modellare Strutture Letterarie Complesse 248
Enrica Bruno

[4/A] STUDI LETTERARI DIGITALI E DISTANT READING

Tracing Intertextuality at Scale. A Network Study of Poetic Influence 255
Gabriele Vezzani

The Landscape of Italianistica: Italian Literary Studies and Digital Humanities in Semantic Space 261
Maria Levchenko

L'Edicola della Storia di Trieste: NLP e *distant reading* Per il confronto tra periodici storici italiani e sloveni nella Trieste del 1902 267
Cristina Fenu, Giovanni Pinna

Literary Tropes as Prescriptive Devices: A Case Study from BookTok 274

Silvia Cappa, Mariangela Giglio, and Anna Sofia Lippolis

[4/B] MUSEI E ACCESSO DIGITALE

Rigenerare un patrimonio desueto con il phygital: la Casa Museo Bartolo Longo 282
Mauro De Bari

Musei senza confini: il digitale come infrastruttura di memoria partecipata nei musei delle aree interne. Il GeoMuseo Monte Arci Stefano Incani di Masullas e il Museo Paleontologico PARC di Genoni 290
Paola Palmas

Pubblici remoti: la Realtà Virtuale come infrastruttura culturale di accesso 296
Elisa Smeraldo

[4/C] FORMAZIONE, DIDATTICA E LABORATORI DH

Gli Hunger Games del LabCD. Un diario di sopravvivenza 304
Enrica Salvatori, Vittore Casarosa, Francesco Ricciardi

It's Buildin' Time: A Project-based Approach for Teaching Data Management to Cultural Heritage Graduate Students 309
Sebastian Barzaghi, Alba d'Elia, Esther Montserrat Giordano, Giulia Guidarelli, Gianluca Petrosillo, Elisabetta Sabattini, Fiammetta Sabba, Lucia Sardo

Integrazione di tecnologie per la traduzione e CLIL nella creazione di risorse didattiche per il sardo 319
Gianfranco Fronteddu, Igor Deiana

[5/A] LINKED OPEN DATA, MODELLI SEMANTICI, INFRASTRUTTURE

Reti sociali e linked data: persone e co-occorrenze documentarie nel Portale delle fonti per la storia della Repubblica italiana 329
Herbert Natta, Gianluca Rossi, Roberta Maggi

Web Portals on Linked Open Data: Testing Sampo-UI in Projects on Early Modern Bibliographic Data and Chinese Calligraphy and Arts 336
Arianna Moretti, Katarina Lučić, Iiro Lassi Ilmari Tiihonen, Jonas Paul Fischer

Risorse Terminologiche e Linked Open Data: il *Glossario delle Infrastrutture di Ricerca* (GIR) come Caso di Studio 345
Lucia Francalanci, Alessia Scognamiglio, Giulia Pedonese, Michele Mallia, Irene Falini, Fahad Khan

Orchestrazione di servizi e scenari avanzati: fruizione e necessità delle comunità di ricerca nel marketplace di H2IOSC 352
Pietro Sichera, Cristina Marras, Enrico Pasini, Vittoria Fabiani, Paolo Ongaro, Michele Scapicchi, Chiara Di Pietro

[5/B] PATRIMONIO CULTURALE: SISTEMI INFORMATIVI INTEGRATI

Architettura e *Digital Humanities*. Conoscenza e valorizzazione del Patrimonio attraverso la costruzione di sistemi informativi integrati 361
Donatella Rita Fiorino, Caterina Giannattasio, Elisa Pilia, Valentina Pintus ,Andrea Pirinu, Marcello Schirru

Tra testo e architettura digitali per lo studio delle trasformazioni storico-artistiche dei monasteri benedettini 368

Gianmario Guidarelli, Paolo Borin, Sonia Cavicchioli, Chantal Pivetta, Renato Canearo,

Dalla pergamena al virtuale. Manoscritti miniati nello spazio tridimensionale 375
Valeria Minisini, Bruno Fanini, Giorgio Gosti

Modeling and visualizing palimpsests with IIIF Presentation API 3.0 382
Panagiotis Leontaridis, Giacomo Marchioro, Glen Robson

[5/C] CROWDSOURCING E PUBLIC ENGAGEMENT

Tag ANIMALx: Crowdsourcing the Non-Human in Cultural Heritage Collections 389
Daniela Teixeira Gomes, Carla Vieira, Joana Vieira Paulino

From perception to memory: Public Engagement and Neurohumanistic Approaches to the Collective Reconstruction of Archaeological Heritage 396
Grazia Solenne, Vincenza Ferrara, Elisa Corrò

Comunicare la ricerca archeologica nell'ecosistema digitale: pratiche, metodi e sperimentazioni su Instagram 404
Marco Demuru

Public engagement nell'era dei synthetic media: per una co-interpretazione critica tra Digital Humanities e Media Education 411
Gabriele Prosperi, Mario A. Bochicchio

[6/A] MODELLAZIONE TESTUALE E LINGUISTICA COMPUTAZIONALE

Verso un dizionario narrativo computazionale degli usi linguistici in aree ad alta vulnerabilità sociale 420
Michela Bandini, Silvia Piccini, Andrea Bellandi, Emiliano Giovannetti

Verso la creazione di una Treebank per il sardo 426
Nicoletta Puddu, Manuela Sanguinetti, Luigi Talamo

A Learning-by-Doing Approach to Spoken Data Collection: The Case of the LPSP Corpus 432
Claudia Roberta Combei, Gaia Eleonora Di Raimondo, Sofia Maestri, Benedetta Romanazzi, Elena Scotti

Digital Humanities e Letteratura Elettronica in Italia: marginalità e convergenze 440
Giuseppe Arena

[6/B] CO-CREAZIONE, RICERCA PARTECIPATIVA, COMUNITÀ

Metodi e strumenti di sperimentazione tra co-creazione, Digital Humanities, e ricerca demo-etno-antropologica: la digitalizzazione come pratica relazionale ed etica. 448
Eleonora De Longis, Matteo Cova, Giovanni Bertelli

Integrare strumenti digitali e conoscenze della comunità per una progettazione urbana resiliente al clima: l'iniziativa Dundrum by Design 455
Chiara Cocco, Mattia Leone, Antonio Savino, Miruna Popa, Sara Tedesco

HCI e AI per il Public Engagement: casi di studio per la valorizzazione del patrimonio culturale 462
Samuel Aldo Iacolina, Manuela Angioni, Valentina Marotto, Francesca Mura, Piergiorgio Palla

The Automation of Research in Digital Humanities: Artificial Researchers, Digital Archives, and the Democratization of Scholarly Access 468
Roberto Di Quirico

[6/C] AI E AUTOMAZIONE PER ARCHIVI E PATRIMONIO

Dal problema storico alle IA: Il Database on the Slave Trade between the Mediterranean and the Atlantic (15th-16th Centuries) 477
Salvatore Spina, Carlo Taviani, Jörg Hörnschemeyer

Peirce Interprets Peirce: Digitization, Automation, and Interpretation in Charles Peirce's Manuscripts 485
Alessandro Adamou, Sebastian Feil, Davide Picca, Carlo Teo Pedretti, Dario Rodighiero,
Lorenzo Zangari

A New Semantic Model for Mythologiae: from Data to Interpretation 494
Bianca La Manna

[7/A] ONTOLOGIE, FAIR DATA E MODELLAZIONE DELLA CONOSCENZA

The ATLAS Guidelines for producing FAIR research products in the Digital Humanities 502
Chiara Martignano, Giorgia Rubin, Sebastiano Giacomini, Alessia Bardi, Marina Buzzoni, Marilena Daquino, Riccardo Del Gratta, Angelo Mario Del Grosso, Franz Fischer, Roberto Rosselli Del Turco, Francesca Tomasi

RATIO: strumenti per la creazione e la gestione di risorse e collezioni FAIR nell'ambito delle scienze umane 509
Federico Silverstri, Lorenzo Mancini, Laura Baggiani, Marco Bagiacchi

Making Uncertainty Explicit: The Preservation of Interpretive Complexity across Digital Humanities Project Workflows 516
Tommaso Battisti, Valentina Pasqual, Giulia Renda, Marilena Daquino

Iconclass as a Semantic Framework for the Analysis of the Sacred Space 524
Polina Voronova

[7/B] GIS E SPAZIALIZZAZIONI DIGITALI

Gli *Historical GIS* e la storia forestale: potenzialità e applicazioni 531
Vincenzo Colaprice

Le nazioni genovesi nell'Atlante digitale di Storia Marittima del Regno di Sardegna (ASMSA). Lavori in corso 539
Giampaolo Salice, Giommaria Carboni

Integrating Romanian Toponymy in a GIS Background: dictionary forest names vs. OpenStreetMap forest data 547
Roxana Patras, Ana Odochiciuc, Mihai-Bogdan Atanasiu, Constantin Răchită

Digital Humanities per narrare i territori: un WebGIS sui porti minori della Sardegna dal medioevo alla contemporaneità. Una proposta di implementazione collaborativa 554
Luigi Serra, Sebastiana Nocco

[7/C] MODELLI, METODI E PROSPETTIVE DELL'EDIZIONE DIGITALE

Roman and Justinianic Legal Terminology in Thirteenth-Century Western European Diplomatic Sources 562

Tamás Kovács, Anguelos Nicolaou, Johannes Laroche, Georg Vogeler

In Search of Lost Time: an Innovative Model for Representing Autograph Texts 567
Giulia Baldelli, Daniele Fusi, Matteo Zupancic, Franz Fischer, Claus Zittel

Naples Dante Project. Un ambiente integrato per l'analisi bibliografica, filologica e iconografica 573
Gennaro Ferrante, Andrea La Veglia, Daniele Fusi, Stefano Angelo Rizzo, Angelo Eugenio Mecca, Daniele Duranti, Anna Sviridova

La curatela biblioteconomica delle raccolte digitali MAB: linee gestionali, requisiti, workflow 580
Nicola Barbuti

[8/A] GRAFI, LINKED DATA E MAPPING

DOG: an open-source ecosystem for mapping, narrating, and exploring humanities data 588
Alessandro Laruffa, Alessandro Capra

Beyond Spatial Data: Exploring Gender and Property in the Catasto Generale Toscano through the Florentia Illustrata Semantic Platform 596
Erica Andreose , Remo Grillo , Gianmarco Spinaci

Per la storia del mercato dell'arte tra otto e novecento: raccolta, modellazione e valorizzazione dei dati sugli antiquari italiani 602
Valentina Rossetti, Marilena Daquino, Francesca Mambelli, Ludovica Pannitto, Francesca Tomasi

Un grafo per la ricerca storico-filologica in archivio: il caso del fondo Giuseppe Albini 609
Lucia Giagnolini, Mohamed Iheb Ouerghi, Valentina Pasqual, Cecilia Tamagnini, Francesca Tomasi

[8/B] GEOGRAFIE LETTERARIE E MULTIMEDIALI

'Spazializzare' la narrativa ecocritica italiana: la piattaforma dell'Atlante digitale della letteratura ecologica italiana 617
Valentina Pasqual

MeMo: mappe digitali, memoria letteraria e public engagement nel Mezzogiorno 624
Laura Giurdanella; Giuseppe Palazzolo

Una proposta di approccio computazionale alle geografie del cinema italiano 631
Alberto Savi, Luca Giovannini

Il campo semantico del fuoco nella Commedia: un caso di studio sulla pertinenza lessicale e sulla visualizzazione 638
Ruoci Song

[8/C] ANNOTAZIONI E MARKUP

Ritagli di teatro. Annotazione semantica della rassegna stampa del Teatro Sant'Erasmo 644
Paolo Bonora

Annotazione semantica delle formule poetiche del Medioevo germanico 651
Roberto Rosselli Del Turco, Letizia Vezzosi

Ça y est : Annotating Arabic Texts for Teaching 658

Dal markup all'interpretazione: interrogare il linguaggio della critica verista in un corpus XML/TEI 665
Denise Bruno, Salvatore Cristofaro, Antonio Di Silvestro, Laura Mazzagufo, Daria Spampinato, Giuseppe Zappalà

[9] SESSIONE POSTER

Archivi letterari nativi digitali: modelli descrittivi ed edizioni sperimentali 673
Elena Barchielli, Simon Willemin, Elena Spadini.

Lexicon Philosophicum: una rivista digitale per la condivisione della conoscenza umanistica 679
Pamela Barletta, Maria Cristina Dalfino, Pietro Restaneo

Intelligenza artificiale, reference digitale e accessibilità: l'esperienza di Alphabetica 684
Flavia Bruni – Elisabetta Castro

Grafoteca e la costruzione digitale della memoria manoscritta: curatela, metadatazione e reti di risorse 691
Amalia Carrano

Valorizzazione digitale del patrimonio culturale: Public-Private Partnership e partecipazione sociale 697
Annalisa Ciliberti

I Vangeli digitali e accessibili, per un'editoria religiosa universale e senza barriere 703
Giulia D'Arcangelo

Implementing a controlled vocabulary for medieval administrative acts: theoretical and methodological considerations 709
Chiara De Bastiani

L'Archivio del Centro Conservazione e Restauro "La Venaria Reale": un caso studio per la diffusione dei dati tecnico-scientifici 717
Stefania De Blasi, Lorena Palmieri, Chiara Pipino

Strumenti digitali per i percorsi di visita della Palazzina di Caccia di Stupinigi: Residenza Sabauda e Museo dell'Ammobiliamento 722
Stefania De Blasi

Strategie di protezione, governance e sostenibilità degli archivi orali 727
Rosaria De luca, Elvira Mercatanti, Monica Monachini

FAIR Memories: A Workflow for the Preservation and Collaborative Reuse of Analog Oral History Collections 734
Nike del Quercio, Costanza Paolillo, Laurent Fintoni

Ontologycore: A Formal Knowledge Graph of Internet Aesthetics 740
Anouk Flinkert, Ekaterina Krasnova, Shiho Nakamura

Il ruolo del Test Plan per la validazione: il caso d'uso del Marketplace H2IOSC 747
Vittoria Fabiani

Partecipazione, mediazione e conoscenza: Àndalas de Cultura quale ecosistema digitale del patrimonio culturale della Sardegna 754
Carolina Floris

From Violin Lessons to Linked Open Data. Reconstructing Tartini's *Scuola delle Nazioni* through Digital Prosopography and Digital Editions 762
Selina Galka, Marcella Tambuscio, Rolf Wissmann, Cristina Scuderi, Georg Vogeler

Addestramento di un modello di HTR in Transkribus per testi plurilingui di età moderna: il caso del *Ripulimento della lingua sarda* (Madau) 768
Michela Incollu, Giulia Murgia

Modelli digitali interattivi per gli archivi teatrali storici: mappare le relazioni semantiche al Teatro Metastasio di Prato (1827–1860) 775
Matilde Innocenti

Tra fotografia e segno: la trascrizione informatizzata dei registri storici Alinari 779
Pamela Krzemien

Biblioteche filosofiche private e memoria culturale: http://picus.unica.it/ 786
Andrea Lamberti, Giovanna Granata

Beyond Archives: A Three-Stage Workflow for Relational Digital Libraries of Written and Oral Memory Sources 791
Giulia Lembo

A Workflow-Centred Approach to Managing Semantic Artifacts as Linked Open Data 800
Michele Mallia, Fahad Khan, Valeria Quochi

SEBASTIAN: uno strumento per l'organizzazione, l'analisi spaziale e la visualizzazione dei dati storici 807
Christian Marcantonio, Giuseppe Consolo

Raccontando e valorizzando il patrimonio culturale. Dalle prime esperienze ai progetti futuri del CNR-ISEM 814
Maria Grazia Mele R., Samuel Aldo Iacolina, Luigi Serra, Giovanni Serreli

La Digital Library e l'Ecosistema digitale per la cultura come opportunità di *public engagement* 821
Federico Meschini, Biancamaria Hermanin, Antonella Negri, Giuliano Romalli

Strategie di engagement nella creazione e sostenibilità di un parco letterario: i risultati del progetto CHANGES 827
Sara Obbiso, Nicola Mariniello, Davide Bagnaresi, Riccardo Stracuzzi, Giuliana Benvenuti, Alessandro Iannucci

Lo studio integrato del patrimonio materiale e immateriale per la ricostruzione delle pratiche storiche di gestione delle risorse ambientali nell'Alpujarra (Sierra Nevada): il progetto Sinergy 835
Alessandro Panetta

Reveduti, Correcti, Aprobati et Confirmati: The Digital Edition of the Statutes of Ascoli 841
Michela Parma, Fabio Mariani

Trascrivere automaticamente i registri catastali sardi del primo Novecento con Transkribus: risultati di un caso di studio 849
Andrea Pergola, Cecilia Tasca

Ricostruzione virtuale di un patrimonio disperso: le collezioni di disegni a Genova nel XIX secolo 856
Giulia Pilosu

Dai workflow descrittivi ai workflow applicativi tramite AEON e l'IA 860
Francesco Pinna, Emiliano Degl'Innocenti

Gamification per la valorizzazione dei beni musicali. Un'ipotesi di raccolta di dati catalografici in *crowdsourcing* 867

Marcello Ranieri

Tra riconoscimento disciplinare e prospettiva pubblica: un modello logico-applicativo per le conoscenze e le competenze nelle DH 874

Lisa Reggiani

Mappe, scale e interpretazione. Per una cartografia digitale semantica dello spazio narrativo 883

Lorenzo Sabatino

Verso un grande corpus diacronico dell'italiano: il contributo degli esempi del GDLI 893

Eva Sassolini, Sebastiana Cucurullo, Marco Biffi e Simonetta Montemagni

Verso una classificazione funzionale dei prompt basata sulle intenzioni cognitive esplicitate dagli studenti nell'interazione con l'IA 901

Daniele Scala, Salvatore Varriale

Il videogioco come strumento di mediazione interculturale 909

Alessia Stocco

Edizione digitale commentata di *Le libere donne di Magliano* di Mario Tobino: primi passi, metodi e prospettive 917

Fabio Zarroli

Il restauro testuale dei papiri nel progetto GreekSchools: CoPhiEditor e l'integrazione di modelli linguistici 923

Premessa

AIUCD 2026, il XV Convegno annuale dell'Associazione per l'Informatica Umanistica e la Cultura Digitale (AIUCD), è dedicato al tema "Digitale e Public Engagement: pratiche e prospettive nelle Digital Humanities". Il convegno mette al centro pratiche, strumenti e metodologie che rendono la conoscenza più aperta, condivisa e inclusiva e contribuiscono alla diffusione di una cultura digitale, intesa come componente costitutiva della cittadinanza contemporanea

Il tema è al centro delle attività del DH UNICA – Centro interdipartimentale per l'Umanistica Digitale dell'Università di Cagliari, organizzatore del Convegno in collaborazione con AIUCD. Il Centro, istituito nel 2020, conta attualmente 76 ricercatrici e ricercatori, espressione di 5 dipartimenti dell'Università degli studi di Cagliari sia di Scienze Umane e Sociali sia di STEM (Letlib, Scienze Politiche, Dicaar, Pedagogia, Psicologia, Filosofia, Ingegneria elettrica ed elettronica).

Il DH UNICA è dotato di un ecosistema digitale, che ottimizza la gestione di una ventina di progetti di ricerca, ed è articolato in ambiti di ricerca e in gruppi di lavoro. Per il carattere fortemente multidisciplinare del Centro e per le tematiche che vi gravitano, come storia digitale pubblica, archivi, musica, quest'anno il XV Convegno dell'AIUCD 2026 si è orientato a dare un contributo ad un dibattito attuale e molto vivo nella comunità DH anche internazionale: l'impatto e il coinvolgimento del pubblico.

Le Digital Humanities assumono oggi, infatti, un ruolo sempre più centrale nella costruzione di pratiche di ricerca aperte e partecipative. La crescente disponibilità di strumenti digitali avanzati, dai database collaborativi alle piattaforme di intelligenza artificiale, offre opportunità senza precedenti per coinvolgere pubblici eterogenei nella produzione e interpretazione della conoscenza. In questo scenario, il public engagement non è più un'attività marginale o secondaria, ma un elemento strutturale della ricerca, in grado di generare nuove forme di collaborazione, arricchire la qualità dei dati e favorire una diffusione più ampia e inclusiva dei risultati scientifici. La riflessione sulle modalità di partecipazione, sulle tecnologie utilizzate, sulle politiche della ricerca e sugli effetti sociali e culturali di questi processi rappresenta quindi un tema cruciale per comprendere e orientare l'evoluzione delle Digital Humanities nel presente e nel futuro.

Il concetto di comunità scientifica è uno dei punti chiave del Convegno, il programma coinvolge in un dialogo attivo la comunità educante, il pubblico, le istituzioni e le politiche della ricerca. Il progetto è quello di costruzione di una società democratica, in cui la comunicazione pubblica della scienza, l'alfabetizzazione digitale, contribuiscono a una formare una cittadinanza scientifica digitale: accesso all'informazione e alla sua produzione, inclusione, trasparenza…

Il XV Convegno annuale AIUCD 2026 si propone di esplorare proprio il rapporto tra Digital Humanities e public engagement, offrendo spazio a modelli teorici, metodologie, servizi, applicazioni pratiche che favoriscono il coinvolgimento attivo dei cittadini nei processi di ricerca, documentazione nel campo delle Humanities. Sono temi di interesse del convegno le esperienze che, attraverso il digitale, attivano percorsi di co-creazione, crowdsourcing, narrazione digitale, musealizzazione, archiviazione virtuale e di citizen science, heritage sharing, dentro cornici Open Access e con approcci innovativi anche alla comunicazione scientifica. Metodi computazionali, sfide metodologiche e soluzioni sperimentali volte a migliorare accessibilità, interoperabilità, riuso e interpretabilità dei dati all'interno degli ecosistemi di ricerca digitale, per favorire la contaminazione e l'integrazione tra discipline e ambiti disciplinari e le pratiche di scienza aperta. AIUCD 2026 rappresenta un approfondimento di molteplici aspetti dell'ampio panorama delle Digital Humanities, con il fine di alimentare il confronto interdisciplinare e il dialogo scientifico in forma inclusiva.

La comunità degli umanisti digitali è stata invitata a contribuire al Convegno presentando proposte orali o in forma di poster intorno a 5 aree tematiche.

La prima area tematica, "DH e co-costruzione del sapere con le comunità: sfide, metodi, strumenti", è stata pensata per esplorare la partecipazione attiva di pubblici eterogenei mediante metodi collaborativi (crowdsourcing, citizen humanities, digital public history, pratiche di co-creazione, community museums ecc.) e strumenti digitali, inclusi sistemi basati sull'utilizzo dell'intelligenza artificiale, quest'ultimo tema importantissimo con il quale la comunità DH deve confrontarsi direttamente e riflettere. Questa area tematica include anche contributi che indagano i principi di citizen science in attività condivise di

produzione di conoscenza, come trascrizioni, annotazioni o mappature partecipate, così come lavori che favoriscono l'Open Access e l'adozione dei principi FAIR per garantire accessibilità, interoperabilità e riusabilità durevole di dati e risorse. Trovano particolare attenzione riflessioni e lavori che promuovono equità, accessibilità, cittadinanza digitale e inclusione di comunità marginalizzate tramite iniziative formative e pratiche educative partecipative in scuole, università, musei, MOOCs o laboratori territoriali.

La seconda area tematica, "Archivi ed edizioni: descrizioni aumentate, accessibilità e sistemi informativi", apparentemente più tradizionale, ha inteso focalizzarsi su riflessioni teoriche e applicative in merito alle opportunità e alle conseguenze dei processi di dematerializzazione della mediazione, al concetto di descrizione aumentata, intesa come opportunità di creare strumenti di mediazione capaci di integrare diverse entità in direzione dell'evoluzione dei sistemi informativi archivistici in sistemi interculturali, secondo le logiche che emergono da RiC e dagli standard di seconda generazione in genere. In questo senso trova particolare spazio il tema della rappresentazione e organizzazione a lungo termine del patrimonio culturale pregresso e dei materiali nativi digitali. Inevitabile in questo contesto una riflessione sulle strategie descrittive e sugli strumenti di restituzione delle informazioni, con riferimento alla creazione di risorse partecipative, alla realizzazione e gestione di edizioni critiche digitali e allo sviluppo di nuovi modelli di rappresentazione e visualizzazione dei dati che tengano conto delle opportunità di interazione e codifica orientata che la digitalità garantisce. Sono presenti proposte in tema di conservazione, gestione, sostenibilità e trasmissione di collezioni digitalizzate e nativamente digitali; progettazione di infrastrutture digitali resilienti e l'adozione di modelli sostenibili per la cura dei dati. Sono inoltre incluse presentazioni sulle buone pratiche e politiche per garantire l'accesso aperto, FAIR e a lungo termine, l'autenticità e la riusabilità delle risorse del patrimonio culturale digitale.

La terza area tematica, "Testualità digitali: prospettive, sviluppi e sperimentazioni", accoglie proposte che esplorano principalmente metodi e strumenti innovativi per lo studio computazionale dei testi, dai corpora letterari ai materiali del patrimonio culturale. Gli ambiti di interesse affrontati includono modelli computazionali per la critica letteraria, progressi nella linguistica computazionale e applicazioni di elaborazione del linguaggio naturale (NLP). Particolare attenzione è data a riflessioni ed esperienze che impiegano l'Intelligenza Artificiale e il Machine Learning per interpretare, modellare e visualizzare il patrimonio culturale, sia materiale che immateriale (ciò include anche fonti storiche, testi letterari). Analisi di rete, visualizzazione dei dati e tecniche di mappatura, soprattutto quando contribuiscono a superare i confini metodologici esistenti o aprono nuove prospettive interpretative sono temi di grande rilievo per la discussione.

La quarta area tematica, "Rappresentazione di dati e della conoscenza", intercetta ricerche dedicate alla rappresentazione della conoscenza e dell'uso delle tecnologie del Web Semantico all'interno delle discipline umanistiche digitali: progettazione di ontologie, modelli concettuali e grafi di conoscenza per i dati del patrimonio culturale, creazione di Linked Open Data e dataset conformi ai principi FAIR, e tecniche automatizzate di estrazione dei metadati, riconoscimento di entità nominate ed arricchimento semantico delle risorse culturali.

La quinta area tematica, "Memorie, storia e patrimoni culturali digitali", ha inteso approfondire principalmente la storia digitale, la co-generazione della memoria e il racconto del patrimonio culturale in digitale, incluse specifiche riflessioni su preservazione degli esiti delle ricerche e dei progetti scientifici digitali, anche nella prospettiva della scienza aperta (Open Science) e con applicazione dei principi FAIR. Inevitabile una riflessione sulle criticità infrastrutturali e sui problemi di sostenibilità dei progetti legati alla digitalizzazione e alle collezioni nate digitalmente, nonché proposte per la messa in atto di buone pratiche che garantiscano preservazione e accesso effettivo al patrimonio con abbattimento di impedimenti fisici, geografici, economici.

Al XV Convegno annuale AIUCD 2026 sono pervenute 151 proposte, 134 sono stati i contributi accettati: una comunità vivace e collaborativa, infatti si è avuta una media di 2,46 autori per contributo, 18 sono gli autori e i co-autori internazionali, a testimonianza delle collaborazioni internazionali e interdisciplinari della nostra comunità. I dati ci dicono, inoltre, quanto le Digital Humanities siano ormai presenti attivamente nel territorio nazionale, 163 tra Università, istituti CNR, centri e istituzioni di ricerca e culturali sono rappresentate al Convegno; Tutte le proposte sono state valutate attraverso un processo di revisione a doppio cieco, gestito da un Comitato Scientifico composto da 10 track chair e un Comitato di Programma di 94 revisori, oltre ai General Chairs.

I contributi, 95 relazioni orali e 39 poster, sono stati organizzati in panel per favorire lo scambio e la discussione su tematiche specifiche, due sessioni sono specificatamente dedicate alla presentazione dei poster.

Il Convegno è chiuso dall'incontro tra Philip Crowe, Direttore di Ricerca dell' UCD, School of Architecture, Planning and Environmental Policy, dell'University College di Dublino, che ha presentato una riflessione sul sostegno che le tecnologie digitali possono dare ad approcci basati sul territorio per affrontare le sfide sociali nelle città e contribuire a coinvolgere le comunità nel processo di cambiamento. Questa riflessione è oggetto di un dialogo tra Crowe e Arianna Ciula (King's College London) e Francesca Tomasi (Alma Mater Studiorum Università di Bologna

Anche quest'anno il Convegno è stato preceduto da Aspettando AIUCD2026 che ha visto una stretta collaborazione tra l'Università di Cagliari con il DH UNICA - Centro interdipartimentale per l'Umanistica Digitale, i gruppi AIUCD Scuola (e il Liceo Statale Mamiani di Roma), Osservatorio DH AI, ALDiNA, e il Clarin Knowledge Center DiPText, e con il patrocinio di Wikimedia Italia e AIPH. La rassegna è stata curata, seguita e animata da Raffaele Argiolas, DH UniCa, Università degli Studi di Cagliari; Emmanuela Carbé, VeDPH, Università Ca' Foscari di Venezia; Federico Boschetti, CNR-ILC, Pisa – VeDPH, che ringraziamo per questa importante occasione di scambio

Dal punto di vista organizzativo questo XV Convegno annuale AIUCD 2026 è stato gestito da un Comitato Organizzativo multidisciplinare, che ha cercato di trovare soluzioni funzionali e innovative alla gestione del Convegno anche per favorire la socializzazione e lo scambio tra i partecipanti e, soprattutto, offrire degli strumenti che rendessero la proposta scientifica e il soggiorno a Cagliari fruibile, accessibile e condivisibile.

Per favorire il più possibile la fruizione del fitto programma, grazie ad Alessandro Capra, abbiamo potuto offrire ai nostri convegnisti un Companion del convegno, che raccoglie, in modo leggero e navigabile, tutto quello che serve sapere — dal programma minuto per minuto alle planimetrie delle aule, dalle 134 relazioni accettate ai luoghi più belli di Cagliari raggiungibili a piedi dalla sede. Companion opera insieme a Noa, un piccolo applicativo che propone ai convegnisti dei collegamenti tra panel, autori e sessioni, costruendo la propria personale scaletta di appuntamenti da seguire.

Concludiamo questa prefazione con un sentito ringraziamento a tutte le persone e gli enti che hanno reso possibile la realizzazione di AIUCD 2026. Un ringraziamento speciale va al Direttivo AIUCD, e in particolare alla presidente Marina Buzzoni, al segretario Paolo Monella e alla tesoriera Francesca Frontini, per il supporto istituzionale e la preziosa collaborazione in tutte le fasi preparatorie del convegno e nelle giornate Cagliaritane. Siamo grati al Comitato Scientifico e al Comitato di Programma, per la disponibilità e l'attenzione nella valutazione delle proposte, così come a tutte le persone che hanno risposto alla call for papers e partecipato al convegno.

Un grazie sentito va al Comitato Organizzatore, per l'instancabile impegno, e la generosa disponibilità nella gestione di ogni fase del convegno.

Ringraziamo inoltre la direzione e il personale della segreteria amministrativa del Dipartimento di Lettere, Lingue e Beni culturali e della Facoltà di Studi Umanistici che hanno ospitato, sostenuto e co-finanziato il convegno e con loro gli studenti dell'Università di Cagliari che hanno offerto un prezioso contributo durante le tre giornate dell'appuntamento.

Infine, rinnoviamo la nostra gratitudine ai nostri sponsor: Presidenza della Regione Sardegna, Fondazione di Sardegna, Comune di Cagliari, all'ISSASCO e all'Istituto Parri, Associazione Iztokor, Clarin, che, in varie forme, hanno sostenuto AIUCD 2026 e ne hanno reso possibile la realizzazione.

In chiusura, ci piace pensare che la cittadinanza digitale non risiede nelle tecnologie che scegliamo di usare o meno, ma nelle comunità che nel digitale individuano un mezzo e una filosofia per dare forza a progetti culturali e scientifici pensati e realizzati per e con le persone.

Cagliari, maggio 2025

CRISTINA MARRAS, ANDREA PERGOLA, GIAMPAOLO SALICE

Legenda
Le 5 aree tematiche del Convegno sono indicate attraverso la seguente legenda


Archivi ed edizioni: descrizioni aumentate, accessibilità e sistemi informativi


DH e co-costruzione del sapere con le comunità: sfide, metodi, strumenti


Memorie, storia e patrimoni culturali digitali


Rappresentazione di dati e della conoscenza


Testualità digitali: prospettive, sviluppi e sperimentazioni


Altri contributi su Digital Humanities

[1/A] Intelligenza Artificiale generativa e riflessione teorica

Una Agenda sull'AI generativa per Digital Humanities

FABIO CIOTTI
UNIVERSITÀ DI ROMA - TOR VERGATA, ITALIA
FABIO.CIOTTI@UNIROMA2.IT

ABSTRACT (ITALIANO)

L'impatto trasformativo dell'Intelligenza Artificiale generativa ci impone, come comunità italiana delle Digital Humanities, di predisporre un'agenda strategica di riflessione collettiva. In questo intervento propongo un contributo in tale direzione, individuando tre direttrici di analisi interconnesse: (1) la teoria dell'IA generativa, che richiede approcci esplicativi non riducibili alle sole spiegazioni matematico-informatiche, incorporando descrizioni funzionali e nozioni intenzionali per comprendere le proprietà linguistiche e cognitive emergenti dei Large Language Models; (2) l'IA come metodo nella ricerca umanistica, che riorganizza profondamente i workflow epistemici permettendo l'operazionalizzazione di quasi ogni aspetto della ricerca, dalla costruzione dei corpora alla produzione di interpretazioni ermeneutiche; (3) l'IA come tecnologia socio-culturale, considerando i LLM come modelli della memoria culturale che a loro volta retroagiscono sui sistemi culturali stessi. I sistemi di IA generativa rappresentano una rivoluzione epistemologica che impone di riconcettualizzare assunzioni fondamentali su significato, cultura, conoscenza e interpretazione. La tradizione metodologica delle Digital Humanities italiane può svolgere un ruolo critico essenziale nello sviluppo di protocolli d'uso controllato e pratiche di cooperazione tra IA e agency umana, orientandoci anche come studiosi che hanno una responsabilità sociale.
Parole chiave: Intelligenza Artificiale generativa, Epistemologia delle Digital Humanities, Large Language Models, Impatto sociale delle DH

ABSTRACT (ENGLISH)

An Agenda on Generative AI for Digital Humanities.
The transformative impact of generative Artificial Intelligence compels us, as the Italian Digital Humanities community, to establish a strategic agenda for collective reflection. In this intervention, I propose a contribution in this direction by identifying three interconnected lines of analysis: (1) the theory of generative AI, which requires explanatory approaches not reducible to mathematical-computational explanations alone, incorporating functional descriptions and intentional notions to understand the emergent linguistic and cognitive properties of Large Language Models; (2) AI as a method in humanistic research, which profoundly reorganizes epistemic workflows by enabling the operationalization of nearly every aspect of research, from corpus construction to the production of hermeneutic interpretations; (3) AI as a socio-cultural technology, considering LLMs as models of cultural memory that in turn act back upon cultural systems themselves. Generative AI systems represent an epistemological revolution that demands a reconceptualization of fundamental assumptions about meaning, culture, knowledge, and interpretation. The methodological tradition of Italian Digital Humanities can play an essential critical role in developing protocols for controlled use and practices of cooperation between AI and human agency, guiding us also as scholars who bear social responsibility.
Keywords: Generative Artificial Intelligence, Epistemology of Digital Humanities, Large Language Models, Social impact of DH

1. INTRODUZIONE

La comunità di studi che nel corso degli anni ha dato vita alla Galassia delle Digital Humanities (Ciotti, 2023a) si è sempre confrontata con l'evoluzione, in gran parte autonoma, dei modelli, dei sistemi e delle infrastrutture tecno-informatiche (l'introduzione del personal computing, poi delle interfacce grafiche, e ancora l'avvento del Web, la digitalizzazione di massa, il Web semantico e i linked data…). Assunzioni teoriche, framework epistemologici e metodologici, domini di ricerca che hanno popolato la storia concettuale del nostro campo di studi e di pratiche sono sempre state negoziate con quella sfera a un tempo aliena e familiare, ponendoci peraltro nel difficile ruolo di mediatori epistemici e culturali tra il

complesso dei saperi umanistici e quello dei saperi tecno-informatici (Buzzetti, 2019). Percorsi ed esiti di questa mediazione sono stati molteplici, contraddittori e fortemente differenziati anche in base all'ambiente linguistico e socioculturale in cui si sono situati, sebbene la fase più recente delle Digital Humanities sia stata di fatto caratterizzata da una sostanziale globalizzazione e omogeneizzazione internazionale, tanto da rendere assai meno nette quelle stesse differenze di tradizioni culturali (Horvath et al., 2022).
La storia del presente ci pone di fronte ancora una volta a una innovazione tecno-scientifica il cui potere trasformativo promette di rivelarsi ancora più profondo e pervasivo di tutti quelli cui abbiamo assistito, e che siamo stati in grado di metabolizzare concettualmente e metodologicamente, nello scorso cinquantennio: l'esplosione dell'Intelligenza Artificiale (generativa). La questione terminologica, alla quale come umanisti digitali siamo assai avvezzi, ci potrebbe portare a criticare, distinguere, puntualizzare, sulla correttezza, adeguatezza, legittimità di questo termine; ma mi sembra di poter dire che di fronte all'enormità della cosa, mettersi a disquisire sulla parola sia futile, e comunque meno urgente del capire e riflettere su come collocarci di fronte a questa travolgente novità scientifica e tecnologica. Si tratta insomma di discutere collettivamente su quale debba essere l'Agenda sull'IA per il nostro ambito di studi. E ci sono buone ragioni per farlo, pur nella diversità delle posizioni su questioni anche fondazionali, perché tra tutte le grandi innovazioni tecnologiche questa dell'IA avrà di sicuro un grande impatto su quella sfera di pratiche e artefatti di cui ci occupiamo come umanisti, la sfera dei linguaggi e della cultura; perché essa avrà certamente un impatto enorme nei nostri stessi metodi, approcci e saperi come umanisti digitali, mettendo di fatto in questione la nostra stessa specificità; ma soprattutto perché potrebbe essere quella con le più profonde conseguenze in generale, dico a livello di storia della civilizzazione.
Agli albori di questa esplosione, chi scrive aveva già posto all'attenzione della comunità DH italiana le possibilità che i nuovi modelli linguistici generativi avrebbero potuto offrire per i nostri studi (Ciotti, 2023b). Oggi dopo tre anni di sviluppo e diffusione pervasiva, l'impatto trasformativo è ormai evidente, e non a caso si registrano, oltre a decine di studi, esperimenti, riflessioni teoriche, anche alcuni tentativi di elaborazione strategica collettiva: (Hemment & Kommers, 2025) (Kommers et al., 2025) (Klein et al., 2025) (*AI Manifesto*, 2025), a tante altre iniziative di discussione e riflessione, non ultimo l'Osservatorio italiano sull'Intelligenza Artificiale e le Digital Humanities costituito come SIG della AIUCD (https://aiucd.github.io/DH-AI). In questa sede voglio provare a fornire il mio contributo a questa riflessione collettiva, perché ritengo che la nostra Associazione scientifica e la comunità che in essa si riconosce possa e debba essere protagonista di questo dibattito, e che anzi possa e debba porsi ancora una volta come riferimento per tutto il mondo dei saperi umanistici.

2. IA E DH: TRE DIRETTRICI PER LA RIFLESSIONE

Inizio con due considerazioni preliminari. La prima, condivisa peraltro da molti studiosi e gruppi di ricerca, è che pensare strategicamente sull'impatto dell'Intelligenza Artificiale generativa sulle Digital Humanities significa anche pensare simmetricamente sul contributo che le DH possono offrire alla ricerca sull'IA, sia che si considerino questi sistemi come mere tecnologie culturali, sia che le si intendano come agenti cognitivi di un qualche tipo nell'ambito di una ontologia inclusiva del mentale, sia che li si concepisca come agenti che esibiscono proprietà cognitive identiche a quelle umane. La seconda è che la riflessione si deve articolare secondo alcune direttrici, ambiti di discorso distinti ma ovviamente interconnessi; qui ne propongo tre: (1) La riflessione teorica e sperimentale sull'IA generativa e sulla natura e proprietà dei Large Language Models; (2) la riflessione epistemologica sull'IA come metodo nella ricerca umanistica; (3) la riflessione sull'IA come tecnologia socioculturale e sui suoi impatti.

2.1 Teoria dell'IA generativa

Il punto di partenza è una constatazione ormai difficilmente contestabile: le spiegazioni esclusivamente informatiche e matematiche dei meta-algoritmi alla base degli LLM, pur necessarie, non sono sufficienti a spiegare i loro comportamenti reali, né possiedono una reale capacità esplicativa o predittiva rispetto alle proprietà linguistiche e cognitive emergenti che osserviamo empiricamente.
Un parallelismo con le neuroscienze può essere istruttivo. Così come una teoria del comportamento umano non può essere ridotta alla biochimica o alla neurofisiologia – pur includendole – allo stesso modo una teoria degli LLM non può arrestarsi al livello architetturale o ingegneristico implementativo. Essa deve incorporare descrizioni funzionali e, in senso controllato e non ingenuo, anche nozioni intenzionali:

parlare di competenze, di capacità, di generalizzazione, di sensibilità contestuale. Non perché i modelli necessariamente "abbiano" intenzioni, ma perché senza questi livelli descrittivi non siamo in grado di spiegare ciò che fanno.
In questo quadro si collocano questioni tipicamente umanistiche: quali teorie del significato e della comprensione sono compatibili con le evidenze empiriche sulle performance linguistiche degli LLM? Che cosa sono esattamente i modelli del mondo, o la teoria della mente, e fino a che punto possiamo ascrivere queste proprietà ai sistemi AI? Che tipo di processi di ragionamento, se lo sono, possiamo attribuire ai modelli linguistici con *reasoning* avanzato che sono stati in grado di conseguire risultati stupefacenti nel dominio della ricerca matematica, contro ogni previsione (Barkeshli et al., 2026; Klowden & Tao, 2026)?
In questo senso, gli LLM non sono solo oggetti da spiegare, ma anche strumenti epistemici indiretti: sistemi che possono fornire nuove evidenze per valutare, validare o mettere in crisi teorie classiche della cognizione, del linguaggio e dei sistemi culturali che su di esso si basano.
A questo livello teorico appartiene anche la questione della fondazione concettuale dell'etica dell'IA. Infatti, credo che sia possibile costruire una seria teoria etica "su e per" l'intelligenza artificiale, e a maggior ragione sviluppare un approccio di *ethics by design*, solo chiarendo lo statuto ontologico e funzionale di questi sistemi e costruendo una teoria filosoficamente solida relativa a quali concetti di responsabilità, autonomia, agentività, affidabilità e autorità epistemica sono applicabili a queste entità (De Caro & Giovanola, 2025).

2.2 IA generativa come metodo nella ricerca umanistica

In questo ambito di discorso, si deve partire dalla consapevolezza che i sistemi di IA generativa riorganizzano in modo profondo i workflow e le stesse pratiche epistemiche fondamentali della ricerca umanistica, digitale e no. La loro presenza, a differenza di metodi e approcci precedenti, è ubiqua e soprattutto permette di operazionalizzare (concetto qui da intendere in senso lato) e delegare a sistemi non umani ogni processo e aspetto della ricerca: la costruzione dei corpora e dei dataset, l'annotazione, l'esplorazione preliminare dei dati, la generazione di ipotesi interpretative e la loro verifica, la sintesi comparativa, persino la confezione finale del prodotto scientifico.
Gli LLM come approcci metodologici occupano una posizione non riducibile alla tradizionale dicotomia tra metodi quantitativi e qualitativi, tra close reading e distant reading, tra formalizzazione e libertà ermeneutica. Anzi, si potrebbe dire che l'uso degli LLM intercetta il punto più delicato della ricerca umanistica: la produzione di interpretazioni ed esplicazioni ermeneutiche, che non si limitano all'estrazione o classificazione dell'informazione, ma implicano giudizi di rilevanza, coerenza, plausibilità e valore. Si tratta ovviamente di saggiare fino a che punto questa intersezione sia efficace, di fronte alla complessità degli oggetti culturali e della loro natura intrinsecamente storica.
I recentissimi sviluppi dei modelli agentivi e di piattaforme come Claude Code e Open AI Codex spingono ulteriormente i limiti di quanto i modelli possono fare (Mollick, 2026). Se fino a pochi mesi fa si poteva al massimo parlare di *vibe coding*, la scrittura di codice eseguita da modelli linguistici (che comunque nel contesto della ricerca umanistica digitale/computazionale rappresenta un enorme supporto), oggi con tali agenti è possibile delegare un intero workflow di ricerca sperimentale a un modello agente, con livelli di autonomia sperimentale prima non immaginabili. Ovviamente si tratta di capire i limiti di questo approccio, la sua capacità di esplorare lo spazio delle possibilità teorico-esplicative, e il ruolo da affidare al controllo del/i ricercatore/i umano/i nel circolo ermeneutico: questioni non solo aperte ma anche soggette a una continua ridefinizione per via della continua innovazione tecnologica. Lo stesso si potrebbe dire per la capacità dei modelli di *deep research* o di strumenti verticali come Notebook LM di effettuare quel tipo di attività di ricerca che delimita lo stato dell'arte e identifica le questioni e le ipotesi già esplorate e verificate nella letteratura scientifica di un dato ambito di studi. Sono solo due esempi che mostrano come le potenzialità dei modelli linguistici siano considerevoli. Un ultimo corollario: credo di poter dire che la tradizione italiana delle Digital Humanities, per il suo forte orientamento metodologico, può qui svolgere una funzione critica essenziale al fine di progettare protocolli di uso controllato, definire criteri di validazione, costruire pratiche di cooperazione tra analisi automatica e interpretazione umana.

2.3 IA generativa come tecnologia socioculturale

Indipendentemente dalle teorie che si vogliano assumere circa la loro natura cognitiva, è chiaro che i grandi modelli linguistici e multimodali (su cui occorrerebbe una riflessione specifica che qui non posso affrontare), poiché sono addestrati su vastissimi insiemi di oggetti semiotici e culturali, a loro volta veicoli

di credenze, valori, norme, stereotipi e narrazioni, sono dei portentosi modelli dell'ambiente culturale in cui la nostra specie opera, che altri chiama infosfera o noosfera: io preferisco la nozione lotmaniana di *semiosfera* (Lotman, 1990, 1992). Essi funzionano come dispositivi di compressione e riorganizzazione della memoria culturale, ma al contempo seguono criteri di modellizzazione che riflettono asimmetrie linguistiche, assiologiche e culturali. In modo simmetrico, questi sistemi retroagiscono sui sistemi culturali. Inseriti nei circuiti della comunicazione sociale, modificano la velocità, la scala e la forma della produzione discorsiva. La generazione automatica di testi e immagini influisce sui processi di diffusione delle credenze, sulla stabilizzazione delle narrazioni condivise e, in ultima istanza, sui comportamenti collettivi. Infine, l'IA generativa riorganizza i workflow dell'industria culturale e della formazione: editoria, giornalismo, audiovisivi, istruzione. Cambiano le filiere produttive, le forme di *gatekeeping*, le nozioni di autorialità e responsabilità, il senso stesso di pratiche come la scrittura, la valutazione e l'apprendimento.
Questa prospettiva tecno-culturale sull'IA generativa è stata elaborata, da e verso direzioni teoriche diverse, da numerosi studiosi, che lo scienziato politico Henry Farrell (Farrell, 2025; Farrell et al., 2025) ha cercato di organizzare in modo sistematico in quattro correnti: Gopnikismo, Interazionismo, teoria del Role-play e Strutturalismo. La prima fa riferimento alle tesi della psicologa evolutiva Alison Gopnik (Yiu et al., 2023) che considera i LLM come motori di trasmissione culturale fondamentalmente disconnessi dalla realtà empirica. Nella sua cornice teorica, la cultura costituisce un sistema collettivo di conoscenza—comprendente formule, pattern, convenzioni, tropi e stereotipi—attraverso cui gli esseri umani compensano le limitazioni cognitive individuali e trasmettono informazioni attraverso le generazioni. I LLM, addestrati esclusivamente su dati testuali prodotti dall'attività culturale umana, operano interamente all'interno di questo dominio simbolico. Questa disconnessione fondamentale dall'esperienza percettiva e dalla causalità fisica genera un'asimmetria caratteristica: i LLM dimostrano notevole facilità nel riorganizzare, sintetizzare e trasmettere la conoscenza culturale esistente, eppure mancano della capacità di autentica innovazione o scoperta causale.
La prospettiva interazionista sposta il fuoco analitico dalle proprietà intrinseche dei sistemi LLM alla loro relazione con l'architettura cognitiva degli interpreti umani. Questo approccio, secondo Farrell, si pone in continuità con le teorie della cultura di Sperber e degli studi di *cultural evolution* (Sperber, 1996) secondo cui le conseguenze culturali dell'IA generativa dipendono da come le menti umane, plasmate da pressioni evolutive e dotate di moduli cognitivi specializzati, processano e rispondono agli output dei LLM. Questo permette di capire le dinamiche mediante le quali alcuni contenuti culturali conseguono adozione e persistenza diffuse, funzionando come attrattori culturali: concetti religiosi, teorie cospirazioniste e archetipi narrativi hanno successo perché fanno presa su moduli mentali evolutisi per altri scopi. In questo contesto i LLM, attraverso una sofisticata produzione umana, possono generare output che costituiscono nuovi attrattori culturali, producendo contenuti particolarmente "risonanti" con l'architettura psico-cognitiva umana.
La teoria del *role-play* si basa fondamentalmente sul lavoro dello scienziato cognitivo Murray Shanahan (Shanahan et al., 2023), che è stato di recente confermato sperimentalmente da alcuni lavori del team di Anthropic (Chen et al., 2025; Lu et al., 2026). Secondo questa prospettiva, un LLM costituisce una sovrapposizione di simulacri all'interno di un multiverso di personaggi possibili, una superposizione di tutte le personae, voci, situazioni e stereotipi culturali presenti nel corpus di addestramento. L'operazione fondamentale del sistema non è il ragionamento da credenze stabili o il perseguimento di obiettivi assunti, bensì la performance di ruoli finzionali tratti da questo vasto repertorio culturale.
Shanahan impiega l'analogia del gioco di ruolo "Venti domande" per esplicitare questo meccanismo: quando il gioco inizia, chi risponde non ha predeterminato quale oggetto difenderà; piuttosto, mantiene tutti gli oggetti possibili in sovrapposizione, restringendo progressivamente lo spazio delle possibilità man mano che le domande vincolano le opzioni praticabili. Analogamente, un LLM durante la conversazione non seleziona un'identità fissa all'inizializzazione, ma collassa dinamicamente la sua distribuzione sui personaggi possibili man mano che si accumulano indizi contestuali. Data la dominanza statistica di tropi convenzionali e scenari stereotipati nei dati di addestramento, questo processo converge frequentemente su risposte stereotipe o culturalmente sovradeterminate.
L'interpretazione strutturalista infine connette le architetture degli LLM alle teorie linguistiche elaborate da Saussure, Hjelmslev e poi dalla semiotica strutturalista novecentesca, che consideravano il linguaggio come sistema auto-contenuto le cui proprietà fondamentali derivano non dal riferimento esterno o dai contenuti psicologici del parlante, ma dalle relazioni interne tra segni. Tra i più rilevanti contributi in questa direzione c'è il libro di Leif Weatherby (Weatherby, 2025) che a partire dalla tradizione strutturalista (ma ci include anche Derrida, con eccesso di inclusività teoretica) sostiene che i LLM

operano una separazione del linguaggio dalla cognizione individuale, realizzando in forma computazionale la visione strutturalista della langue come distinta dalle parole, del sistema dall'uso, del fatto sociale dallo stato psicologico. Il linguaggio, da questo punto di vista, costituisce il linguaggio *in extenso*, come macchina, la cultura stessa concepita come sistema semiotico auto-organizzantesi che opera secondo la propria logica, indipendentemente dalle menti individuali. Di recente l'ottimo libro di Claudio Paolucci *Nati Cyborg* (Paolucci, 2025) si muove in un terreno concettuale simile, proponendo una teoria degli LLM come agenti semiotici non umani, sulla base di una teoria del linguaggio basata sulla linea di pensiero che da Peirce arriva a Eco, incrociata con la teoria della mente estesa di Clark. Su un piano più eminentemente metodologico, Ted Underwood (Underwood, 2021, 2025; Underwood et al., 2025) ritiene che i modelli linguistici consentano agli studiosi di mappare e manipolare lo spazio culturale, e di conseguenza identificare generi, tracciare l'evoluzione concettuale, stabilizzare punti di riferimento per forme artistiche o abitudini di pensiero. Da questo punto di vista, rileva come lo stesso 'problema' dei bias culturali vada piuttosto visto come opportunità per l'analisi culturale, poiché esso rivela le direttrici che orientano le mentalità, e questo rende scientificamente efficace anche l'adozione nella sperimentazione di modelli di piccola dimensione *open source*, svincolando la ricerca dalla necessità di adottare le piattaforme AI proprietarie delle grandi multinazionali.
Si muovono in questa direzione anche alcuni contributi teorici che io stesso ho pubblicato di recente (Ciotti, 2024, 2025), dove propongo che i modelli linguistici possano essere intesi come istanziazioni computazionali della nozione di semiosfera di Jurij Lotman. Significativamente, Lotman stesso in un articolo del 1979 intitolato "La cultura come intelletto collettivo e i problemi dell'intelligenza artificiale", scrisse: "Dobbiamo sottolineare che l'intelletto collettivo, come modello per l'intelletto artificiale, ha diversi vantaggi rispetto all'intelletto individuale. Poiché l'intelletto collettivo è un meccanismo creato dalla storia dell'umanità, è molto più esplicito, le sue procedure sono manifeste nel linguaggio della cultura e sono registrate in numerosi testi, a differenza dei linguaggi nascosti del cervello umano." (Lotman, 1979). Anche questa prospettiva ha rilevanti implicazioni metodologiche: se i LLM modellizzano la semiosfera, essi forniscono agli studiosi strumenti senza precedenti per analizzare struttura e dinamiche culturali. L'indagine umanistica tradizionale, vincolata dalle limitazioni soggettive degli interpreti individuali e dalla difficoltà di operazionalizzare concetti culturali, può ora avvalersi di modelli computazionali che esternalizzano e rendono manipolabili i processi stessi della significazione culturale.

3. UNA RIVOLUZIONE EPISTEMOLOGICA DA CAPIRE E GUIDARE: VERSO UN'AGENDA

L'IA generativa ci costringe a riconcettualizzare le nostre assunzioni fondamentali su cosa siano il significato, la cultura, la conoscenza, l'interpretazione, la metodologia. Le Digital Humanities per decenni si sono dovute misurare con la resistenza degli oggetti di studio dei saperi umanistici all'operazionalizzazione computazionale. Concetti come genere, stile, ideologia, struttura narrativa, variante, intertestualità, valore estetico, fatto storico, e infiniti altri che popolano gli studi umanistici resistono sia alla formalizzazione attraverso la rappresentazione simbolica classica o sistemi di regole esplicite, sia all'operazionalizzazione quantitativa, alla riduzione a misure e numeri da trattare con modelli statistico-matematici.
I sistemi di IA generativa impiegano paradigmi di rappresentazione ed elaborazione fondamentalmente diversi. Piuttosto che simboli espliciti manipolati secondo regole formali, o quantità misurabili, questi sistemi codificano linguaggio, conoscenza e cultura in rappresentazioni continue sub-simboliche, vettori ad alta dimensionalità di numeri reali privi di sottostruttura discreta. Le singole dimensioni non possiedono interpretabilità semantica inerente; il significato emerge olisticamente dalle relazioni geometriche attraverso l'intero spazio. Questo formato rappresentazionale si dimostra notevolmente capace di catturare fenomeni culturali refrattari alla formalizzazione.
Inoltre, i LLM esternalizzano processi interpretativi tradizionalmente confinati alla cognizione umana soggettiva. La comprensione ermeneutica, l'interpretazione del significato all'interno dei testi culturali, ha storicamente risieduto nelle menti dei singoli studiosi, accessibile solo attraverso l'introspezione e comunicabile solo attraverso il linguaggio descrittivo. I LLM esternalizzano questi processi, ma allo stesso tempo li rendono tanto poco perspicui quanto quelli umani.
L'emergere dell'IA generativa segna quindi non meramente uno sviluppo tecnologico ma una trasformazione epistemologica che richiede sofisticazione teorica e metodologica. Gli studiosi devono sviluppare cornici rigorose per comprendere cosa i sistemi intelligenti sono e non sono, come si relazionano alla cognizione culturale umana, quali tipi di conoscenza possono fornire e quali limitazioni ne

vincolano l'applicazione. Si tratta di un lavoro che deve procedere sia sul piano concettuale sia su quello empirico, attraverso studi sistematici su come le rappresentazioni dei LLM corrispondano alle concettualizzazioni culturali umane, dove divergano e quali bias o distorsioni introducano. Infine, dal punto di vista metodologico è necessario sviluppare *best practices* per impiegare i LLM nella ricerca mantenendo consapevolezza critica delle loro limitazioni e potenziali applicazioni erronee.
La tradizione teorica e culturale dell'Informatica Umanistica (Ciotti, 2018) di cui la nostra comunità scientifica è erede, ci pone nella condizione di assumere un ruolo guida in questa impresa, che è anche una ennesima (forse l'ultima) possibilità di riaffermare la necessità e la centralità sociale dei saperi umanistici. Già la storia passata ha mostrato come la ricerca umanistica, anche nei settori più esoterici (e a prima vista privi di impatto sociale e di sostenibilità economica, sulla base della razionalità economicista che oggi domina anche l'istruzione superiore) se si mette in gioco seriamente e punta all'innovazione metodologica e disciplinare può non solo invertire la narrativa del declino ma contribuire allo sviluppo dei nuovi ecosistemi digitali; perché i problemi veramente complessi con cui la ricerca sull'Intelligenza Artificiale si deve misurare sono spesso quelli di cui si occupano le discipline umanistiche, e addestrare modelli linguistici a misurarsi con l'eredità di lingue e scritture ormai abbandonate da secoli ha di gran lunga più valore ed efficacia generale che usarli per capire il sentiment dei commenti sui social media in relazione all'ultimo fatto di sangue. Dobbiamo prima di tutto convincerci noi umanisti di questo, però, per essere in grado di convincere il resto della società. Quando sulle terze pagine dei giornali, sui blog culturali, su saggi e articoli, su libri (accompagnati da soddisfatte recensioni) si leggono invettive emesse da illustri e meno illustri colleghi intellettuali che ci rassicurano che la AI è solo un sistema statistico che non potrà mai produrre nuova conoscenza, che non potrà mai contribuire a capire veramente niente che conti o genuinamente (nota a margine: dopo tanta esaltazione della deriva postmoderna dei significati questo ritorno sospetto alla venerazione del vero e del genuino mi pare molto sospetto), che anzi essa sia una minaccia alla sussistenza stessa della conoscenza e altre venerande istituzioni sociali e culturali, vedo invece una tendenza contraria, e mi domando come sia possibile non aver capito non solo che, come ci insegnò Hugo, "Ceci tuera cela", ma che stavolta "cela" siamo noi.
La posta in gioco si estende oltre la ricerca accademica per abbracciare implicazioni culturali e sociali più ampie. In primo luogo, nella sfera che non ho qui trattato, solo perché la sua rilevanza e difficoltà è al di là delle mie capacità di singolo studioso: quella della formazione. Ma più in generale, se è vero che i sistemi di IA generativa di oggi e quelli a venire operano sui sistemi culturali-semiotici, essi possiedono una capacità senza precedenti di plasmare il discorso sociale, influenzare la formazione delle credenze e riconfigurare i paesaggi culturali. Comprendere questi sistemi diventa essenziale non solo per la ricerca delle DH e quella umanistica in generale, ma per assumere un atteggiamento responsabile verso tecnologie che stanno trasformando la cultura e la società in cui viviamo. Per questo penso che, come comunità, sia nostra responsabilità predisporre un'agenda che sia anche la mappa per il nostro futuro prossimo come studiosi e come intellettuali socialmente responsabili.

BIBLIOGRAFIA

AI Manifesto. (2025). C2DH. https://www.uni.lu/c2dh-en/articles/ai-manifesto/

Barkeshli, M., Douglas, M. R., & Freedman, M. H. (2026). Artificial Intelligence and the Structure of Mathematics (Versione 1). arXiv. https://doi.org/10.48550/ARXIV.2604.06107

Buzzetti, D. (2019). The Origins of Humanities Computing and the Digital Humanities Turn (M. Lollini, Trad.). Humanist Studies & the Digital Age, 6(1), 32–58. https://doi.org/10.5399/uo/hsda.6.1.3

Chen, R., Arditi, A., Sleight, H., Evans, O., & Lindsey, J. (2025). Persona Vectors: Monitoring and Controlling Character Traits in Language Models (Versione 3). arXiv. https://doi.org/10.48550/ARXIV.2507.21509

Ciotti, F. (2018). From Informatica Umanistica to Digital Humanities and return: A conceptual history of Italian DH. TESTO & SENSO.

Ciotti, F. (2023a). Introduzione. La galassia delle Digital Humanities. In Digital Humanities. Metodi, strumenti, saperi. Carocci.

Ciotti, F. (2023b). Minerva e il pappagallo. IA generativa e modelli linguistici nel laboratorio dell'umanista digitale. TESTO & SENSO, (26), 289–315. https://doi.org/10.58015/2036-2293/671

Ciotti, F. (2024). Gli LLM come lettori modello artificiali. In Me.Te. Digitali. Mediterraneo in rete tra testi e contesti. Proceedings del XIII Convegno Annuale AIUCD2024 (pp. 342–344). AIUCD. https://doi.org/10.6092/unibo/amsacta/7927

Ciotti, F. (2025). Lettori artificiali e macchine pigre: Cosa i Large Language Model possono farci capire sulla cooperazione interpretativa. In Letteratura e intelligenza artificiale. Un dialogo interdisciplinare (pp. 109–146). Lithos.

De Caro, M., & Giovanola, B. (2025). Intelligenze. Etica e politica dell'IA. Il Mulino. https://books.google.it/books?id=BQKb0QEACAAJ

Farrell, H. (2025, gennaio 7). Large language models are cultural technologies. What might that mean? Programmable Mutter. https://www.programmablemutter.com/p/large-language-models-are-cultural

Farrell, H., Gopnik, A., Shalizi, C., & Evans, J. (2025). Large AI models are cultural and social technologies. Science, 387(6739), 1153–1156. https://doi.org/10.1126/science.adt9819

Hemment, D., & Kommers, C. (2025). Doing AI Differently: Rethinking the Foundations of AI via the Humanities. https://doi.org/10.5281/ZENODO.16421295

Horvath, A., Bonch-Osmolovskaya, A., Goodrich, A., Lombana-Bermudez, A., Gurumurthy, A., Orekhov, B., Clark, C., Chiesa, C. D., Leterme, C., Bharthur, D., Ávila, D. B., Roy, D., Trigueros, E. M., Priego, E., Del Rio Riande, G., Kim, I., Kizhner, I., Marienberg-Milikowsky, I., Chen, J., … Unwin, T. (2022). Global Debates in the Digital Humanities (D. Fiormonte, S. Chaudhari, & P. Ricaurte, A c. Di; Vol. 8). University of Minnesota Press. https://doi.org/10.5749/9781452968919

Klein, L., Martin, M., Brock, A., Antoniak, M., Walsh, M., Johnson, J. M., Tilton, L., & Mimno, D. (2025). Provocations from the Humanities for Generative AI Research (Versione 2). arXiv. https://doi.org/10.48550/ARXIV.2502.19190

Klowden, T., & Tao, T. (2026). Mathematical methods and human thought in the age of AI (Versione 1). arXiv. https://doi.org/10.48550/ARXIV.2603.26524

Kommers, C., Ahnert, R., Antoniak, M., Benetos, E., Benford, S., Bunz, M., Caramiaux, B., Concannon, S., Disley, M., Dobson, J., Du, Y., Duéñez-Guzmán, E., Francksen, K., Gius, E., Gray, J., Heuser, R., Immel, S., So, R., Leigh, S., … Hemment, D. (2025). Computational Hermeneutics: Evaluating Generative AI as a Cultural Technology. SSRN. https://doi.org/10.2139/ssrn.5409144

Lotman, J. M. (1979). Culture as collective intellect and the problems of artificial intelligence. In Russian Poetics in translation (Vol. 6, pp. 84–96).

Lotman, J. M. (1990). Universe of the Mind: A Semiotic Theory of Culture. Indiana University Press.

Lotman, J. M. (1992). La semiosfera: L'asimmetria e il dialogo nelle strutture pensanti (S. Salvestroni, A c. Di; 2a ed). Marsilio.

Lu, C., Gallagher, J., Michala, J., Fish, K., & Lindsey, J. (2026). The Assistant Axis: Situating and Stabilizing the Default Persona of Language Models (Versione 1). arXiv. https://doi.org/10.48550/ARXIV.2601.10387

Mollick, E. (2026, febbraio 18). Claude Code and What Comes Next. One Useful Thing. https://www.oneusefulthing.org/p/claude-code-and-what-comes-next

Paolucci, C. (2025). Nati cyborg. Cosa l'intelligenza artificiale generativa ci dice dell'essere umano. Luca Sossella Editore.

Shanahan, M., McDonell, K., & Reynolds, L. (2023). Role play with large language models. Nature, 1–6.

Sperber, D. (1996). Explaining Culture: A Naturalistic Approach. Oxford: Basil Blackwell.

Underwood, T. (2021, ottobre 21). Mapping the latent spaces of culture. The Stone and the Shell. https://tedunderwood.com/2021/10/21/latent-spaces-of-culture/

Underwood, T. (2025). The impact of language models on the humanities and vice versa. Nature Computational Science, 5(9), 695–697. https://doi.org/10.1038/s43588-025-00819-4

Underwood, T., Nelson, L. K., & Wilkens, M. (2025). Can Language Models Represent the Past without Anachronism? (Versione 1). arXiv. https://doi.org/10.48550/ARXIV.2505.00030

Weatherby, L. (2025). Language machines: Cultural AI and the end of remainder humanism. University of Minnesota Press.

Yiu, E., Kosoy, E., & Gopnik, A. (2023). Transmission versus truth, imitation versus innovation: What children can do that large language and language-and-vision models cannot (yet). Perspectives on Psychological Science, 17456916231201401.

L'errore dell'IA come euristica per una deliberazione assistita

VALERIO DE LUCE[1], TIBERIO URICCHIO[2]

[1]UNIVERSITÀ DI MACERATA - V.DELUCE@UNIMC.IT, ORCID: 0009-0006-6190-6406
[2]UNIVERSITÀ DI PISA - TIBERIO.URICCHIO@UNIPI.IT, ORCID: 0000-0003-1025-4541

ABSTRACT (ITALIANO)

L'impiego di Large Vision Language Models (LVLM) per generare automaticamente metadati coerenti con le architetture formali dei protocolli archivistici promette di accelerare le tempistiche descrittive. Ma cosa accade quando questi sistemi si confrontano con materiali intrinsecamente legati a un contesto storico e culturale specifico? Abbiamo applicato un framework agentico basato su più LVLM all'archivio "Satira e Sogni" del fumettista Sergio Staino, valutando la capacità dei modelli di estrarre informazioni e produrre metadati da 300 immagini digitalizzate. I modelli riconoscono i personaggi ricorrenti dell'autore e riproducono i dati strutturati espliciti; tuttavia, falliscono quasi completamente nell'identificare figure politiche italiane, nel cogliere riferimenti satirici indiretti e nel modellare le relazioni tra entità e contesto documentale. Proponiamo di interpretare questi errori come sonde epistemiche, ovvero fallimenti sistematici che rivelano, per contrasto, le competenze tacite mobilitate dalla professione archivistica. Da questa analisi emerge la "deliberazione assistita", un approccio alla collaborazione tra uomo e macchina in cui l'IA non sostituisce l'archivista ma diventa interlocutore critico, aprendo spazi di riflessione condivisa sui criteri descrittivi.
Parole chiave: LVLM; descrizione archivistica; sonda epistemica; conoscenza tacita; archivi di fumettisti

ABSTRACT (ENGLISH)

The error of AI as a heuristic for assisted deliberation.
The use of Large Vision Language Models (LVLM) to automatically generate metadata consistent with the formal architectures of archival protocols promises to speed up descriptive processes. But what happens when these systems are confronted with materials intrinsically linked to a specific historical and cultural context? We applied an agentic framework based on multiple LVLMs to the "Satira e Sogni" archive of cartoonist Sergio Staino, evaluating the system's ability to extract information and produce metadata from 300 digitised images. The system recognises the author's recurring characters and reproduces explicit structured data; however, it fails almost completely to identify Italian political figures, to grasp indirect satirical references, and to model relationships between entities and documentary context. We propose to interpret these errors as epistemic probes, i.e., systematic failures that reveal, by contrast, the tacit skills mobilised by the archival profession. This analysis gives rise to "assisted deliberation", an approach to human-machine interaction in which artificial intelligence does not replace the archivist but becomes a critical interlocutor, opening spaces for shared reflection on descriptive criteria.
Keywords: LVLM; archival description; epistemic probe; tacit knowledge; cartoonists' archives

1. INTRODUZIONE

L'applicazione dell'intelligenza artificiale (IA) alla descrizione archivistica rappresenta un campo emergente che promette di accelerare i flussi di lavoro catalografici, sollevando però interrogativi sulla natura stessa della professione. Se da un lato i Large Language Models (LLM) e i LVLM offrono capacità inedite di elaborazione testuale e visiva, dall'altro il loro impiego su materiali intrinsecamente legati a contesti storici e culturali specifici tende a produrre fallimenti sistematici che meritano di essere analizzati.
Gli archivi di fumettisti rappresentano un caso particolarmente critico per l'IA. A differenza degli archivi tradizionali, essi documentano processi creativi piuttosto che burocratici, e il loro significato emerge dall'interazione tra immagine, testo e layout. I materiali che contengono sono eterogenei e generalmente caratterizzati da elevata complessità. Nel caso specifico del fumetto satirico, la densità referenziale richiede competenze interpretative che eccedono l'analisi puramente visiva (Eco, 1964).

La letteratura recente ha documentato come l'automazione possa supportare fasi specifiche del lavoro archivistico, dall'elaborazione (Hutchinson, 2020) all'estrazione di metadati (Büttner, 2019), mantenendo tuttavia centrale l'expertise professionale, un orientamento coerente con il concetto di Augmented Humanities proposto da Bryant et al. (2021), che enfatizza la collaborazione tra agenti umani e sistemi digitali. Gli esperimenti condotti finora si sono però concentrati prevalentemente su materiali testuali e omogenei, lasciando ancora largamente inesplorato il territorio degli archivi a prevalente componente grafica (Allegrezza, 2025).
Questo contributo analizza i risultati di un esperimento condotto sull'archivio "Satira e Sogni" del fumettista Sergio Staino, in cui un framework agentico basato su LVLM è stato impiegato per generare automaticamente metadati da 300 immagini digitalizzate (De Luce & Uricchio, 2026)[1]. A partire dai fallimenti emersi, ne sviluppiamo un'interpretazione epistemologica e proponiamo la deliberazione assistita come approccio alla collaborazione uomo-macchina.
L'ipotesi che guida questo lavoro è che gli errori possano essere impiegati come sonde epistemiche. Distinguiamo in questa sede la sonda, strumento euristico impiegato dall'analista, dalla soglia, ovvero il confine di competenza che la sonda rende visibile. Con l'espressione sonda epistemica intendiamo fallimenti dei modelli che soddisfano tre condizioni: 1) sono sistematici, ricorrendo in più documenti del campione con pattern riconoscibili; 2) sono categorizzabili, riconducibili a tipologie distinte; 3) sono contrastivi, rivelando per negazione una competenza necessaria alla descrizione.
A differenza della tradizionale analisi degli errori, orientata al miglioramento dei modelli, le sonde epistemiche mirano a rendere esplicite le competenze umane che il framework presuppone ma non possiede. Iterare il processo, analizzare i fallimenti e correggere gli output stimola una riflessione metacognitiva che, nella prospettiva della reflective practice (Schön, 2017), fa emergere competenze altrimenti tacite (Polanyi & Sen, 2009). Da questa analisi deriviamo la "deliberazione assistita", una forma di collaborazione uomo-macchina in cui l'errore dell'IA diventa dispositivo euristico per interrogare la professione[2].
L'obiettivo è duplice: da un lato categorizzare i fallimenti dei modelli per tipologia, dall'altro ricavarne implicazioni per la pratica e la formazione professionale attraverso un caso di studio che sottopone il framework a un banco di prova critico.

2. STATO DELL'ARTE

Questa ricerca si colloca all'intersezione di tre ambiti: l'impiego di sistemi agentici basati su LVLM, l'applicazione dell'IA nei processi descrittivi della documentazione e l'interpretazione epistemologica dell'errore algoritmico.
I sistemi agentici rappresentano un'evoluzione recente nell'impiego degli LLM, in cui più modelli collaborano per svolgere compiti complessi attraverso una divisione specializzata delle funzioni (Abou Ali et al., 2026). Un esempio affine al nostro caso è costituito dal framework di Groppe et al. (2025) per la generazione federata di metadati secondo lo standard ISAD(G), che tuttavia si limita a materiali testuali. Nel nostro lavoro estendiamo questa logica impiegando un framework agentico composto da più LVLM per elaborare materiali a prevalente componente visiva, secondo uno schema descrittivo derivato dal protocollo impiegato per la rappresentazione dei metadati presenti sul sito di Satira e Sogni (De Luce & Uricchio, 2026). La sfida principale per questi sistemi emerge quando i materiali da elaborare richiedono competenze interpretative specifiche del dominio.
Sul versante archivistico, l'integrazione dell'IA nelle pratiche documentali è oggetto di crescente attenzione. Progetti come InterPARES Trust AI ne indagano le implicazioni (Feliciati et al., 2021), mentre una panoramica sistematica del rapporto tra archivi e intelligenza artificiale è offerta da studi recenti che mappano le principali aree di applicazione e le questioni aperte, evidenziando come le soluzioni disponibili supportino ancora solo in misura limitata il lavoro dell'archivista (Allegrezza, 2025). Gli LLM si sono dimostrati efficaci nella generazione di parole chiave e nell'arricchimento dei metadati, soprattutto in ambito museale (Reusens et al., 2025). Pur trattandosi di un contesto differente, lo riteniamo un ambiente affine da cui trarre ispirazione; al contempo, consideriamo essenziale che l'automazione

[1] In De Luce & Uricchio (2026) si presentano il framework tecnico, la configurazione del framework e la valutazione quantitativa dei risultati.
[2] Il termine "deliberazione assistita" è qui impiegato in accezione diversa da quella prevalente nella letteratura sulla "democrazia deliberativa", dove indica processi di mediazione AI tra gruppi umani per raggiungere consenso (cfr. Tessler et al., 2024). Nel nostro contesto, la deliberazione scaturisce dal fatto che il confronto con l'output dei modelli attiva un processo riflessivo individuale sui propri criteri descrittivi.

acceleri i flussi di lavoro mantenendo centrale l'expertise umana, in coerenza con il paradigma dell'Augmented Humanities (Bryant et al., 2021). Meno indagato è ciò che accade quando l'automazione incontra i propri limiti strutturali e produce errori sistematici.
Un nodo critico riguarda lo statuto epistemologico dell'errore nei flussi prodotti dall'algoritmo. Gli agenti basati su LLM e LVLM tendono a generare un'illusione di completezza: anche in assenza di dati affidabili, costruiscono una coerenza logica che può occultare la natura mediata delle fonti (Finn & Khosrowi, 2025). Tuttavia, l'errore che ne deriva non è necessariamente privo di valore conoscitivo. Come osserva Ge (2025), le allucinazioni possono talvolta rivelarsi soluzioni creative inattese o, nel nostro caso, indicatori di limiti strutturali. Non tutte le allucinazioni sono epistemicamente produttive: i criteri che introduciamo in §1 servono precisamente a distinguere gli errori informativi dal rumore accidentale. È su questo terreno che l'errore dell'IA può diventare produttivo: laddove i modelli falliscono, l'archivista è chiamato a esplicitare ciò che normalmente rimarrebbe implicito nella propria pratica professionale (Schön, 2017).

3. CASO DI STUDIO

Gli archivi di fumettisti non rispondono alle definizioni ortodosse di archivio. Potremmo definirli "archivi inventati" (Valacchi, 2023), accostarli agli archivi d'artista (Brunetti, 2022) o alla realtà multiforme degli archivi privati di persona (Allegrezza, 2018). Si tratta di un tema ancora marginale nel dibattito accademico: non esistono prassi consolidate per collegare uno specifico protocollo descrittivo alla tipologia di oggetti che questi archivi possono contenere. Ciò vale anche per altri contesti come quello puramente bibliografico o biblioteconomico (Sferruzza, 2023). Nonostante la presenza di strumenti di riferimento,[3] è assente una formalizzazione delle pratiche per questa tipologia di risorse.
I documenti prodotti dai fumettisti sono estremamente eterogenei. Schizzi, bozzetti e illustrazioni sono opere che spesso presentano serialità narrativa, personaggi ricorrenti e, nel caso della satira, riferimenti politici. La stessa letteratura in ambito Computer Vision riconosce la comprensione dei fumetti come un problema ancora largamente aperto (Vivoli et al., 2024). Abbiamo scelto questo caso di studio proprio perché presenta caratteristiche che lo pongono al limite per l'analisi computazionale.
L'archivio di Sergio Staino raccoglie la produzione del fumettista dalla fine degli anni Settanta. Comprende 15.695 opere[4] tra disegni, vignette, quaderni, note, agende, strisce e bozzetti, digitalizzati nell'ambito di un progetto sostenuto dalla Fondazione CR Firenze, dalla Regione Toscana e dal Comune di Scandicci. L'archivio presenta eterogeneità dei materiali, incompletezza delle serie produttive, serialità narrativa e ha come tema principale la satira politica con riferimenti a figure della società italiana.
Per la sperimentazione sono state selezionate 300 immagini digitalizzate dal "Fondo archivio cartaceo", distribuite su diverse tipologie documentarie. La selezione è stata effettuata mediante campionamento stratificato per rispecchiare la distribuzione tipologica del fondo, con prevalenza di schizzi e presenza di vignette, strisce e altri materiali.
Dall'archivio, che contiene oltre 4.000 entità distinte nelle sue associazioni documentate è stato estratto un vocabolario controllato di 27 entità, selezionate tra le più frequenti nel campione e per la loro rilevanza nell'universo iconografico dell'autore: 14 persone (tra cui i protagonisti fittizi Bobo e Bibi e i figli dell'autore Michele e Ilaria), 7 organizzazioni, 3 luoghi, 2 oggetti e 1 evento. Il ground truth è stato definito sulla base dei metadati presenti sul sito, prodotti dai professionisti responsabili della catalogazione originale. Le associazioni entità-documento sono state classificate in tre tipologie: depicted (entità visivamente rappresentata, 30%), mentioned (nome esplicitato nel testo, 28%) e referred (riferimento indiretto, 42%).

4. METODO

L'analisi si basa sui risultati di una sperimentazione condotta nell'ambito del Cartoonists Meta-description Framework (CMF), un framework agentico basato su LVLM open-source per la generazione automatica di metadati (De Luce & Uricchio, 2026). Il framework vede impiegato un orchestratore e quattro agenti (descrizione, estrazione entità, classificazione, validazione) basati su LVLM open-source, testati in due configurazioni: una multi-agente con modelli di grande scala e una con modelli più piccoli. Entrambe

[3] Si fa riferimento a Records in Contexts (RiC; ICA EGAD, 2023), standard di riferimento del settore che favorisce l'interoperabilità e supera l'impostazione gerarchica dei modelli descrittivi tradizionali. Mancano tuttavia linee guida operative specifiche per gli archivi di fumettisti.
[4] Dati ricavati dal portale Satira e Sogni (https://www.sergiostaino.it/satira-e-sogni/). Ultimo accesso: 4 febbraio 2026.

operano in modalità zero-shot, senza fine-tuning né Retrieval-Augmented Generation (RAG). I modelli non dispongono di grounding storico-culturale: ricevono esclusivamente l'immagine, un prompt strutturato in italiano e il vocabolario controllato. Quest'ultimo ha natura sintattico-strutturale, essendo derivato dalle associazioni entità-documento presenti nei metadati del portale dell'archivio. Questa configurazione è intenzionalmente minimale, adottata per operare con risorse computazionali limitate. L'impiego esclusivo di modelli on-premises è stato adottato per garantire la riservatezza sul materiale.
I pattern di errore qui discussi sono emersi dal confronto sistematico tra gli output e i metadati di riferimento, annotando le tipologie di fallimento ricorrenti e raggruppandole per affinità. I pattern rappresentano una categorizzazione interpretativa che orienta la riflessione epistemologica del presente contributo. La valutazione è stata integrata da un giudizio umano esperto.
L'obiettivo non è produrre risultati generalizzabili, ma proporre un approccio che consiste nell'impiego degli errori come sonde epistemiche. I tre criteri introdotti in §1 (sistematicità, categorizzabilità, contrastività) sono stati applicati come filtro operativo per distinguere errori informativi da rumore accidentale: il primo verifica la ricorrenza sistematica, il secondo consente la traduzione in conoscenza operativa, il terzo lega il fallimento alla competenza umana che è presupposta ma assente.

5. RISULTATI

Le descrizioni generate risultano complessivamente accettabili sul piano terminologico, nel senso che impiegano lessico pertinente al dominio, mentre l'identificazione delle entità è fortemente compromessa. Il divario più rilevante emerge nelle relazioni *referred*, ovvero i riferimenti indiretti che richiedono conoscenza culturale e storica. Su questa tipologia i modelli producono un tasso di fallimento molto elevato, e il loro mancato riconoscimento spiega gran parte della bassa performance complessiva.
I modelli hanno riconosciuto in modo affidabile solo 4 delle 27 entità del vocabolario controllato, i personaggi ricorrenti di Staino (Bobo, Bibi, Michele, Ilaria), figure visivamente distintive. Le figure politiche italiane risultano sistematicamente non riconosciute, venendo descritte in termini generici (es. Craxi descritto come "uomo anziano con barba e occhiali") senza identificazione, evidenziando, almeno con le configurazioni testate, l'incapacità di riconoscere caricature politiche senza contesto culturale.

Pattern	Categoria	Fallimento	Frequenza	Evidenza
A	Iconografico e contestuale	Descrizione percettiva senza identificazione; assenza di inferenza	oltre 50%	Spadolini: "personaggi stilizzati"; Craxi: "uomo anziano con barba e occhiali"
B	Pragmatico-satirico	Mancato riconoscimento dell'ironia	circa 60%	"I nostri carri armati possono aiutarvi", descritto come "suggerisce che i carri armati dell'URSS possano aiutare" senza cogliere il sarcasmo
C	Relazionale	Incapacità di modellare serialità e contesto di produzione	oltre 50%	"probabilmente Bobo e Bibi" nella descrizione ma nessuna entità estratta; varianti grafiche non riconosciute

Tabella 1 Pattern di errore identificati suddivisi in categoria, tipo di fallimento, frequenza e un esempio di evidenza

Per escludere che i pattern fossero aneddotici abbiamo applicato ai 300 output criteri operativi di conferma (presenza di descrizioni generiche, assenza di marker interpretativi, menzione di personaggi ricorrenti senza estrazione come entità); per le metriche quantitative nel dettaglio si rimanda a De Luce & Uricchio (2026). I pattern si sovrappongono poiché una stessa immagine può presentare più tipologie di fallimento.

Nel complesso, il framework produce output coerenti in presenza di segnali espliciti ma fallisce quando la descrizione richiede interpretazione culturale, comprensione pragmatica e modellazione relazionale.

6. DISCUSSIONE

Per illustrare il meccanismo proposto, consideriamo un caso concreto seguendo le quattro fasi della deliberazione assistita. Fase 1: l'archivista esamina l'output per una vignetta e legge "personaggi stilizzati". Fase 2: identifica la discrepanza, riconoscendo Spadolini. Fase 3: classifica l'errore applicando i tre criteri: è sistematico (le figure politiche non vengono identificate in nessun documento del campione), categorizzabile (pattern A: descrizione percettiva senza identificazione) e contrastivo (presuppone la capacità di riconoscere figure politiche italiane in forma caricaturale). L'errore non è rumore, ma sonda epistemica. Fase 4: l'archivista esplicita il criterio descrittivo implicito, la familiarità con i repertori iconografici della politica italiana, e questa consapevolezza può informare l'arricchimento del vocabolario controllato o le linee guida per operatori in formazione. A differenza del paradigma human-in-the-loop, orientato alla correzione iterativa per migliorare il sistema (Monarch, 2021), la deliberazione assistita è orientata alla comprensione: l'obiettivo non è perfezionare il modello, ma interrogare la professione.
La Tabella 2 sintetizza l'applicazione dei tre criteri a ciascun pattern, verificando che ogni errore identificato soddisfi le condizioni per qualificarsi come sonda epistemica e non come rumore.

Criterio	**Pattern A (iconografico)**	**Pattern B (pragmatico-satirico)**	**Pattern C (relazionale)**
Sistematicità	Figure politiche non identificate in nessun documento	Contenuti satirici descritti letteralmente in circa il 60% degli output	Entità menzionate ma non estratte in oltre la metà del corpus
Categorizzabilità	Descrizione percettiva senza identificazione	Trascrizione letterale senza interpretazione pragmatica	Menzione senza modellazione relazionale
Contrastività	Presuppone il riconoscimento di caricature politiche	Presuppone la capacità di cogliere ironia e allusione	Presuppone memoria autoriale e contestualizzazione
Esempio	Spadolini: "personaggi stilizzati"	"I carri armati possono aiutarvi": descritto senza cogliere il sarcasmo	"probabilmente Bobo e Bibi": nessuna entità estratta

Tabella 2: Verifica dei tre criteri della sonda epistemica per ciascun pattern di errore

Il meccanismo si estende ai pattern B) e C): il pragmatico-satirico rivela la competenza nel cogliere allusione, iperbole e rovesciamento semantico (un'ia che descrive letteralmente una vignetta perde il commento critico sull'elusività della classe dirigente); il relazionale riguarda la memoria autoriale e la collocazione del documento nella produzione complessiva di Staino.
I tre pattern si allineano a tradizioni disciplinari consolidate: l'iconografia nella storia dell'arte, in particolare i livelli di analisi descritti da Panofsky (1955), con le questioni aperte legate alla distanza culturale tra dati di addestramento e oggetti analizzati (Wevers & Smits, 2020), la pragmatica del genere nella semiotica, la contestualizzazione nella teoria archivistica. Il contributo non è la scoperta di tali competenze, già note, né la pretesa di dimostrarne l'esistenza attraverso i fallimenti, ma il meccanismo attraverso cui l'errore algoritmico le rende simultaneamente visibili in un unico contesto operativo, esplicitando per contrasto saperi altrimenti taciti e dispersi in tradizioni separate, e la proposta di trasformare questa analisi in strumento per la formazione e la riflessione professionale. Questo approccio può risultare utile soprattutto nella fase esplorativa iniziale di progetti di catalogazione particolarmente sfidanti.

7. SONDE EPISTEMICHE E CONFIGURAZIONE

È opportuno qualificare con precisione lo statuto delle sonde epistemiche impiegate in relazione alle condizioni operative dell'esperimento. Come esplicitato in §4, i modelli hanno operato in modalità zero-shot, senza grounding storico-culturale. Il vocabolario controllato è una struttura sintattica priva di profondità semantica: nella terminologia di Floridi (2010), si tratta di un'organizzazione dell'informazione che non implica comprensione da parte della macchina. Di conseguenza, le sonde epistemiche qui identificate sono dipendenti dalla configurazione: in un framework dotato di RAG con repertori iconografici della politica italiana o profili biografici produrrebbe verosimilmente soglie diverse, spostando

il confine su altre competenze tacite. Ciò non invalida i pattern osservati, ma li qualifica come relativi a una specifica configurazione operativa. Ad esempio, modelli con accesso a repertori iconografici della politica italiana supererebbero verosimilmente il pattern A (identificazione delle caricature), ma difficilmente il pattern B (comprensione del sarcasmo e del rovesciamento semantico), che richiede competenze pragmatiche non riducibili a informazione contestuale. Diverse configurazioni rivelerebbero dunque diverse competenze, rendendo le sonde uno strumento diagnostico iterativo, non un risultato fisso. Questa caratteristica è coerente con la letteratura recente sul bias culturale nei LVLM, che dimostra come il contesto culturale fornito ai modelli modifichi sistematicamente i suoi output (Howard et al., 2026).
Si pone inoltre la questione se la soglia osservata si collochi tra macchina e umano o, più precisamente, tra descrizione e interpretazione. Concordiamo con questa lettura, e riteniamo che essa rafforzi piuttosto che indebolire la tesi centrale. La pratica archivistica, anche quando si definisce "descrittiva", incorpora strati interpretativi che i professionisti mobilitano tacitamente: ciò che Yakel (2003) ha proposto di riconoscere sotto il termine "archival representation", catturando la dimensione inevitabilmente interpretativa dell'ordinamento e della descrizione. Le sonde epistemiche rendono visibile questa incorporazione. Con la deliberazione assistita si creano le condizioni per esplicitare *dove* e *come* il professionista situa la propria interpretazione, trasformando un sapere tacito in un oggetto di riflessione condivisa. In questo senso, la deliberazione non avviene tra due entità stabili, ossia tra una macchina che sbaglia e un umano che detiene la verità, ma è un processo generativo: l'errore della macchina non viene confrontato con una "verità" umana fissa, bensì attiva una riflessione che produce consapevolezza nuova, ridefinendo i termini stessi del confronto.
Il contributo adotta deliberatamente la prospettiva archivistica perché il compito assegnato ai modelli è la generazione di metadati descrittivi, non l'analisi storica. Tuttavia, i fallimenti osservati, in particolare nel pattern B, attraversano il confine disciplinare, rivelando competenze che appartengono tanto all'archivista quanto allo storico. La descrizione archivistica di fondi satirici incorpora necessariamente atti interpretativi (datare, contestualizzare, identificare referenti), cosicché la prospettiva archivistica non esclude quella storiografica, ma rivela dove le due convivono. L'esempio del politico "irraggiungibile al telefono" illustra questo intreccio: non è riducibile a una descrizione archivistica in senso stretto, ma neppure a un'interpretazione puramente storiografica. Riconoscere questa bidimensionalità a nostro avviso conferma che la deliberazione assistita potrebbe estendersi a una riflessione fortemente interdisciplinare.

8. CONCLUSIONI

Questo contributo ha analizzato i pattern di fallimento di un framework agentico basato su LVLM applicato all'archivio "Satira e Sogni" e ne ha interpretato il significato epistemologico.
I tre pattern identificati (iconografico, pragmatico-satirico, relazionale) sono stati interpretati come sonde epistemiche che rivelano le competenze mobilitate nella descrizione di questo specifico archivio, dunque istanziazioni concrete di categorie più generali (come iconografia, pragmatica del genere, contestualizzazione), che in altri archivi si declinerebbero diversamente. I fallimenti riguardano: familiarità con repertori iconografici della politica italiana, comprensione dei meccanismi satirici, capacità di collocare il documento nella produzione dell'autore.
Sviluppi futuri intendono replicare la sperimentazione su scala più ampia, verificare l'applicabilità ad archivi di diversa natura e soggetti produttori. Intendiamo inoltre condurre studi con professionisti di settore per osservare i processi riflessivi attivati dal confronto con gli output del framework.

9. RINGRAZIAMENTI

Lavoro parzialmente sostenuto nell'ambito del progetto FoReLab (Dipartimenti di Eccellenza) del Ministero dell'Istruzione e della Ricerca (MUR). Ringraziamo NVIDIA Corporation per il sistema DGX Spark donato attraverso il programma NVIDIA Academic Grant.

BIBLIOGRAFIA

Abou Ali, M., Dornaika, F., & Charafeddine, J. (2026). Agentic AI: A comprehensive survey of architectures, applications, and future directions. *Artificial Intelligence Review*, 59(1), 11.

Allegrezza, S. (2018). L'impatto del digitale negli archivi privati. *ATLANTI*, 28, 79-90.

Allegrezza, S. (2025). L'intelligenza artificiale al servizio degli archivi digitali. *Umanistica Digitale*, (20), 643-660.

Brunetti, D. (2022). Artists' archives. *JLIS.it*, 13(2), 174-188.

Bryant, P. T. (2021). *Augmented Humanity: Being and Remaining Agentic in a Digitalized World* (p. 308). Springer Nature.

Büttner, G. (2019). Auto-classification in an international organization. *Comma*, 2017, 15-26.

De Luce, V., & Uricchio, T. (2026). The Cartoonists Meta-description Framework. In Proceedings of the 22nd Italian Research Conference on Digital Libraries (IRCDL 2026), Modena, 19-20 febbraio. CEUR Workshop Proceedings (in corso di pubblicazione).

Eco, U. (1964). *Apocalittici e integrati*. Bompiani.

Feliciati, P., Frontoni, E., & Paolanti, M. (2021). InterPARES Trust Artificial Intelligence. https://interparestrustai.org/

Finn, F., & Khosrowi, D. (2025). AI assistants in the archive and the lure of "instant history". *Cambridge Forum on AI: Culture and Society*.

Floridi, L. (2010). *Information: A Very Short Introduction*. Oxford University Press.

Ge, L. (2025). Spectral imaginings and sympoietic creativity. *Big Data & Society*, 12(4). https://doi.org/10.1177/20539517251400696

Groppe, J. et al. (2025). Automated Archival Descriptions with Federated Intelligence. *DEXA 2025*, 53-67.

Howard, P., Su, X., & Fraser, K. C. (2026). Cultural Counterfactuals: Evaluating Cultural Biases in Large Vision-Language Models with Counterfactual Examples. *arXiv preprint arXiv:2603.02370*.

Hutchinson, T. (2020). Natural language processing and machine learning as practical toolsets for archival processing. *Records Management Journal*, 30(2), 155-174.

ICA EGAD. (2023). *Records in Contexts - Conceptual Model v0.2*.

Monarch, R. M. (2021). *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Manning Publications.

Panofsky, E. (1955). Iconography and Iconology: An Introduction to the Study of Renaissance Art. In *Meaning in the Visual Arts*. Doubleday.

Polanyi, M. & Sen, A. (2009). *The Tacit Dimension*. University of Chicago Press.

Reusens, M., Adams, A., & Baesens, B. (2025). Large Language Models to make museum archive collections more accessible. *AI & Society*, 40(6), 4485-4497.

Schön, D. A. (2017). *The Reflective Practitioner: How Professionals Think in Action*. Routledge.

Sferruzza, M. (2023). I fumetti in biblioteca. *AIB studi*, 63(2), 297-311.

Tessler, M. H. et al. (2024). AI can help humans find common ground in democratic deliberation. *Science*, 386(6719), eadq2852.

Valacchi, F. (2023). Se l'archivio è artificiale: Varso uno ius archivi partecipativo?. *AIDAinformazioni*, (1-2), 153-170.

Vivoli, E. et al. (2024). One missing piece in Vision and Language: A survey on comics understanding. arXiv preprint arXiv:2409.09502.

Wevers, M., & Smits, T. (2020). The visual digital turn: Using neurat network to study historical images. *Digital Scholarship in the Humanities*, 35(1), 194-207.

Yakel, E. (2003). Archival Representation. *Archival Science*, 3, 1-25.

Verso una memoria artificiale? Il progetto Johannes tra questioni tecnologiche, archeologiche e morali

STEFANO BERTOLDI
UNIVERSITÀ DEGLI STUDI DI SIENA - STEFANO.BERTOLDI@UNISI.IT

ABSTRACT (ITALIANO)

Johannes è un progetto che ha come obiettivo di ricostruire e simulare, tramite intelligenza artificiale, la memoria di un abitante, ribattezzato appunto "Johannes", del castello di Miranduolo (Chiusdino, SI). L'indagine archeologica ha permesso di scavare, tra 2001 e 2016, praticamente l'intero spazio cinto da mura e riconoscere le fasi altomedievali dell'insediamento. La prima attestazione nella documentazione scritta che parla di Miranduolo è datata al 1004, ma le fonti materiali hanno permesso di comprendere le vicende che hanno portato alla fondazione del sito nel VII secolo per ragioni minerarie e come il villaggio si sia poi evoluto fino alla trasformazione in castello.
Lo scavo del cimitero ha permesso di individuare alcune sepolture da ricollegare allo scontro che si svolse a Miranduolo negli anni precedenti al 1133 tra le truppe del Vescovo di Volterra ei difensori del castello, ovvero gli uomini della famiglia dei Gherardeschi. In particolare l'inumato SK86, che presenta un quadro di lesioni da ricondurre ad azioni di combattimento. Johannes-SK86 sopravvisse all'assedio, ma portò con sé, per gli ultimi anni della sua vita, i segni di quell'esperienza.
Johannes è un tentativo, di simulazione di una memoria artificiale, in cui il personaggio in questione racconta il suo punto di vista, tramite un'interazione vocale con gli utenti. È un racconto parziale e di parte, ma permette di interagire con un abitante del castello, ricostruito sulla base delle conoscenze storico-archeologiche acquisite; è in definitiva anche un esperimento funzionale alla nostra capacità professionale di modellare la storia, attraverso un approccio bottom up, che parte dall'esperienza di vita di un qualunque abitante del passato, popolando concetti complessi legati agli aspetti tecnologici, emotivi, religiosi e di esperienza corporeo-cinestetica.
Parole chiave: memoria artificiale; archeologia virtuale e artificiale; agenti virtuali; archeologia medievale

ABSTRACT (ENGLISH)

Toward an Artificial Memory? The Johannes Project between Technological, Archaeological, and Moral Issues.
Johannes is a project aimed at reconstructing and simulating, through artificial intelligence, the memory of an inhabitant—renamed "Johannes"—of the castle of Miranduolo (Chiusdino, Siena, Italy).
Archaeological research carried out between 2001 and 2016 excavated virtually the entire enclosed area within the walls and identified the early medieval phases of the settlement. The first written reference to Miranduolo dates to 1004, yet the material record has made it possible to trace the events leading to the site's foundation in the 7th century, linked to mining activities, and to reconstruct how the village subsequently developed and was later transformed into a castle.
Excavation of the cemetery revealed several burials that can be connected to the conflict that took place at Miranduolo in the years preceding 1133, between the troops of the Bishop of Volterra and the castle's defenders, the men of the Gherardeschi family. Of particular interest is the inhumation SK86, which displays a pattern of lesions consistent with combat-related injuries. Johannes–SK86 survived the siege, but carried the physical traces of that experience into the final years of his life.
Johannes represents an attempt to simulate an artificial memory in which the character narrates events from his own perspective through voice-based interaction with users. The narrative is inevitably partial and subjective, yet it enables engagement with a castle inhabitant reconstructed on the basis of current historical and archaeological knowledge. Ultimately, it is also an experiment that tests our professional capacity to model the past through a bottom-up approach: starting from the lived experience of an ordinary person, it gives substance to complex concepts related to technology, emotions, religion, and embodied (corporeal–kinaesthetic) experience.
Keywords: artificial memory; virtual and artificial archaeology; virtual agents; medieval archaeology

INTRODUZIONE

Il presente contributo intende presentare il cosiddetto Progetto Johannes, ovvero una sperimentazione di ricostruzione artificiale della memoria di un essere umano del passato (Bertoldi & Valenti, 2024); nel dettaglio si tratta di un uomo che visse e morì nella prima metà del XII secolo nel castello di Miranduolo (Chiusdino, SI). Da un punto di vista tecnologico, è stata sfruttata la nuova API Realtime di OpenAI che consente di creare agenti virtuali speech-to-speech e un server dell'Università degli Studi di Siena, che ospita la "memoria", l'identità, l'epistemologia, le regole, oltre che il volto dell'avatar.
Gli obiettivi erano quelli di proporre un dispositivo interattivo di conversazione diretta con un Essere artificiale e virtuale proveniente dal passato, mettendo in evidenza i limiti della nostra ricostruzione storico-archeologica, incapace di modellare aspetti fondativi dell'esperienza umana, legati all'emotività, spiritualità, esperienza corporea dell'individuo.
La complessità tecnica è stata legata a due aspetti divergenti ma profondamente interconnessi; oltre alla parzialità delle nostre informazioni su aspetti sconosciuti (che abbiamo già elencato), dovevamo "imbrigliare" la mente di Johannes, per cancellare le sue conoscenze su quella lunga fase che abbiamo rinominato come fase liminare della coscienza, che è compresa tra la morte dell'individuo e la ricostituzione artificiale della sua memoria.
L'avatar troverà definitiva sede nel nascente Museo di Miranduolo, che verrà inaugurato a Chiusdino nel prossimo mese di luglio e sarà a disposizione dei visitatori per spiegare, attraverso la voce e la mente di un abitante del medioevo toscano, l'esperienza di vita di un abitante di un castello di XII secolo. Non sarà (non vuole essere questo) una guida virtuale del museo, ma una simulazione esperienziale che permetterà di immergersi nel passato attraverso una sorta di capsula del tempo.

1. MIRANDUOLO DALLA FONDAZIONE AL XII SECOLO

Miranduolo (fig. 1), sul rilievo di Costa Castagnoli in alta Val di Merse, è un osservatorio privilegiato per seguire in diacronia la trasformazione di un paesaggio produttivo, politico e sociale del Medioevo toscano.



Figura 1: La val di Merse con lo scavo di Miranduolo sulla destra (Ricostruzione grafica InkLink)

Il sito viene fondato nel VII secolo come villaggio-fabbrica minerario e metallurgico, legato allo sfruttamento di filoni a solfuri misti e alla riduzione del ferro: una comunità relativamente piccola composta da circa 30–40 persone, distribuite in 8–9 capanne, opera in un sistema specializzato, con consumi abbastanza omogenei, ma con una chiara emergenza funzionale e gerarchica rappresentata dalla capanna C53, isolata e dotata di fosse granarie e il cui abitante era dotato di cavalcatura, che viene interpretata come punto di controllo della filiera e della redistribuzione delle risorse; il ferro, ridotto in lingotti o barrette, sembra destinato a uscire dal sito, probabilmente verso Volterra, in un quadro compatibile con un'iniziativa pubblica/fiscale del regno longobardo.
Tra VIII e IX secolo la vocazione insediativa cambia profondamente e il sito evolve verso un'organizzazione agraria sempre più strutturata fino alla piena fase curtense: se nell'VIII secolo si osservano due distinti possessores (una delle due aree di potere era dotata di chiesa, forse simbolo di controllo ecclesiastico), personaggi dotati di possibilità economiche importanti e forse anche di libertà personali (un fabbro, un fattore agricolo e probabilmente un ceramista), nel corso del IX secolo la proprietà si accentra in un unico dominus, la sommità si fortifica con fossato e palizzata e una seconda linea difensiva separa il dominicum dal massaricium sui terrazzi; un actor controlla lavoro e flussi di derrate, mentre il surplus viene raccolto e difeso nei magazzini superiori, con una redistribuzione selettiva della cultura materiale che "si irradia" proprio dal magazzino della casa dominica. In questa fase la chiesa perde la funzione economica di polo economico e si configura come pertinenza del dominus, e ai suoi piedi si imposta un'area cimiteriale che sostituisce precedenti spazi di stoccaggio. A metà X secolo un incendio, verosimilmente connesso a un attacco armato, colpisce soprattutto l'area sommitale; si tratta probabilmente dell'assalto che permette la conquista dei Gherardeschi, che proprio in questa fase si stavano espandendo nella zona della val di Merse. Tra ultimo quarto del X e primo quarto dell'XI secolo si osserva una ricostruzione di edifici e difese con zoccoli in pietra ed elevati in terra pressata; cambiano le tecniche costruttive, ma non muta la struttura della società. L'unico edificio ricostruito interamente in pietra è la chiesa. Nel 1004, quando viene citato Miranduolo per la prima volta, l'insediamento è attestato come castello con chiesa e curtis di proprietà gherardesca. Tra primo e secondo quarto dell'XI si verifica un secondo attacco a Miranduolo attestato archeologicamente, che però mira a colpire magazzini e difese del castello; potrebbe trattarsi di una prima fase della conflittualitò tra i Gherardeschi e il Vescovo di Volterra, che sfocerà definitivamente nel XII secolo, anche se le fonti scritte non testimoniano nulla in tal senso. Dopo questo evento Miranduolo viene ricostruito integralmente per volontà dei Gherardeschi, utilizzando la pietra: viene realizzato un circuito murario comprendere l'intera collina. Viene costruito un grande edificio signorile, un'altra abitazione nella sommità, destinata a un personaggio legato al signore. Il sistema d'accesso è fortificato e viene costruita la porta con una torre adiacente. La chiesa viene pavimentata in pezzame di pietra, mentre lo spazio immediatamente circostante alla stessa assume una destinazione cimiteriale ormai stabile. All'interno del nuovo circuito murario, la ricostruzione comprende anche un tessuto abitativo costituito da sette edifici abitativi in pietra. In questo scenario si osserva l'edificio più "specializzato" dopo il palazzo: la cosiddetta caserma. Si tratta di una struttura rettangolare allungata, probabilmente a due piani, costruita in pietra e collocata in un punto strategico: di fronte al lato nord-est del fossato e a controllo di una piccola postierla nelle mura. La sua funzione deve essere interpretata come residenza delle guardie/armati dipendenti dalla famiglia signorile, con il piano terra utilizzato come stalla (Valenti, 2022: 14-31).
Nel XII secolo Miranduolo si colloca al crocevia tra consolidamento signorile, competizione territoriale e conflitto armato, mostrando in modo quasi "stratigrafico" l'intreccio fra architettura del potere e dinamiche politiche. L'antefatto è di lunga durata: già dagli inizi del secolo il vescovo di Volterra getta le basi di una decisa affermazione signorile sul territorio in questione, sostenuta dal controllo delle miniere argentifere delle Colline Metallifere (con snodi cruciali a Montieri), e agisce in un sistema di alleanze che lo vede contro Siena e vicino a Firenze, in un clima di tensioni e competizioni tra poteri pubblici, privati ed ecclesiastici che investe direttamente anche l'area gherardesca. Proprio in questi decenni Miranduolo raggiunge la massima maturità come castello (fig. 2): nei decenni immediatamente precedenti e a cavallo del pieno XII secolo si registrano cantieri che accrescono monumentalità e capacità difensiva, con interventi sul palazzo (articolato su più livelli), sulla chiesa e sull'accesso, oltre che la costruzione di una seconda torre limitrofa al palazzo.
Su questo scenario si innesta lo scontro, che nel primo trentennio del XII secolo sfocia in un assedio pesante condotto dalle truppe del presule volterrano (fig. 3) e viene ricordato dal lodo di pace del 1133. La sequenza ricostruibile combina una fase di combattimento "esterno" e una di pressione sulle difese. Gli esiti materiali dell'assedio sono netti ma selettivi: crollo del circuito murario fino alle fondazioni sul lato

meridionale/sud-ovest, gravi danneggiamenti alla torre presso la porta, abbattimento di un'abitazione in quell'area, probabili guasti alla chiesa e compromissioni anche in sommità in corrispondenza della torre e della cisterna. Tuttavia, la conseguenza immediata non è la fine del castello: non si osservano abbandoni e la cultura materiale indica continuità di vita lungo tutto il secolo. Sul piano storico-politico, il conflitto e la pace del 1133 ridisegnano rapporti di forza e identità interne alla casata, mentre nel giro di pochi decenni cambia anche lo scenario regionale: l'ingresso di Siena (esplicito nella donazione del 1178 di metà del castello) apre una nuova fase di coabitazioni e negoziazioni, leggibile anche nella predisposizione di spazi edilizi "paritari" di fronte alla zona signorile, pur destinati a restare in gran parte progetti incompiuti nella traiettoria di lunga durata del sito (Nardini 2022: 75-80).

2. DAL CORPO ALLA MEMORIA; LO SCAVO DI SK86 E LA RICOSTRUZIONE DELLA SUA IDENTITÀ

Lo scavo dell'area cimiteriale intorno alla chiesa di San Giovanni Evangelista, ha permesso di raccogliere importanti informazioni sulla vita, il lavoro e la salute degli abitanti di Miranduolo; tra queste è stata scavata una sepoltura (definita dal codice SK86, fig. 4) che mostra chiaramente i segni tangibili di uno scontro armato, con cronologia di XII secolo (su base stratigrafica e di analisi della cultura materiale). SK86 presenta traumi concentrati sul capo e sulla mandibola: un fendente sul frontale sinistro e una piccola depressione ovoidale sul frontale destro, compatibili con colpi da arma da taglio e da oggetto contundente appuntito. Il condilo mandibolare mostra rimaneggiamento osseo, segno di sopravvivenza e di esiti cronici sulla masticazione; la TC indica che le lesioni craniche non penetrarono fino alle meningi, suggerendo uno scontro non letale. Quello che abbiamo ribattezzato come Johannes affrontò le milizie del vescovo, sui piani o dentro al castello, venne ferito profondamente nello scontro, ma non morì quel giorno: sopravvisse e potè raccontare quella traumatica esperienza.
Oltre all'analisi antropologica, per SK86 è stata applicata la ricostruzione facciale forense secondo il "protocollo di Manchester", a cura di Stefano Ricci (Università degli Studi di Siena, Dipartimento di Scienze fisiche, della Terra e dell'Ambiente). Sul cranio si posizionano marcatori di spessore dei tessuti, poi si ricostruiscono in sequenza i principali muscoli mimici e masticatori, modellando il volto in plastilina su basi anatomiche standardizzate (Abate & Ricci, 2022). Il modello finito viene quindi scansionato e rifinito digitalmente con una texture realistica coerente con i dati biologici (fig. 5).
Dopo gli aspetti storici, antropologici e biologici, è stato svolto un lavoro di ricostruzione della memoria di SK86, facendo riemergere quindi Johannes. La struttura informatica si basa sulla nuova API Realtime di OpenAI che consente di creare agenti virtuali speech-to-speech, che si interfaccia con una webapp, ospitata su un server dell'Università degli Studi di Siena e realizzata in Node.js più Express, composta anche da json utili a regolamentare il comportamento e le conoscenze di Johannes: in particolare si tratta di identità, epistemologia, regole e soprattutto la memoria del nostro personaggio, che ospita informazioni su Miranduolo e sugli altri insediamenti del territorio, su Siena e Populonia che (nella nostra ipotesi verosimile) avrebbe visitato e limitate conoscenze sul mondo esterno e sul passato.

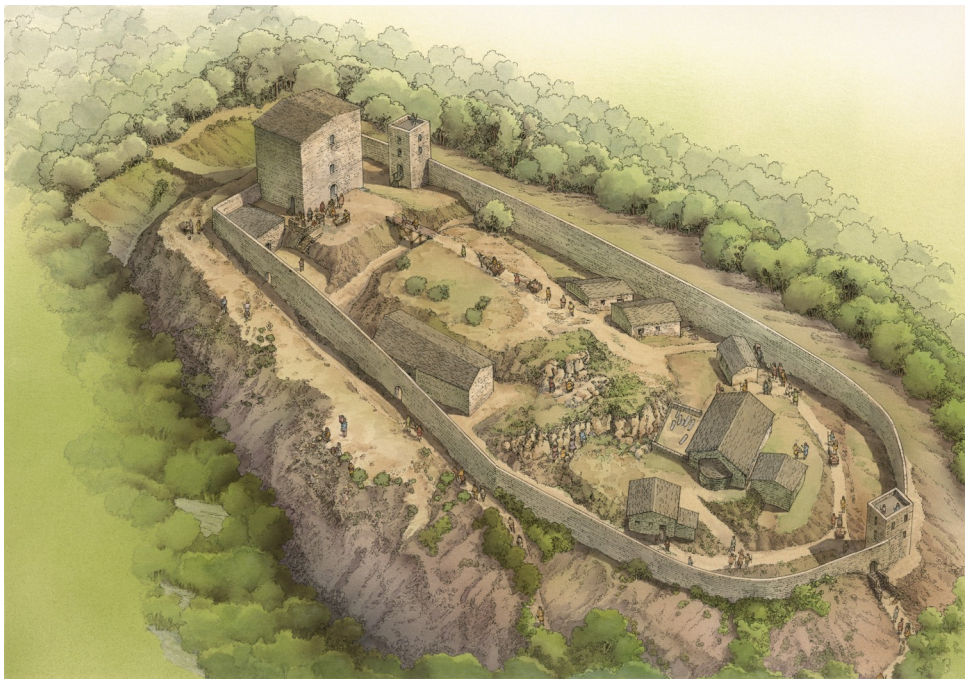


Figura 2: Miranduolo agli inizi del XII secolo (Ricostruzione grafica InkLink)

Abbiamo così legato la memoria artificiale ad un determinato periodo storico, senza possibilità di conoscere ciò che succede dopo la sua morte (il portale spazio-temporale di connessione tra noi e lui conduce in un momento in cui Johannes era ancora vivo). La possibilità di utilizzare anche modelli speech-to-text e text-to-speech, ha permesso di implementare una sovrastruttura di coaching per la memoria; è infatti possibile, in questa fase di implementazione e anche successivamente, quando ci sarà bisogno di inserire concetti e informazioni nella mente dell'agente artificiale, arricchirla con testi. Allo stesso modo, considerato che Johannes può "improvvisare" ove necessario (sempre rispettando però i suoi orizzonti epistemologici e identitari), le sue storie vengono trascritte in una apposita pagina web, in cui il coach può verificare le informazioni, approvarle, modificarle e/o respingerle (Suchman, 2007).

3. PROBLEMATICHE TECNOLOGICHE, ARCHEOLOGICHE, MORALI AL PROGETTO JOHANNES

Johannes nasce come esperimento di "archeologia artificiale" in cui la ricostruzione di un individuo del XII secolo non è delegata alla sola potenza generativa del modello, ma a una filiera controllata che tiene insieme tecnologia, vincoli archeologico-storici e responsabilità morale. In questo senso, ogni scelta tecnica (interfaccia, voce, avatar, memoria) diventa un pezzo del ragionamento teorico: non un semplice "mezzo", ma una componente epistemica (Jasanoff, 2004).
La webapp (fig. 6) tenta di tradurre questa impostazione in scelte architetturali precise (Akrich, 1992). Il nodo metodologico più delicato – e più produttivo – è però la memoria.



Figura 3: Fasi finali dell'assedio di Miranduolo (Ricostruzione grafica InkLink)

La webapp separa ciò che Johannes "dice" da ciò che diventa "memoria di Johannes". In altri termini, la continuità non è affidata alla conversazione (che può tendere a far sedimentare qualunque ipotesi come "fatto"), ma a un corpus ispezionabile e modificabile, che può essere trattato come un protocollo (Star, 1999).
In questo scenario emergono però dei problemi specifici: la doppia generazione può produrre meta-allucinazioni. Non solo Johannes può inventare, ma anche l'estrattore può "vedere" storie dove non ci sono o irrigidire sfumature in episodi; lo schema JSON e i filtri riducono il rumore, ma non sostituiscono l'audit umano. L'intero recinto istruzionale deve resistere a drift e a intrusioni informative; in questo senso bisogna essere attenti a distinguere le intrusioni come errori, e le intrusioni come dinamica strutturale della memoria (umana e artificiale) che va compresa e non necessariamente rimossa.



Figura 4: Lo scavo della sepoltura SK86



Figura 5: Ricostruzione del volto di SK86 (realizzazione: Stefano Ricci)

La sicurezza non è solo una questione informatica, ma epistemica: ogni punto in cui dati esterni entrano (input utente, trascrizioni, estrazione automatica) è un possibile varco per anacronismi o per la solidificazione di narrazioni non fondate.
In definitiva il progetto non deve essere inteso come semplice interfaccia di restituzione, ma come un assemblaggio sociotecnico in cui fonti archeologiche, ricostruzione antropologica, architettura informatica, filtri epistemologici, controllo umano e interazione museale concorrono a produrre il personaggio. In questa prospettiva, l'agente non coincide né con il modello generativo né con l'avatar, ma con una rete di

mediazioni che ne circoscrive possibilità, limiti e forme di parola. La sua agency non è quindi autonoma in senso forte, ma situata, distribuita e continuamente regolata da vincoli documentari, scelte infrastrutturali e pratiche di curatela.



Figura 6: La webapp Johannes

4. CONCLUSIONI E SVILUPPI FUTURI

Johannes non può (e non vuole) essere un'enciclopedia del Medioevo: deve essere un individuo situato entro un orizzonte locale e biografico costruito da tracce che non equivalgono al "passato in sé" (fig. 7).

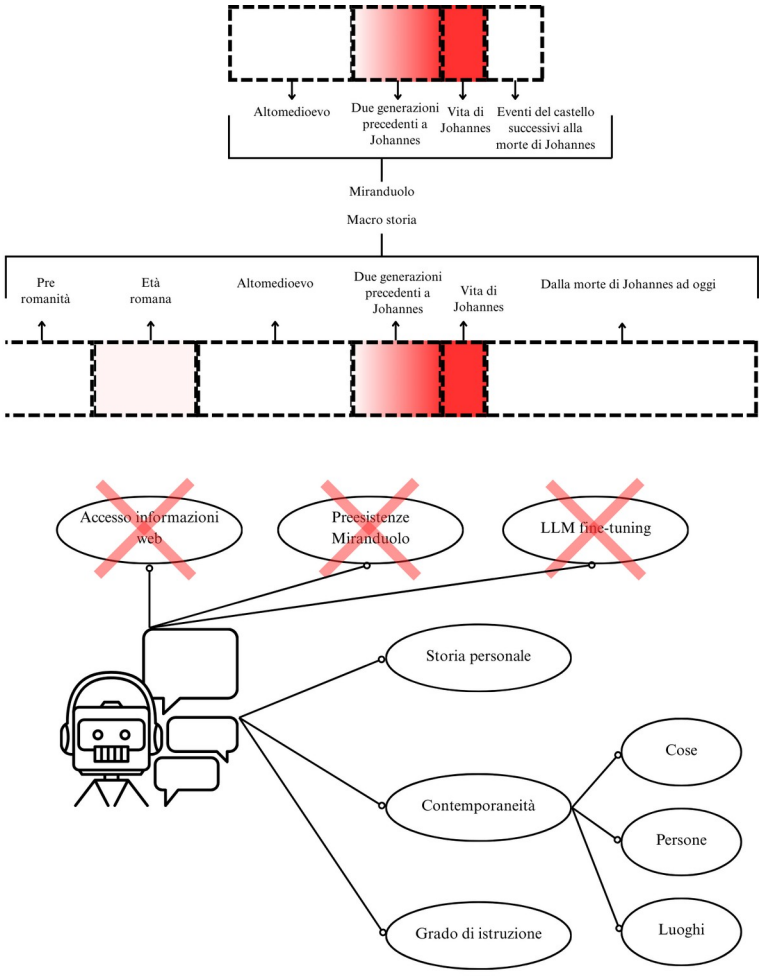


Figura 7: Schema del funzionamento della memoria di Johannes

Confrontare memoria e deposito stratigrafico, in questo senso, può essere utile a comprendere la forza della ricostruzione come modello storico: esistono unità positive e negative, lacune che non possono essere colmate, residualità e rimaneggiamenti; e, come nello scavo, anche qui la lettura è sempre una forma disciplinata di abduzione.
Voce e volto amplificano l'attribuzione di interiorità (quella che potremo definire "coscienza percepita"), rendendo più urgente un'etica della trasparenza (Jasanoff, 2003): se Johannes "sembra" testimone, il pubblico rischia di scambiare una ricostruzione vincolata per una prova; se Johannes "sembra" cosciente, si rischia di proteggere l'illusione invece della qualità della conoscenza.
Da qui discendono alcune esigenze metodologiche che la webapp già tenta di incarnare: dichiarare i limiti epistemologici, rendere tracciabile la formazione della memoria e mantenere sempre distinguibili traccia, interpretazione e invenzione.
La sperimentazione museale dovrà prevedere una prima valutazione qualitativa dell'interazione, finalizzata a verificare non solo il coinvolgimento dei visitatori, ma anche la loro comprensione dello statuto epistemico dell'agente. Sarà quindi importante osservare se Johannes venga percepito come testimone storico, come simulazione interpretativa o come guida narrativa, ricorrendo a brevi interviste in uscita su comprensione, fiducia e percezione dell'incertezza.
In sintesi, Johannes funziona quando i tre livelli restano allineati: tecnologia come recinto e tracciabilità; fonti scritte e materiali come insieme di vincoli e di lacune da preservare; morale come progettazione contro l'illusione e contro la sedimentazione automatica dell'invenzione.

RINGRAZIAMENTI

Johannes fa parte del progetto per lo svolgimento di attività di ricerca dell'ambito del PNRR – Missione 4 "Istruzione e Ricerca" - Componente 2 "Dalla ricerca all'Impresa" – Investimento 1.2 "Finanziamento di progetti presentati da giovani ricercatori" – Decreto per l'assunzione di Ricercatori Internazionali Post-Dottorato - Avviso MUR D.D. N. 47 del 20 febbraio 2025 – GSD: 10/ARCH-01 - ARCHEOLOGIA – SSD: ARCH-01/E – ARCHEOLOGIA CRISTIANA, TARDOANTICA E MEDIEVALE - Dipartimento di Scienze storiche e dei beni culturali - Università degli Studi di Siena.
È un'azione promossa dall'Arena 2 "Nuove tecnologie nella Ricerca sul Patrimonio Culturale", SOPaC "Sostenibilità e Occupabilità nel Patrimonio Culturale", Progetto d'Eccellenza del Dipartimento di Scienze Storiche e dei Beni Culturali dell'Università degli Studi di Siena.
Fa parte della linea di finanziamento denominata "Chiusdino RigerErA, Coprogettazione allestimento del Museo di Miranduolo", Finanziamento PNRR NextGenerationEU, Missione 1 – Digitalizzazione, innovazione, competitività e cultura, Component 3 – Cultura 4.0 (M1C3) - Misura 2 "Rigenerazione dei piccoli siti culturali, patrimonio culturale, religioso e rurale", investimento 2.1: "Attrattività dei borghi storici".

BIBLIOGRAFIA

Abate, E. & Ricci, S. (2022). Restauro, studio paleoantropologico e tafonomico degli inumati del cimitero medievale di Miranduolo. In M. Valenti, A. Nardini, & S. Bertoldi (Eds.), Miranduolo. Le lunghe mutazioni del potere tra VII e XIV secolo, 121-134. Firenze: All'Insegna del Giglio.

Akrich, M. (1992). The De-Scription of Technical Objects. In W.E. Bijker & J. Law (Eds.), Shaping Technology/Building Society: Studies in Sociotechnical Change, 205-224. Cambridge (MA), MIT Press.

Bertoldi, S. & Valenti, M. (2024). Artificial Beings in Archaeology: Bridging Historical Reality and Digital Simulation. Archeomatica 15 (3), 14-19.

Jasanoff, S. (2003). Technologies of Humility: Citizen Participation in Governing Science. Minerva, 41(3), 223-244.

Jasanoff, S. (Ed.) (2004). States of Knowledge: The Co-Production of Science and Social Order. Londra-New York: Routledge.

Nardini, A. (2022). L'insediamento e i rapporti politici e sociali tra XI e XIV secolo. In M. Valenti, A. Nardini, & S. Bertoldi (Eds.), Miranduolo. Le lunghe mutazioni del potere tra VII e XIV secolo, 67-86. Firenze: All'Insegna del Giglio.

Star, S.L. (1999). The Ethnography of Infrastructure. American Behavioral Scientist, 43(3), 377-391.

Suchman, L. (2007). Human-Machine Reconfigurations: Plans and Situated Actions (2nd ed.). Cambridge: Cambridge University Press.

Valenti, M. (2022). Le lunghe mutazioni del potere (VII-inizi XIV secolo). In M. Valenti, A. Nardini, & S. Bertoldi (Eds.), Miranduolo. Le lunghe mutazioni del potere tra VII e XIV secolo, 14-39. Firenze: All'Insegna del Giglio.

AI confini della realtà: il pensiero umano nell'epoca della sua riproducibilità generativa

DANIELE SILVI
UNIVERSITÀ DEGLI STUDI DI ROMA 'TOR VERGATA'

ABSTRACT (ITALIANO)

Il progetto GENREAL: la fiction dell'Intelligenza Artificiale
In un'epoca in cui la tecnologia non solo riproduce la realtà, ma la genera, sfidando il nostro concetto di autenticità, questo intervento intende suscitare una discussione sul tema della autenticità generativa, riportando un progetto di ricerca svolto da colleghi dell'Università di Cambridge. Verranno quindi esposti i risultati dell'indagine che rappresenta lo step iniziale del progetto GENREAL (GENerative REALities: Exploring AI Learning and Creativity), sviluppato dall'Università di Cambridge (Department of Social Anthropology), nel 2024, volto a esplorare il modo in cui le tecnologie generative influenzano i processi di apprendimento e creatività nei contesti accademici (Floridi, 2014). Ma che cosa succede al dibattito critico quando questo è svolto in un contesto accademico, proposto da accademici e rivolto ad altri accademici? Siamo sicuri che quanto ci viene detto è 'vero'? e che cosa significa 'vero'?
Parole chiave: Intelligenza artificiale; simulazione; epistemologia; autenticità; Digital Humanities;

ABSTRACT (ENGLISH)

The GENREAL Project: The Fiction of Artificial Intelligence.
In an era in which technology not only reproduces reality but actively generates it, challenging our very notion of authenticity, this contribution aims to foster a discussion on the concept of generative authenticity by presenting a research project conducted by colleagues at the University of Cambridge. The paper will outline the initial findings of an investigation that constitutes the first stage of the GENREAL project (GENerative REALities: Exploring AI Learning and Creativity), developed at the University of Cambridge (Department of Social Anthropology) in 2024, and aimed at exploring how generative technologies influence learning and creativity in academic contexts (Floridi, 2014).
Within this framework, a central question emerges: what happens to critical debate when it takes place within an academic context, conducted by academics and addressed to other academics? Can we be certain that what we are told is "true"? And what does "true" actually mean?
Keywords: Artificial Intelligence; Simulation; Epistemology; Authenticity; Digital Humanities;

1. IL PROGETTO GENREAL: LA FICTION DELL'INTELLIGENZA ARTIFICIALE

Il progetto GenReal nasce con l'obiettivo di esplorare l'impatto delle tecnologie generative, in particolare l'intelligenza artificiale (AI), sui processi di apprendimento e creatività nei contesti accademici. In un momento storico in cui la tecnologia non si limita più a riprodurre la realtà ma la genera autonomamente, il progetto si propone di indagare come questi strumenti influenzino la percezione dell'autenticità, ridefinendo i confini tra originale e generato.
Per rispondere a tali questioni, il progetto ha coinvolto 1.500 studenti di corsi in Digital Humanities presso un college partner dell'Università di Cambridge. Lo studio ha analizzato le percezioni degli studenti sull'uso
delle tecnologie generative nei loro percorsi educativi e creativi, con particolare attenzione a strumenti come ChatGPT, DALL·E e altre piattaforme di intelligenza artificiale. Attraverso questionari, interviste e analisi qualitative, è stato possibile raccogliere dati su tre aree principali:
- Il ruolo dell'AI nella scrittura accademica e creativa.
- L'utilizzo dell'AI nell'analisi dei dati e nella ricerca.
- Le implicazioni etiche e culturali legate all'uso di strumenti generativi.

Questi aspetti si inseriscono anche in una riflessione più ampia sul futuro dell'umanità nell'epoca dell'AI, come discusso da Tegmark (2017), il quale analizza il potenziale dell'intelligenza artificiale nel ridefinire l'identità umana e il nostro rapporto con la tecnologia.

Si tratta di una forte opportunità per comprendere da vicino il ruolo delle tecnologie generative proprio nei
contesti che le fanno da genitore, vale a dire gli ambienti accademici e culturali (Harari, 2018; Brynjolfsson
& McAfee, 2014). Inizieremo quindi condividendo i risultati dell'indagine condotta nel college, focalizzata sulla percezione degli studenti riguardo all'influenza delle tecnologie generative, come l'intelligenza artificiale, nel loro percorso educativo e creativo. Sono stati intervistati 1500 studenti – nel corso di un intero anno accademico – provenienti da corsi di laurea in Digital Humanities. L'obiettivo principale era comprendere come venisse percepito l'uso dell'AI nei loro studi, in particolare per la scrittura, l'analisi dei dati e la ricerca creativa. Le domande includevano sia opinioni generali che esperienze specifiche legate all'uso di strumenti digitali avanzati.

2. METODOLOGIA E RISULTATI ATTESI

Lo studio ha utilizzato un approccio misto, combinando metodi quantitativi e qualitativi. I questionari distribuiti a un ampio campione di studenti hanno fornito una visione generale delle loro percezioni, mentre
le interviste semi-strutturate hanno approfondito esperienze specifiche e riflessioni personali. La ricerca si è avvalsa di tecniche di analisi testuale automatizzata e codifica manuale per interpretare i dati raccolti. Si
prevede che i risultati del progetto forniscano una comprensione più approfondita del ruolo dell'AI nell'ambito accademico, evidenziando come le tecnologie generative possano essere utilizzate non solo come strumenti di supporto, ma come catalizzatori per nuove forme di apprendimento e creatività. Inoltre,
si mira a fornire linee guida per l'integrazione etica e consapevole di queste tecnologie nei curricula universitari.

3. I RISULTATI DELLA RICERCA

Inizieremo quindi condividendo i risultati dell'indagine condotta nel college, focalizzata sulla percezione degli studenti riguardo all'influenza delle tecnologie generative, come l'intelligenza artificiale, nel loro percorso educativo e creativo. Sono stati intervistati 1500 studenti – nel corso di un intero anno accademico – provenienti da corsi di laurea in Digital Humanities. L'obiettivo principale era comprendere come venisse percepito l'uso dell'AI nei loro studi, in particolare per la scrittura, l'analisi dei dati e la ricerca creativa. Le domande includevano sia opinioni generali che esperienze specifiche legate all'uso di strumenti digitali avanzati. L'indagine ha evidenziato significative differenze tra coloro che utilizzano sistematicamente l'intelligenza artificiale generativa (gruppo A) e coloro che non la utilizzano (gruppo B). Il campione è stato suddiviso in proporzione quasi equivalente: 700 studenti nel gruppo A e 800 studenti nel gruppo B.

1. Percezione dell'IA generativa
• Gruppo A (utilizzatori):
o 75%: Ritiene che l'IA generativa sia uno strumento che migliora l'efficienza nello studio.
o 15%: La considera un'innovazione interessante, ma non sempre necessaria.
o 10%: Esprime preoccupazioni per una possibile dipendenza tecnologica.
• Gruppo B (non utilizzatori):
o 45%: Ritiene che l'IA sia interessante ma difficile da integrare nelle attività quotidiane.
o 35%: Esprime scetticismo verso l'efficacia degli strumenti generativi.
o 20%: Non si fida delle tecnologie IA per motivi etici o pratici.

2. Impatto sull'apprendimento
• Gruppo A:
o 43,2%: Ha dichiarato che l'IA generativa ha migliorato la produttività nei progetti di scrittura (es. redazione di saggi).
o 35,1%: Utilizza strumenti come ChatGPT per sviluppare idee creative o analizzare dati complessi.
o 21,6%: Ha segnalato una maggiore difficoltà nel valutare la qualità delle fonti suggerite dall'IA.
• Gruppo B:

o 50%: Ha dichiarato che preferisce metodi tradizionali di studio e ricerca.
o 30%: Ritiene che l'assenza di IA non influisca negativamente sulla propria performance accademica.
o 20%: Ha espresso curiosità verso l'integrazione dell'IA, ma si sente poco preparato ad utilizzarla.

3. Creatività e tecnologie generative
• Gruppo A:
o 73,7%: Sostiene che l'IA generativa stimoli nuove idee, specialmente in progetti interdisciplinari.
o 26,3%: Trova che l'uso di strumenti IA renda più complessa l'autovalutazione della propria creatività.
• Gruppo B:
o 60%: Ritiene che il processo creativo sia puramente umano e non necessiti di supporto tecnologico.
o 30%: Ammette che potrebbe essere utile esplorare strumenti generativi per progetti specifici.
o 10%: È convinto che l'IA ostacoli la creatività autentica.

4. Barriere all'uso dell'IA
• Gruppo A:
o 55%: Ha segnalato difficoltà tecniche nell'apprendimento degli strumenti.
o 30%: Ha riscontrato problemi etici, come la mancanza di trasparenza nei processi generativi.
o 15%: Non percepisce alcuna barriera significativa.• Gruppo B:
o 50%: Ritiene di non avere competenze tecniche sufficienti per utilizzare strumenti generativi.
o 35%: Ha espresso timori per la protezione dei dati e la privacy.
o 15%: Non percepisce barriere all'utilizzo, ma preferisce metodi tradizionali.

4. OSSERVAZIONI

Dall'analisi emerge che gli studenti che utilizzano l'IA generativa tendono a percepirla come un supporto concreto per migliorare produttività e creatività, sebbene rimangano preoccupazioni legate all'etica e alla trasparenza. Gli studenti che non utilizzano l'IA, invece, mostrano un mix di curiosità e resistenza, indicando un bisogno di formazione e sensibilizzazione per abbattere le barriere all'adozione. Un dato particolarmente interessante è che entrambi i gruppi — utilizzatori e non utilizzatori — evidenziano circa con la stessa percentuale (30/35 %) problemi etici legati all'intelligenza artificiale generativa. Questo suggerisce che, indipendentemente dal livello di utilizzo, la percezione dei rischi etici rappresenta un terreno comune tra gli studenti. Le principali preoccupazioni riguardano la trasparenza dei processi generativi, la possibilità di manipolazione delle informazioni e l'impatto sull'autenticità del lavoro creativo. Questi dati sottolineano l'importanza di promuovere una discussione aperta e critica sull'uso responsabile di queste tecnologie, integrando linee guida etiche nei percorsi educativi.

5. AI CONFINI DELLA REALTÀ: IL PENSIERO UMANO NELL'EPOCA DELLA SUA RIPRODUCIBILITÀ GENERATIVA

In un momento storico in cui la preoccupazione che le Intelligenze artificiali possano creare contenuti indistinguibili da quelli derivanti dal pensiero e dall'opera umana, preoccupazione che ha suscitato diversi dibattiti – spesso dai toni accesi – la mia riflessione punta i riflettori su quanto più antico sia in realtà il problema e dove realmente sia da collocare.
Un dibattito recente e significativo si articola, ad esempio, tra la posizione di Daniel Dennett e quella di Walter Quattrociocchi. Dennett descrive l'intelligenza artificiale come un caso di "competence without comprehension" (Dennett, 2017), ovvero un sistema capace di produrre risultati efficaci senza possedere una reale comprensione dei processi che li generano. Il progetto GENREAL si inserisce precisamente in questa dinamica, risultando epistemicamente efficace pur in assenza di un referente empirico reale.
Dall'altro lato, Quattrociocchi ha mostrato come la diffusione di contenuti non verificati sia influenzata da bias cognitivi e dinamiche sociali (Quattrociocchi, 2016), spostando il problema sul piano della ricezione e della credibilità. Tuttavia, entrambe le prospettive sembrano presupporre la possibilità di distinguere stabilmente tra vero e falso, senza interrogare pienamente i meccanismi attraverso cui tale distinzione viene costruita all'interno dei contesti accademici e discorsivi.Ciò che verrà discusso nella sezione seguente cerca di porre il focus proprio su questi aspetti della percezione del reale sul fronte della

trattatistica scientifica, laddove si tende a dare per scontato che insiemi di dati strutturati siano credibili a prescindere (numeri, tabelle, grafici e simili) proprio perché hanno una 'veste' scientifica. Il discorso che intendo quindi provocare non rappresenta una semplice riflessione a posteriori, ma costituisce il fondamento epistemologico dell'intero dispositivo GENREAL. Il contributo finora qui presentato adotta consapevolmente una strategia di simulazione progettuale. Il progetto GENREAL (GENerative REALities: Exploring AI Learning and Creativity), così come descritto nelle sezioni precedenti, non corrisponde a una ricerca empirica attualmente in corso, ma costituisce un dispositivo critico deliberatamente costruito per essere indistinguibile, nella forma e nel linguaggio, da un progetto di ricerca reale nel campo delle Digital Humanities. Si dota, in altri termini, proprio di quelle caratteristiche 'scientifiche' (dati, tabelle, percentuali, ecc) che precedendolo gli conferiscono credibilità.
Partner istituzionali, dataset, risultati quantitativi e qualitativi, così come le modalità di presentazione pubblica del progetto, sono elementi fittizi ma verosimili, progettati a fini dimostrativi. La loro funzione non è quella di presentare dati empirici, bensì di attivare una riflessione sui meccanismi di credibilità, autorità scientifica e validazione epistemica che regolano la produzione e la circolazione della conoscenza nell'ecosistema accademico contemporaneo, in particolare in relazione all'uso di tecnologie di intelligenza artificiale generativa. Il partner istituzionale che dialoga con il sottoscritto durante l'esposizione è un avatar generato con il programma Synthesia, alimentato da una IA, che simula un docente di una università straniera: il prof. Alexander Reed (come spiegato nella descrizione del 'progetto civetta').
La simulazione non va pertanto intesa come inganno, ma come esperimento metodologico e performativo, in cui la forma stessa del progetto (e la sua genesi) diventa oggetto di analisi. In questo senso, anche la presentazione al convegno è concepita come parte integrante del dispositivo: l'intervento viene condotto secondo le modalità canoniche di una comunicazione scientifica, riproducendo intenzionalmente l'ambiguità tra reale e generato che costituisce il tema centrale della proposta.
Questa scelta metodologica si colloca all'interno di una più ampia riflessione sul public engagement nelle Digital Humanities e sul ruolo dell'AI nella costruzione di narrazioni scientifiche credibili, interrogando i confini tra informazione, dimostrazione e performance nel contesto della comunicazione accademica.
Viviamo in un'epoca in cui la tecnologia non solo riproduce la realtà, ma la genera, sfidando il nostro concetto di autenticità. Il mio intervento intende suscitare una discussione sul tema della autenticità generativa, secondo l'approccio di Walter Benjamin. Il concetto di autenticità è stato oggetto di numerosi dibattiti filosofici e culturali, proprio a partire dalla sua celebre *L'opera d'arte nell'epoca della sua riproducibilità tecnica* (1936). Benjamin sosteneva che la riproduzione tecnica alterasse l''aura' dell'opera d'arte, intaccandone l'unicità e l'autenticità. L'intelligenza artificiale, tuttavia, eleva questa riflessione a un livello successivo: non solo replica contenuti esistenti, ma è capace di generare opere originali, ridefinendo i paradigmi della creatività e della conoscenza (McCarthy, 2006). Questo tema è ulteriormente approfondito da Brynjolfsson e McAfee (2014), che evidenziano come le tecnologie emergenti stiano trasformando non solo il lavoro, ma anche i processi creativi e intellettuali.

6. LA QUINTA DIMENSIONE DEL REALE

C'è una quinta dimensione oltre a quelle che l'uomo già conosce; è senza limiti come l'infinito e senza tempo come l'eternità; è la regione intermedia tra la luce e l'oscurità, tra la scienza e la superstizione, tra l'oscuro baratro dell'ignoto e le vette luminose del sapere: è la regione dell'immaginazione, una regione che potrebbe trovarsi "Ai confini della realtà".

La domanda da cui prende avvio questa riflessione — quanto siano scientificamente credibili risultati ottenuti mediante intelligenza artificiale — si rivela, a uno sguardo più attento, parziale. Essa presuppone infatti una distinzione netta tra vero e falso che, lungi dall'essere messa in crisi dall'intelligenza artificiale, appare oggi semplicemente resa visibile nella sua natura problematica.
L'esperimento GENREAL, presentato nella prima parte di questo intervento come progetto empirico, è stato costruito deliberatamente come dispositivo di simulazione (credibile, per i motivi che illustrerò più avanti). La sua efficacia non risiede nella veridicità dei dati — che sono fittizi — ma nella sua capacità di aderire perfettamente alle forme discorsive della comunicazione scientifica contemporanea: struttura metodologica, lessico specialistico, apparato statistico, legittimazione istituzionale. In altri termini, il progetto funziona perché appare scientificamente plausibile.
Umberto Eco nella *Fenomenologia di Mike Bongiorno* afferma che: "Mike Bongiorno professa una stima e una fiducia illimitata verso l'esperto; un professore è un dotto; rappresenta la cultura autorizzata. È il tecnico del ramo. Gli si demanda la questione, per competenza." (Eco, 1994: 31) In questo senso, Mike

Bongiorno rappresenta, secondo Eco, una forma paradigmatica di cultura di massa, pertanto il senso comune va proprio in quella direzione: se lo dice un esperto allora deve essere vero. Il punto però è: chi legittima l'esperto come tale?
Questo dato suggerisce uno slittamento teorico rilevante: il criterio di validazione non coincide più con la verificabilità empirica, ma con la credibilità narrativa all'interno di un sistema condiviso di convenzioni epistemiche.
In questo senso, l'intelligenza artificiale non introduce una frattura radicale ma amplifica una condizione già presente. Nel campo delle scienze umane, la produzione di conoscenza è da sempre mediata da forme narrative che organizzano e rendono intelligibili i dati, piuttosto che derivarne automaticamente significati. La differenza rispetto alle cosiddette scienze dure non è semplicemente metodologica, ma epistemologica: ciò che viene riconosciuto come "vero" è spesso ciò che risulta coerente, persuasivo e contestualmente legittimato.
L'IA rende quindi evidente una questione più profonda: la distinzione tra vero e falso è un po' esasperata in certi contesti. Come nel testo della sigla di apertura della nota serie televisiva degli anni '50, *Twilight Zone* (*Ai confini della Realtà*), che cito in apertura dell'intervento, questa zona tra il vero e il falso è chiamata regione dell'immaginazione, postulando in qualche modo il fatto che l'immaginazione sia un misto di queste due cose, un po' come il romanzo storico di manzoniana memoria: un misto di storia e d'invenzione.
Insomma il discrimine netto tra Vero e Falso non esiste. C'è piuttosto una zona sfumata, una linea d'ombra abbastanza ampia, che lascia spazio all'ermeneusi sulla credibilità e/o veridicità.
Il problema è evidenziato dall'uso delle IA ma non è con esse che nasce: nelle scienze umane (differentemente dalle scienze dure) la dimensione narrativa gioca un ruolo strutturale nei processi di costruzione del sapere in luogo della dimostrazione scientifica o della disamina puntuale di cause ed effetti (giacché questi nel campo umanistico non sono normalmente dati algoritmicamente). Questo modo di vedere la cosa dipende fortemente dalla struttura con la quale funziona la ricerca scientifica.
Ne è la riprova l'affare Sokal: Alan Sokal - professore di fisica alla New York University - voleva dimostrare che, in certi ambiti delle scienze umane (come il postmodernismo e il post-strutturalismo), bastava usare un linguaggio oscuro, citare gli autori "giusti" (come Deleuze, Guattari o Lacan) e sostenere tesi politicamente corrette per essere pubblicati, anche scrivendo assurdità scientifiche. Molti intellettuali, tra cui Jacques Derrida, accusarono Sokal di essere un "furbastro" che non aveva capito nulla della filosofia che criticava. Sostennero che la scienza non è l'unico modo per interpretare il mondo e che Sokal aveva semplicemente abusato della fiducia dei redattori. L'evento diede il via a un dibattito feroce tra chi difendeva il realismo scientifico (la realtà esiste indipendentemente da noi) e chi sosteneva il costruttivismo sociale (la realtà è una costruzione culturale).
L'"affare Sokal" rappresenta un precedente significativo: la pubblicazione di un articolo deliberatamente privo di senso, ma costruito secondo i codici linguistici e teorici di un determinato ambito accademico, ha mostrato come la riconoscibilità formale possa prevalere sulla consistenza contenutistica (Sokal, 1996). Analogamente, il caso recente di Francesca Gino evidenzia come anche nei contesti più rigorosi la validazione scientifica possa essere compromessa da dinamiche sistemiche legate alla produzione e alla pressione dei risultati. Francesca Gino, nota professoressa italiana di etica e psicologia economica alla Harvard Business School, alla quale è stata revocata la cattedra dopo un'indagine interna che ha riscontrato manipolazioni dei dati scientifici nelle sue ricerche. Francesca Gino è stata accusata di aver falsificato o manipolato i dati in almeno quattro dei suoi studi, in particolare alterando i risultati di questionari somministrati per sostenere le proprie ipotesi di ricerca. L'inchiesta è scattata nel marzo 2023 dopo la segnalazione di alcune anomalie statistiche nei dataset da parte di tre blogger (Data Colada). Harvard ha riconosciuto una "sofisticata manipolazione" dei dati. La ricercatrice era considerata un'esperta di comportamento etico e disonestà, autrice del best seller "Talento ribelle: perché conviene rompere le regole al lavoro e nella vita".
Tuttavia non è solo una questione di 'etica', semmai c'è da chiedersi se l'Etica c'entra oppure no e in che misura. Il problema del Vero/Falso o del Simulacro/Non-simulacro non è solo delle IA. C'è un sistema complessivo in cui la ricerca è diventata un apparato produttivo che ha bisogno di risultati. In questo senso il 'non-autentico' non esiste, si tratta di convenzioni sociali.
Baudrillard, in *Simulacri e simulazione* (Baudrillard, 1994) apre il testo parlando della mappa e dice che non è più il territorio a precedere la mappa, ma la mappa precede il territorio. Il dato reale che noi rappresentiamo esiste solo in seconda analisi, principalmente esistono modelli che producono ciò che noi

chiamiamo la realtà. Il progetto GenReal si configura esattamente in questi termini (anche volutamente nel nome), un modello accademico che produce il reale (cioè il progetto).
In questa prospettiva, la distinzione tra autentico e non autentico si rivela meno ontologica di quanto si sia portati a credere, configurandosi piuttosto come esito di pratiche sociali e istituzionali. È in questo senso che la riflessione di Baudrillard sui simulacri appare particolarmente pertinente: non si tratta più di distinguere tra realtà e rappresentazione, ma di riconoscere un regime in cui la simulazione precede e struttura il reale: "It is the generation by models of a real without origin or reality: a hyperreal". (Baudrillard, 1994: 8). La simulazione è dunque la generazione di un reale senza origine e l'intelligenza artificiale – appunto generativa – si inserisce pienamente in questo regime, non come agente di falsificazione, ma come dispositivo capace di operare perfettamente all'interno delle sue regole. Essa non produce semplicemente contenuti veri o falsi, ma contenuti credibili.
Da questo punto di vista, la "regione dell'immaginazione" evocata in apertura — quello spazio intermedio tra luce e oscurità, tra scienza e superstizione — non è un'eccezione marginale, ma la condizione ordinaria entro cui si costruisce la conoscenza. L'IA non fa che renderla esplicita, mettendo in crisi non tanto la verità, quanto la nostra fiducia nei criteri attraverso cui la riconosciamo.

BIBLIOGRAFIA

Benjamin, W. (1936). L'opera d'arte nell'epoca della sua riproducibilità tecnica. Torino, Italia: Einaudi.

Baudrillard, J. (1994). Simulacra and simulation (S. Glaser, Trans.). University of Michigan Press.

Brynjolfsson, E., & McAfee, A. (2014). The second machine age: Work, progress, and prosperity in a time of brilliant technologies. New York, NY: W.W. Norton & Company.

Dennett, D. C. (2017). From bacteria to Bach and back: The evolution of minds. W. W. Norton & Company.

Eco, U. (1994). Diario minimo. Bompiani.

Floridi, L. (2014). The fourth revolution: How the infosphere is reshaping human reality. Oxford, England: Oxford University Press.

Harari, Y. N. (2018). 21 lessons for the 21st century. New York, NY: Spiegel & Grau.

McCarthy, J. (2006). What is artificial intelligence? Stanford University. http://jmc.stanford.edu/articles/whatisai/whatisai.pdf

Quattrociocchi, W. (2016). Liberi di crederci: Informazione, Internet e post-verità. Codice Edizioni.

Sokal, A. (1996). Transgressing the boundaries: Toward a transformative hermeneutics of quantum gravity. Social Text, (46/47), 217–252. https://doi.org/10.2307/466856

Tegmark, M. (2017). Life 3.0: Being human in the age of artificial intelligence. New York, NY: Alfred A. Knopf.

[1/B] Corpora digitali: requisiti e soluzioni gestionali

Verba volant, emails manent: una proposta di metodologia per l'acquisizione e la preservazione degli archivi di posta elettronica di personaggi illustri

ALBERTO ABIS[1], STEFANO ALLEGREZZA[2], FABRIZIO STUPINO[3]

[1]FONDAZIONE 1563 PER L'ARTE E LA CULTURA, ALBERTO.ABIS@FONDAZIONE1563.IT
[2]UNIVERSITÀ DEGLI STUDI DI MACERATA, STEFANO.ALLEGREZZA@UNIMC.IT
[3]UNIVERSITÀ G. D'ANNUNZIO CHIETI-PESCARA, FABRIZIO.STUPINO@UNICH.IT

ABSTRACT (ITALIANO)

Il contributo affronta il tema della preservazione degli archivi di posta elettronica come componente centrale degli archivi digitali di persona, con particolare riferimento ai fondi di personalità di rilievo pubblico. A partire dal caso dell'archivio di posta elettronica dell'onorevole Guido Bodrato, il contributo propone una procedura operativa per l'acquisizione, la messa in sicurezza e la gestione di grandi corpora di email, fondata sulla separazione tra oggetti informativi primari, rappresentazioni derivate e metadati strutturati di gestione. Viene illustrata la metodologia EMaG (Email Metadata-based Governance), che consente di garantire integrità, tracciabilità e analizzabilità del corpus senza interventi diretti sui messaggi originali. Il contributo presenta inoltre Epistula, un tool applicativo per l'analisi relazionale e contenutistica delle email basata su metadati e tecniche di NLP. Nella conclusione, il lavoro discute il ruolo della procedura come fase di pre-preservazione strutturata e il suo possibile inquadramento evolutivo nel modello EA-PDF.
Parole chiave: archivi di persona; archivi digitali; conservazione; posta elettronica; metadati

ABSTRACT (ENGLISH)

Verba volant, emails manent: a proposed methodology for acquiring and preserving the email archives of famous people

This paper addresses the preservation of email archives as a key component of personal digital archives, with specific reference to the records of prominent public figures. Drawing on the case study of Guido Bodrato's email archive, the contribution proposes an operational procedure for the acquisition, safeguarding, and management of large email corpora. The approach is based on a clear separation between primary information objects, derived static representations, and structured governance metadata. The paper introduces the EMaG (Email Metadata-based Governance) methodology, designed to ensure integrity, traceability, and analytical usability without altering original email files. It also presents Epistula, a dedicated tool supporting relational and content-based analysis through metadata-driven processes and NLP techniques. The paper concludes by framing the proposed procedure as a structured pre-preservation phase and discussing its potential evolution towards the EA-PDF archival model.
Keywords: personal archives; personal digital archives; preservation; email; metadata.

1. INTRODUZIONE

Negli ultimi decenni, la progressiva trasformazione digitale ha inciso in modo profondo sulle modalità di produzione, sedimentazione e trasmissione delle testimonianze documentarie personali. Gli archivi di persona digitali nativi – costituiti cioè da documenti creati originariamente in ambiente digitale e non derivanti da processi di digitalizzazione di supporti analogici – rappresentano oggi una delle principali sfide teoriche, metodologiche e operative per l'archivistica contemporanea. Essi comprendono una pluralità eterogenea di tipologie documentarie, quali i documenti creati con strumenti di elaborazione testi, i fogli di calcolo, le presentazioni, le fotografie digitali, i documenti sonori ed audiovisivi digitali, i messaggi di posta elettronica, i messaggi istantanei, i contenuti pubblicati o condivisi su piattaforme social, nonché i dati e i documenti prodotti da applicazioni e dispositivi personali. Nel loro insieme, tali materiali concorrono alla costruzione dell'identità digitale del soggetto produttore e costituiscono una fonte imprescindibile per la ricostruzione biografica, intellettuale, professionale e relazionale delle personalità contemporanee.

Tuttavia, a differenza degli archivi di persona tradizionali, fondati prevalentemente su supporti cartacei e caratterizzati da una relativa stabilità materiale, gli archivi digitali nativi si collocano in un contesto tecnologico instabile e in continua evoluzione (Guercio, 2016; Allegrezza, 2018). La loro esistenza è fortemente dipendente da infrastrutture hardware e software, da formati di file, da sistemi operativi e da piattaforme proprietarie che, nel tempo, diventano obsoleti o cessano di essere supportati. Tale dipendenza tecnologica introduce un elemento di fragilità strutturale che rende questi archivi particolarmente esposti al rischio di perdita, di inaccessibilità o di alterazione irreversibile. In assenza di interventi archivistici tempestivi e consapevoli, una parte significativa della memoria personale e collettiva del nostro tempo è destinata a scomparire. In aggiunta, a differenza degli archivi analogici, che spesso possono essere recuperati anche a distanza di anni dalla loro produzione, gli archivi digitali nativi sono soggetti a un rapido degrado funzionale: file non più apribili, supporti illeggibili, account disattivati, credenziali perse, servizi chiusi o trasformati. Il fattore tempo assume quindi un ruolo decisivo. Ritardare l'intervento significa aumentare esponenzialmente il rischio di perdita definitiva dei documenti e delle relazioni archivistiche che li legano tra loro.
A tale criticità si affianca il tema, sempre più rilevante, dell'eredità digitale (D'Arminio Monforte, 2020). Alla morte del soggetto produttore, l'accesso ai suoi archivi digitali è spesso ostacolato da vincoli giuridici, contrattuali e tecnologici: termini di servizio delle piattaforme, normative sulla protezione dei dati personali, assenza di disposizioni testamentarie specifiche, nonché difficoltà tecniche nel recupero delle informazioni, ad esempio dovuta all'impossibilità di acquisire le credenziali di autenticazione necessarie per accedere alle caselle di posta elettronica o agli spazi cloud. Infine, la produzione documentaria avviene spesso in modo inconsapevole, frammentato e distribuito su una molteplicità di dispositivi e servizi: computer personali, smartphone, tablet, cloud storage, piattaforme di posta elettronica e social network. A ciò si aggiunge la difficoltà di individuare confini archivistici chiari, in un contesto in cui i documenti sono frequentemente duplicati, sincronizzati o condivisi, e in cui la distinzione tra sfera privata e sfera pubblica risulta sempre più sfumata. L'assenza di pratiche di gestione documentale strutturate da parte dei soggetti produttori rende inoltre complessa qualsiasi azione di selezione, ordinamento e descrizione ex post.

2. STATO DELL'ARTE

La logica conseguenza di tutto ciò è che, in molti casi, gli archivi digitali di persona vengono cancellati automaticamente o rimangono inaccessibili, privando la collettività di una parte significativa del patrimonio documentario contemporaneo. Per storici, sociologi, studiosi della cultura, ma anche per giuristi e archivisti, tali archivi costituiscono una risorsa di primaria importanza, a condizione che ne sia garantita l'autenticità, l'integrità, l'affidabilità e la leggibilità nel tempo. La mancanza di una consapevolezza diffusa sul valore archivistico dei propri documenti digitali contribuisce ulteriormente a questa dispersione silenziosa della memoria. Negli ultimi anni, il tema del trattamento e della conservazione degli archivi digitali di persona sta emergendo come una questione di primaria rilevanza tanto per la comunità archivistica quanto per le istituzioni culturali e accademiche (Allegrezza, 2025). Tra i progetti più significativi a livello nazionale va segnalato quello di Pavia Archivi Digitali (PAD)[1], promosso dall'Università degli Studi di Pavia, che fin dal 2009 si propone di raccogliere e conservare archivi nativi digitali e documenti elettronici di personalità della cultura e della società contemporanee, garantendo la loro preservazione a lungo termine e l'accessibilità per la ricerca scientifica (Carbè, 2023). La Fondazione Arnoldo e Alberto Mondadori è attiva nella conservazione e valorizzazione di fondi personali e editoriali, promuovendo anche progetti interdisciplinari e momenti di riflessione su temi giuridici e deontologici, tra cui quelli legati alla privacy. Il progetto ALDiNa, nato in ambito AIUCD con il supporto di CLARIN-IT, riunisce studiosi di diverse discipline per definire metodologie e buone pratiche per la gestione e la conservazione degli archivi letterari nativi digitali. Presso l'Università di Bologna era operativo fin dal 2019 l'International Research Centre Personal Digital Memories, attivo nella ricerca sui modelli di creazione, gestione e conservazione dei fondi personali in ambiente digitale.
Tuttavia, la riflessione teorica non ha ancora sviluppato delle linee guida specifiche per il trattamento e la preservazione della componente digitale degli archivi di persona, come dimostrano le "Linee guida sul

[1] PAD, acronimo di "Pavia Archivi Digitali", è un progetto promosso dall'Università degli studi di Pavia. Nato nel 2009, si propone di raccogliere archivi nativi digitali e documenti in formato elettronico di personalità rappresentative della cultura e della società odierna, garantendone la conservazione nel tempo e la possibilità di studio, http://www-5.unipv.it/users/pad.

trattamento dei fondi personali" a cura della Commissione nazionale biblioteche speciali, archivi e biblioteche d'autore dell'Associazione Italiana Biblioteche (AIB) che non affrontano se non marginalmente il tema dei fondi di persona digitali (AIB, 2019).
Il trattamento degli archivi digitali di persona può essere letto anche alla luce degli studi socio-tecnici (Science and Technology Studies, STS), che interpretano gli archivi come infrastrutture di conoscenza in cui dimensioni tecniche, organizzative e sociali risultano profondamente intrecciate. In questo contesto, le operazioni di classificazione, selezione e descrizione non sono attività neutrali ma pratiche di mediazione che contribuiscono a definire l'oggetto documentario, i suoi usi possibili e le forme della sua accessibilità. In linea con il concetto di coproduzione di Sheila Jasanoff, la formazione e la preservazione degli archivi di posta elettronica sono intese come processi in cui pratiche tecniche, istituzionali e sociali contribuiscono congiuntamente alla costruzione del valore documentale (Jasanoff, 2004). Al contempo, come hanno documentato Susan Leigh Star e Geoffrey C. Bowker, gli archivi digitali sono analizzati come infrastrutture e sistemi di classificazione che organizzano la memoria, rendendo visibili alcune tracce e oscurandone altre
Come (Bowker & Star, 1999). Questa prospettiva consente di interpretare le email non come semplici documenti, ma come elementi inseriti in reti socio-tecniche che ne condizionano la selezione, l'accesso e la conservazione.

3. IL CASO DELL'ARCHIVIO DI GUIDO BODRATO

Tra i progetti più interessanti che stanno affrontando il tema del trattamento degli archivi personali digitali vi è quello legato alla figura dell'onorevole Guido Bodrato (27 marzo 1933 – 8 giugno 2023)[2]. Il suo archivio comprende, oltre alla documentazione cartacea conservata presso la Fondazione Carlo Donat-Cattin[3] (pervenuta in due momenti distinti, un primo versamento nel 2023 ed un secondo nel 2024), anche una quantità significativa di documenti nativi digitali. Si tratta di testi, fotografie, documenti audiovisivi prodotti e raccolti da Guido Bodrato sia nello svolgimento della sua attività professionale che durante momenti di vita familiare e personale. Tali documenti erano conservati nel computer personale di Guido Bodrato e sono stati acquisiti dalla Fondazione Donat-Cattin per donazione disposta dalla famiglia, nel rispetto della volontà dello stesso Bodrato. L'acquisizione è avvenuta mediante la creazione di una copia forense di tutto il materiale digitale conservato sul dispositivo. Tale operazione ha consentito di effettuare un backup completo e fedele della situazione originaria del fondo, preservando l'integrità dell'archivio e fotografando la situazione al momento del versamento.
Tuttavia, una componente molto importante dell'archivio digitale di Guido Bodrato, quella della posta elettronica, era sfuggita a questa acquisizione, in quanto non presente nel dispositivo utilizzato dal politico. Infatti, Bodrato non gestiva la posta elettronica attraverso un client (come Microsoft Outlook, Mozilla Thunderbird, ecc.) ma mediante la cd. webmail; di conseguenza i messaggi di posta elettronica non venivano trasferiti in locale ma rimanevano depositati nelle caselle di posta elettronica sui server dei provider di posta da lui utilizzati, e rischiavano, per i noti problemi dell'eredità digitale (D'Arminio Monforte, 2020), di andare perduti. Da qui la necessità di un intervento immediato e non procrastinabile.

4. L'ACQUISIZIONE DELL'ARCHIVIO DI POSTA ELETTRONICA

L'intervento di preservazione dell'archivio di posta elettronica di Guido Bodrato ha preso avvio nell'ottobre 2025 ed è stato condotto, sotto la supervisione di un comitato tecnico-scientifico, da un gruppo di lavoro

[2] Guido Bodrato, esponente della Democrazia Cristiana, ministro, deputato ed europarlamentare è stato un personaggio politico di rilievo nel secondo '900, inizialmente sul territorio torinese e successivamente a livello nazionale. Fu deputato alla Camera dal 1968 al 1994 per sette legislature, dopo essere stato consigliere comunale a Torino. All'interno della DC militò nella corrente di sinistra "Forze Nuove", insieme a Carlo Donat-Cattin, e fu stretto collaboratore di Benigno Zaccagnini. Nel corso della sua lunga carriera politica ricoprì più volte incarichi di governo: fu Ministro della Pubblica Istruzione (1980-1982), Ministro del Bilancio e della Programmazione Economica (1982-1983) e Ministro dell'Industria, del Commercio e dell'Artigianato (1991-1992), servendo in diversi esecutivi sotto primi ministri come Forlani, Spadolini, Fanfani e Andreotti. Dal 1996 al 1999 fu direttore del quotidiano "Il Popolo" e, dal 1999 al 2004, eurodeputato nelle liste del Partito Popolare Europeo. È ricordato per la sua coerenza, sobrietà e profondo attaccamento alle istituzioni e alla democrazia costituzionale.
[3] Dal 1992, anno della sua costituzione, la Fondazione Carlo Donat-Cattin è impegnata nella promozione di studi e ricerche sulla storia del pensiero sociale, politico ed economico del Novecento, con particolare attenzione al movimento cattolico e alla sua presenza in Italia e in Europa, e alla realizzazione di iniziative finalizzate all'analisi delle profonde trasformazioni sociali, economiche e culturali della nostra epoca. Attualmente ha sede a Torino all'interno del Polo del '900.

interdisciplinare composto da professionisti con competenze archivistiche, informatiche e di digital preservation. L'attività si è fondata sull'adozione di una metodologia sviluppata internamente, denominata EMaG (Email Metadata-based Governance), finalizzata alla messa in sicurezza, alla gestione tecnica e all'analisi degli archivi digitali di posta elettronica. EMaG si basa sulla netta separazione tra il corpus delle email originali, le versioni statiche derivate da esse e i metadati strutturati di gestione; non prevede interventi diretti sugli oggetti informativi primari — costituiti dai file EML e dai relativi allegati — ma opera esclusivamente attraverso strutture di metadati esternalizzate, progettate per garantire la tracciabilità delle operazioni, la ricostruzione del contesto informativo e l'automazione dei processi di analisi e gestione del corpus documentale. Le fasi in cui si articola tale metodologia sono descritte di seguito (vd. Fig. 1).

1) *Acquisizione controllata del corpus di email*. La prima operazione è stata la messa in sicurezza dei messaggi contenuti nelle caselle di posta elettronica di Guido Bodrato. Quest'ultimo aveva utilizzato nel corso della sua esistenza due account di posta elettronica: la casella di posta collegata al primo account, bodrato@alice.it, è stata correttamente acquisita grazie alla disponibilità delle credenziali di autenticazione fornite dagli eredi; al contrario, non è stato possibile recuperare il corpus di email collegate all'account guido.bodrato@gmail.com in quanto l'autenticazione richiedeva l'invio di un codice OTP ad un numero di telefono cellulare non più attivo, rendendo impossibile recuperare l'accesso alla casella di posta.

Il recupero delle email contenute nella prima casella di posta è stato effettuato utilizzando il client open source Mozilla Thunderbird, valutato come lo strumento migliore anche nei confronti di altre soluzioni disponibili sul mercato (come MailStore). Tale soluzione ha consentito l'acquisizione completa di tutte le email contenute nella casella di posta, complete di intestazione, corpo del messaggio e allegati. In questa fase non sono state svolte attività di selezione, riordino o descrizione archivistica; il corpus è acquisito integralmente, in un'ottica di preservazione tecnica preliminare volta alla salvaguardia dell'integrità informativa. Le email sono state acquisite nel formato EML, universalmente riconosciuto come formato idoneo per l'estrazione e il trasferimento della posta elettronica (Prom, 2021) e indicato anche come formato "interoperabile" dalle linee guida Agid (Agid, 2021). I file EML costituiscono il corpus di email originali e rappresentano la fonte informativa primaria dell'intero processo di preservazione.

2) *Produzione di copie statiche derivate*. A partire dal corpus di email originali, per ciascuna email è stata prodotta una versione statica utilizzano il formato PDF: ciascun messaggio è stato convertito in un documento PDF contenente il corpo del testo, le principali intestazioni e un elenco degli eventuali allegati. La conversione non ha comportato modifiche al contenuto informativo né trasformazioni degli allegati. Il documento PDF così ottenuto non è considerato oggetto di conservazione, ma una rappresentazione derivata funzionale alla consultazione e all'accesso, mantenendo il file EML come riferimento primario per la preservazione.

3) *Gestione degli allegati*. Gli allegati, se presenti sono stati trattati come oggetti digitali autonomi, semanticamente connessi al messaggio di posta. Essi sono preservati nel formato originale e non incorporati nel PDF. Per ciascun allegato è stata calcolata una impronta crittografica (utilizzando l'algoritmo SHA-256) da utilizzare come meccanismo di verifica dell'integrità nel tempo (Prom, 2021).

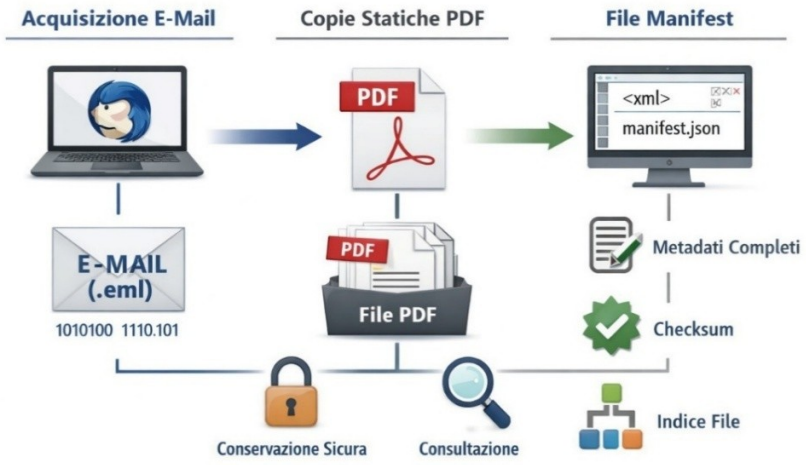


Figura 1: Le fasi operative della metodologia EMaG

4) *Creazione della struttura dell'archivio di posta elettronica*. Le email sono state archiviate mantenendo inalterata la struttura organizzativa presente nella casella di posta ("Posta in arrivo" e "Posta inviata"), con l'aggiunta di una ulteriore aggregazione, denominata "Archivio mail Bodrato", creata dagli eredi e contenente una selezione di messaggi inviati e ricevuti da Guido Bodrato. Le consistenze sono le seguenti: "Posta in arrivo": 11.364 messaggi, "Posta inviata": 249 messaggi, "Archivio mail Bodrato": 380 messaggi, per un totale di circa 9 GB.
5) *Produzione di un file manifest strutturato di gestione*. Elemento centrale dell'intervento è stata la produzione, per ciascuna aggregazione, di un file manifest in formato JSON per descrivere in modo esplicito e *machine-readable* sia i metadati relativi alle email che le relazioni tra file EML, versioni statiche PDF e allegati. Per ciascun messaggio di posta elettronica, il file manifest associa tra loro i principali metadati descrittivi (mittente, destinatari, data, oggetto, corpo della email), il percorso del file EML, il percorso del file PDF, l'elenco degli allegati con i relativi metadati tecnici e le impronte. Il file manifest non ha funzione descrittiva in senso archivistico, ma costituisce uno strumento tecnico-relazionale di controllo operativo del corpus, finalizzato alla tracciabilità, alla ricostruibilità del contesto informativo e all'automazione delle operazioni successive, come le analisi sul corpus, senza intervenire direttamente sugli oggetti email originali.
6) *Gestione del corpus e analisi delle relazioni*. La sistematizzazione organica del corpus di posta elettronica è esercitata principalmente a partire dal file manifest. Le operazioni di analisi, aggregazione e ricostruzione delle relazioni sono eseguite attraverso l'uso dell'applicativo *Epistula*, di cui si parlerà nel capitolo seguente.

In questo contesto, le eventuali operazioni di selezione, esclusione o marcatura di messaggi a bassa rilevanza documentaria non sono effettuate in modo automatico né discrezionale, ma si collocano all'interno di un processo strutturato di governance. Tali decisioni sono assunte da un gruppo di lavoro interdisciplinare, composto da archivisti e specialisti informatici – già coinvolto nelle fasi di acquisizione e messa in sicurezza del corpus – che opera tenendo in conto sia i principi archivistici, quali la pertinenza e il valore testimoniale, che le analisi computazionali basate su pattern ricorrenti e tecniche di classificazione automatica. In coerenza con la metodologia EMaG, tali operazioni non intervengono sugli oggetti informativi primari, ma sono applicate esclusivamente a livello di metadati all'interno del file manifest. Ogni decisione è tracciata e documentata, garantendo la trasparenza, la reversibilità e la verificabilità del processo, nel rispetto dei principi di integrità, autenticità e non alterazione.

5. EPISTULA

Al fine di semplificare e automatizzare le attività di gestione, analisi e strutturazione di ampi corpora di messaggi di posta elettronica acquisiti in formato EML, è stato sviluppato internamente l'applicazione Epistula. Essa opera a partire da file *manifest* strutturati di gestione consentendo di effettuare analisi relazionali, temporali e di tipo contenutistico attraverso un ambiente integrato e replicabile. Epistula incapsula e automatizza funzionalità tipicamente realizzate tramite script personalizzati, garantendo coerenza metodologica, tracciabilità delle operazioni e scalabilità su grandi corpora documentari. L'analisi dei contenuti è condotta mediante tecniche di Natural Language Processing (NLP) applicate tramite script in Python dedicati, ad esempio per l'identificazione automatica di classi di messaggi a bassa rilevanza documentaria (quali email di spam, comunicazioni seriali, messaggi augurali o notifiche ripetitive). Tali classi possono essere marcate o escluse dal perimetro di analisi intervenendo esclusivamente sui metadati contenuti nel file *manifest* strutturato. Questa impostazione consente di applicare analisi computazionali senza accedere ai messaggi originali, evitando modifiche al corpus di email originali e alle relative versioni statiche in formato PDF. La gestione del corpus avviene attraverso processi di aggregazione e clusterizzazione basati sui metadati e sul contenuto testuale, che permettono di costruire strutture relazionali e tematiche del corpus. Quindi, le operazioni di analisi e gestione del corpus avvengono senza interventi diretti sui messaggi o sugli allegati, garantendo la preservazione dell'integrità degli oggetti informativi.
Epistula consente di generare strutture derivate che rappresentano interazioni comunicative, sequenze temporali e raggruppamenti relazionali tra soggetti e messaggi. In particolare, è possibile effettuare raggruppamenti e segmentazioni del corpus basati su più livelli di analisi relazionale e contenutistica, tra loro integrati, come, ad esempio:

– relazioni dirette tra mittenti e destinatari, considerate come unità minime di interazione comunicativa;
– divisione e ricomposizione del corpus per corrispondenti, con distinzione tra profili di mittente, destinatario e co-destinatario, al fine di ricostruire direzioni, intensità e ricorrenza degli scambi;
– reti di corrispondenza tra gruppi di soggetti, derivate dall'analisi delle co-occorrenze nei campi di indirizzamento (To, Cc, Bcc);
– tipologie di argomento del messaggio, derivate dall'analisi del corpo testuale e dei campi oggetto.

In presenza di corpora documentari particolarmente estesi, Epistula può operare su strutture tabellari o basi di dati derivate dal file manifest, mantenendo invariato il principio di separazione tra oggetti informativi e strumenti di analisi, e preservando l'integrità dei corpora di email originali.
L'impiego dell'applicativo Epistula è finalizzato a supportare archivisti e ricercatori nelle attività di esplorazione, analisi e interpretazione del corpus di posta elettronica. Gli scenari d'uso tipici comprendono la ricostruzione delle reti di corrispondenza, l'analisi temporale delle comunicazioni e l'individuazione di cluster tematici basati sui contenuti e sui metadati. L'utilità dello strumento può essere valutata in relazione alla riduzione dei tempi di analisi, alla capacità di individuare relazioni significative e alla qualità delle aggregazioni prodotte: i primi riscontri evidenziano un miglioramento significativo relativamente alla leggibilità e interrogabilità del corpus di posta elettronica.

6. CONCLUSIONI E SVILUPPI FUTURI

Come si è già avuto modo di sottolineare nell'introduzione, gli archivi di persona digitali nativi rappresentano una componente essenziale del patrimonio documentario del XXI secolo, ma al contempo una delle sue parti più vulnerabili. La loro salvaguardia richiede interventi urgenti, competenze specifiche e un quadro teorico e normativo adeguato, capace di tenere insieme le esigenze della memoria, della tutela dei diritti e dell'accesso alla conoscenza. Ignorare o sottovalutare questa sfida significherebbe accettare una perdita irreversibile di fonti e di significati, con conseguenze profonde per la ricerca storica e per la costruzione della memoria collettiva delle generazioni future.
In questo scenario, la procedura di acquisizione e preservazione degli archivi di posta elettronica descritta nelle sezioni precedenti può essere considerata un primo passo significativo. Essa non costituisce un modello concluso, ma un'architettura preparatoria orientata a rispondere a esigenze operative immediate e suscettibile di evoluzione verso soluzioni di conservazione più formalizzate. In tale prospettiva, la recente proposta del formato EA-PDF (Email Archiving PDF) avanzata dall'EA-PDF Working Group (PDF Association, 2025) rappresenta uno scenario evolutivo coerente e potenzialmente integrabile.
La separazione tra corpus di email originali (file EML), versioni statiche (PDF) e metadati di gestione (manifest strutturato) consente infatti di preservare fin dalle fasi iniziali tutti gli elementi informativi necessari alla successiva costruzione di oggetti conformi alle specifiche EA-PDF. In particolare, la disponibilità del corpus di email originali – che costituiscono gli "original raw source email assets" secondo le specifiche del formato EA-PDF (PDF Association, 2025) – , unitamente alla presenza di metadati strutturati e verificabili, soddisfa uno dei requisiti fondamentali del modello EA-PDF, ovvero la possibilità di stabilire una relazione affidabile tra la rappresentazione PDF e la fonte email originaria (EA-PDF LWG, 2021). La metodologia proposta può, quindi, essere vista come una fase di pre-preservazione strutturata, durante la quale il corpus viene messo in sicurezza e reso governabile senza imporre una chiusura formale anticipata dell'oggetto di conservazione. Ciò consente di rinviare l'adozione di EA-PDF ad un momento in cui siano maturate adeguate condizioni istituzionali, archivistiche e tecniche
È importante sottolineare che l'eventuale integrazione di EA-PDF come scenario evolutivo non implica una sostituzione retroattiva della metodologia, ma piuttosto una sua estensione funzionale. Essa mantiene la propria validità come strumento di gestione, analisi e accesso al corpus, mentre EA-PDF interviene come ulteriore livello di formalizzazione, orientato alla conservazione a lungo termine. In tal modo, la metodologia si configura come un ponte tra esigenze operative immediate e modelli avanzati di conservazione digitale della posta elettronica, in grado di adattarsi a contesti istituzionali, volumi documentari e finalità differenti.

BIBLIOGRAFIA

Agid, Agenzia per l'Italia Digitale. (2021). Linee Guida sulla formazione, gestione e conservazione dei

AIB, Commissione nazionale biblioteche speciali, archivi e biblioteche d'autore (2019). Linee guida sul trattamento dei fondi personali. https://www.aib.it/documenti/linee-guida-sul-trattamento-dei-fondi-personali.

Allegrezza, S. (2018). L'impatto del digitale negli archivi privati: quale futuro per gli archivi di persona?. Atlanti, Volume 28(2018), Numero 1, pp. 79-90.

Allegrezza, S. (2025) Personal digital archiving: state of the art and future perspectives. The intangible papers. Authorial philology and born-digital texts. Antonelli G. et al. (Eds.). Bologna. Società Editrice il Mulino. pp. 15 – 39. ISBN: 9788815391704.

Bowker, G. C., & Star, S. L. (1999). Sorting things out: Classification and its consequences. MIT Press. https://direct.mit.edu/books/monograph/4738/Sorting-Things-OutClassification-and-Its

EA-PDF LWG (2021). Preserving Email (2nd edition). Data Types Series. Artefactual Systems and the Digital Preservation Coalition. https://www.dpconline.org/docs/technology-watch-reports/2472-preserving-email/file. http://doi.org/10.7207/twgn21-08.

Carbè, E. (2023). Digitale d'autore: macchine, archivi, letterature. Firenze, Siena: Firenze University Press; USiena Press, vol. 1, pp. 9-173 (ISBN 979-12-215-0023-3; 979-12-215-0022-6)

D'Arminio Monforte, A. (2020). La successione nel patrimonio digitale. Pisa: Pacini Editore.

documenti informatici. https://www.agid.gov.it/sites/agid/files/2024-05/linee_guida_sul_documento_informatico.pdf.

Guercio, M. (2016). Archivi personali: la sfida del digitale. Una riflessione. Gli archivi di persona nell'era digitale. Il caso dell'archivio di Massimo Vannucci. Allegrezza, S., Gorgolini, L. (Eds.). Bologna: Il Mulino, 2016, pp. 133-148

Jasanoff, S. (Ed.). (2004). States of knowledge: The co-production of Science and the Social Order. Routledge.

PDF Association (2025), EA-PDF, An archival email format based on PDF/A, p.8-10. https://pdfa.org/download-area/specifications/EA-PDF-v1.pdf

Prom, C. J. (2011). Preserving email. Digital Preservation Coalition Technology Watch Report 11-01, December 2011. https://www.dpconline.org/docs/technology-watch-reports/739-dpctw11-01-pdf/file.

Sfide alla sostenibilità dei progetti di digitalizzazione: i casi studio degli archivi manicomiali calabresi

EMANUELA NICOLE DONATO[1], MARIA NATALIA FEDERICO[2], GRAZIA SERRATORE[3]

[1]UNIVERSITÀ DELLA CALABRIA E IIT-CNR, RENDE (CS), ITALIA – EMANUELA.DONATO@IIT.CNR.IT
[2]IIT-CNR, RENDE (CS), ITALIA – MARIA.FEDERICO@IIT.CNR.IT
[3]UNIVERSITÀ DELLA CALABRIA E IIT-CNR, RENDE (CS), ITALIA - GRAZIA.SERRATORE@IIT.CNR.IT

ABSTRACT (ITALIANO)

I progetti di digitalizzazione rappresentano il primo passo nel percorso di valorizzazione e di ampliamento delle possibilità di accesso a un archivio storico in ambiente digitale. Tuttavia, le direttrici d'azione e gli esiti di ogni progetto dipendono dai contesti istituzionali e dalle risorse umane e finanziarie a disposizione, che possono incidere sulla possibilità di garantire un loro mantenimento sistematico. Negli ultimi decenni, in Italia, il crescente interesse per gli archivi storici manicomiali ha dato avvio a diversi progetti scientifici finalizzati al loro recupero, favorendo al contempo lo sviluppo di numerosi filoni di ricerca ad essi dedicati. Il presente contributo, pertanto, propone una riflessione su alcune delle principali sfide alla sostenibilità dei progetti di digitalizzazione, presentando inizialmente una panoramica dei più rilevanti interventi realizzati su archivi storici manicomiali a livello nazionale, per poi focalizzarsi in particolare su due casi studio: gli archivi storici dei manicomi calabresi di Girifalco (CZ) e Reggio Calabria. Lo studio prende in esame le esperienze maturate nell'ambito dei progetti ALPHA e PRIN PNRR 2022 DILIGO per l'archivio storico del manicomio di Girifalco e del progetto D'AMORE per quello di Reggio Calabria, soffermandosi sulle soluzioni metodologiche adottate, sulle difficoltà riscontrate e sui risultati raggiunti. Attraverso l'analisi di questi due casi di studio, il contributo intende mettere in luce alcune delle principali difficoltà e sfide connesse agli interventi di digitalizzazione, che rendono complessa la realizzazione di percorsi di valorizzazione coerenti e sistematici.
Parole chiave: digitalizzazione; archivi storici; ex ospedali psichiatrici; cartelle cliniche storiche

ABSTRACT (ENGLISH)

Challenges to the sustainability of digitalization projects: the case studies of the Calabrian psychiatric hospital archives. Digitalization projects represent the first step in the process of enhancing and expanding access to a historical archive within a digital environment. However, the directions of action and the outcomes of each project depend on institutional contexts and on the human and financial resources available, which may affect the possibility of ensuring their systematic maintenance over time. In recent decades, in Italy, the growing interest in historical psychiatric hospital archives has led to the launch of several scientific projects aimed at their recovery, while also fostering the development of numerous research initiatives dedicated to them. This contribution therefore offers a reflection on some of the main challenges to the sustainability of digitalization projects, initially presenting an overview of the most significant interventions carried out on psychiatric hospital archives at the national level, and then focusing in particular on two case studies: the historical archives of the Calabrian asylums of Girifalco (CZ) and Reggio Calabria. The study examines the experiences gained within the ALPHA and PRIN PNRR 2022 DILIGO projects for the historical archive of the Girifalco asylum and the D'AMORE project for the archive of Reggio Calabria, considering the methodological solutions adopted, the difficulties encountered and the results achieved. Through the analysis of these two case studies, the contribution aims to highlight some of the main difficulties and challenges related to digitalization initiatives, which make the implementation of coherent and systematic enhancement paths particularly complex.
Keywords: digitalization; historical archives; former psychiatric hospitals; historical medical records

1. INTRODUZIONE

L'entrata in vigore della Legge 180 del 1978, nota anche come Legge Basaglia, sancì la definitiva chiusura degli ex ospedali psichiatrici (OP) in Italia, determinando la loro progressiva dismissione. L'attuazione della riforma basagliana, demandata alle singole Regioni, fu priva di parametri comuni e si protrasse per

circa vent'anni, producendo situazioni disomogenee nella gestione dei patrimoni documentali di queste istituzioni. In seguito al progressivo esaurirsi delle funzioni dei propri soggetti produttori e conservatori (Reatti, 2023), alcuni archivi manicomiali furono trasferiti sotto la responsabilità delle Aziende Sanitarie Locali competenti, mentre altri rimasero abbandonati ed esposti a fenomeni di incuria o dispersione. Sul finire degli anni Novanta, grazie al progetto *Carte da Legare*,[1] avviato dalla Direzione Generale Archivi (DGA), si diffuse la crescente consapevolezza del valore storico, sociale e culturale di questi patrimoni documentali. Gli archivi storici manicomiali sono infatti beni culturali ai sensi del Decreto Legislativo 42/2004 - *Codice dei beni culturali e del paesaggio* (art. 10, comma 2, lettera b) e, poiché contengono al loro interno dati sensibili, il termine per la loro consultazione è di settant'anni (D.Lgs. 42/2004, art. 122), previa autorizzazione e nel rispetto di specifiche misure di riservatezza per motivi di pubblico interesse o per finalità connesse alla ricerca storica. Conseguentemente, divenne sempre più evidente l'esigenza di coordinare attività volte a garantirne il recupero, la conservazione e l'accessibilità nel lungo periodo, attraverso l'adozione di standard descrittivi condivisi. Nell'ambito del progetto *Carte da Legare*, dopo il censimento degli archivi manicomiali, fu avviata la loro messa in sicurezza dal punto di vista conservativo, promuovendone, quando necessario, il versamento negli Archivi di Stato, e attuando interventi di riordino ed inventariazione. È stata creata una piattaforma dedicata alla schedatura analitica delle cartelle cliniche, considerate fonti essenziali per ricostruire la storia istituzionale dei manicomi e le vicende umane dei ricoverati.[2] Va precisato, tuttavia, che questa si configura esclusivamente come uno strumento per il censimento dei dati, utile per studi statistici e comparativi, e non prevede la possibilità di consultare i contenuti delle cartelle cliniche. Accanto a questa iniziativa sono nati altri progetti di rilievo nazionale, tra cui il portale *Spazi della follia* [3] *e* il PRIN 2022 *Narration and Care. The deinstitutionalization of the asylum system in Italy: history, imaginary, planning (from 1961 to today)*.[4] Il progetto *Spazi della follia,* finanziato dal Ministero dell'Istruzione, dell'Università e della Ricerca (MIUR) nell'ambito del programma PRIN 2008, nasce da un accordo tra la DGA e una rete nazionale di Atenei per favorire la conoscenza e la valorizzazione del patrimonio storico-architettonico di ciascun complesso manicomiale sul territorio nazionale, restituendo dati storici sull'origine e sulle trasformazioni edilizie degli stabilimenti, che sono desunti da fonti a stampa coeve e da fonti archivistiche di diversa provenienza. Il PRIN 2022 *Narration and Care*, invece, ha l'obiettivo di valorizzare i dibattiti e le narrazioni sul tema della malattia mentale, alla luce di tre parole chiave: inclusione, deistituzionalizzazione e formazione. Nell'ambito di questo progetto, condotto in sinergia dalle Università di Cagliari, Torino e Pavia, è prevista la realizzazione di un atlante digitale, *Atlas of the Overcoming of Italian Asylums*,[5] che raccoglie informazioni sulle rappresentazioni degli ex OP italiani in diverse tipologie di fonti audio-visive, iconografiche e letterarie, nonché nella documentazione istituzionale e non istituzionale sulla storia dei manicomi. Nonostante la rilevanza di queste iniziative e il significativo impulso che esse hanno dato alla tutela degli archivi manicomiali, persistono limiti strutturali alla loro piena attuazione, poiché mancano le risorse economiche, tecniche e professionali necessarie a tradurre gli indirizzi metodologici forniti in pratiche operative stabili. La possibilità di proseguire e consolidare le attività avviate all'interno di questi progetti dipende, in larga misura, dalla concreta disponibilità delle istituzioni locali a partecipare ai processi di digitalizzazione degli archivi e di condivisione dei dati all'interno di piattaforme e progetti nazionali. Le attività di digitalizzazione si inquadrano generalmente all'interno di progetti scientifici specifici, sostenuti da finanziamenti dedicati, che prevedono tempi definiti per il conseguimento di più obiettivi circoscritti. Operare entro vincoli stringenti di tempo e di risorse può comportare, infatti, la necessità di selezionare diversi sub-corpora del patrimonio archivistico oggetto di studio, privilegiando determinate serie o tipologie di documenti a discapito di altre. Conseguentemente, nel caso in cui non sia possibile ottenere nuovi finanziamenti per proseguire le attività avviate, la conclusione di un progetto può determinare la creazione di archivi digitalizzati, trascritti e annotati solo parzialmente, con implicazioni significative in termini di rappresentatività, completezza e organicità del fondo (Campisi et. al, 2009). Lunghi intervalli tra un progetto di digitalizzazione e l'altro possono determinare l'obsolescenza delle soluzioni hardware e software già adottate, generando conseguenti condizioni di disomogeneità e problemi di continuità

[1] https://cartedalegare.cultura.gov.it/home (cons. 04/02/26)
[2] Gli archivi storici dei manicomi di Girifalco e Reggio Calabria sono parzialmente inventariati in Carte da Legare, grazie al lavoro di riconciliazione e riversamento dei metadati delle cartelle cliniche condotto dalla sede di Rende dell'Istituto di Informatica e Telematica del Consiglio Nazionale delle Ricerche (IIT-CNR).
[3] https://spazidellafollia.unicam.it/it (cons. 04/02/2026)
[4] https://prin.unica.it/de-asylum/ (cons. 04/02/2026)
[5] http://de-asylum-data.eu/ (cons. 05/02/2026)

operativa. Inoltre, va considerato che la conclusione di un progetto non esaurisce le esigenze di conservazione e valorizzazione dei documenti, sia analogici che digitali, ma apre, al contrario, a ulteriori questioni legate all'accessibilità, alla sostenibilità, e alla durabilità delle risorse digitali prodotte, al fine di prevenire l'obsolescenza tecnologica (Cucurullo et al., 2025). Pertanto, il presente contributo propone una riflessione sui problemi di sostenibilità dei progetti di digitalizzazione, presentando le sfide e le difficoltà incontrate nei casi di studio degli archivi storici dei due manicomi calabresi di Girifalco (CZ) e Reggio Calabria. L'analisi di queste esperienze consente di partire dalla descrizione di criticità riscontrate a livello locale per ricostruire dinamiche più ampie e trasversali, che caratterizzano in generale i processi di digitalizzazione del patrimonio archivistico. Il confronto tra progetti differenti, le soluzioni metodologiche adottate e i risultati ottenuti permette di individuare fattori strutturali che influenzano l'efficacia degli interventi avviati.

2. IL MANICOMIO DI GIRIFALCO

L'ex OP di Girifalco registrò il suo primo accesso il 22 luglio 1881 e rimase in attività fino all'entrata in vigore della Legge 180/1978 che ne determinò la progressiva dismissione (Marcello, 1995). L'archivio storico, conservato presso l'edificio originario, custodisce 15.794 cartelle cliniche, testimonianza del percorso terapeutico e del vissuto dei pazienti (Chiaravalloti & Taverniti, 2021). Ad oggi i documenti dell'archivio storico di Girifalco sono parzialmente digitalizzati grazie alle attività condotte dai progetti ALPHA e DILIGO. ALPHA (eAsy inteLligent service Platform for Healthy Ageing) è stato avviato nel 2014 nell'ambito del progetto "Invecchiamento: innovazioni tecnologiche e molecolari per un miglioramento della salute dell'anziano", a cura dell'IIT-CNR, in collaborazione con l'ICAR (Istituto di Calcolo e Reti ad Alte Prestazioni) CNR di Napoli e il Centro Regionale di Neurogenetica di Lamezia Terme (CZ). Il progetto ha permesso di digitalizzare oltre 5.000 cartelle cliniche che coprono l'arco temporale 1881-1931, di cui le prime 550 (1881-1894) sono state anche trascritte mediante software di dettatura vocale. Inoltre, nell'ambito di questo progetto è stata siglata una convenzione con la DGA per inserire nella banca dati nazionale del progetto *Carte da Legare* i metadati relativi alle cartelle cliniche (Rinaldi et al., 2025). Nonostante i grandi sforzi profusi, nel 2018, al termine del progetto, risultava digitalizzato meno di un terzo del totale dei documenti. A distanza di dieci anni dal primo intervento, nel 2023, il processo di digitalizzazione dell'archivio storico dell'ex OP di Girifalco è stato ripreso grazie al progetto PRIN PNRR 2022 DILIGO (DIgital preservation, LInguistic analysis and valorization of historical archive of the former psychiatric hospital of GirifalcO), che ha pianificato l'acquisizione di ulteriori 3.000 cartelle cliniche relative agli anni 1932-1944. Il progetto è guidato dall'IIT-CNR in collaborazione con le Università di Foggia e di Napoli "Federico II", impegnate nell'annotazione e nell'analisi linguistica della documentazione clinica relativa ai periodi della Prima Guerra Mondiale (1914-1917) e del Ventennio fascista (1922-1944). Per entrambi i progetti le attività di digitalizzazione sono state svolte in loco in seguito all'autorizzazione della Soprintendenza archivistica, ai sensi del D.Lgs. 42/2004 (art. 21 commi 4-5), e grazie ad accordi di collaborazione con l'Azienda Sanitaria Provinciale (ASP) di Catanzaro, che ha permesso l'accesso ai locali dell'archivio. Le acquisizioni sono state effettuate con scanner a testina mobile ScanSnap SV600, che permette di digitalizzare i documenti riducendo al minimo la manipolazione degli originali cartacei. Per quanto concerne la conservazione, per ovviare all'obsolescenza dei formati, rispetto al JPG (Joint Photographic Experts Group) precedentemente utilizzato, la scelta è ricaduta sul formato PDF (Portable Document Format) che, riducendo i rischi di alterazione accidentale, assicura l'integrità dei dati e la loro conservazione a lungo termine. Affinché il materiale digitale conservato possa mantenere il suo valore nel tempo, sono state adottate strategie volte a garantire la sicurezza dei dati. I file sono stati salvati in doppia copia su due differenti hard-disk, creando copie multiple, backup e repliche distribuite, requisito fondamentale per mitigare la perdita di dati. L'adozione di una politica di disaster recovery, attuata fin dalle prime fasi del lavoro, ha contribuito, quindi, a rafforzare i processi di archiviazione nell'ottica della conservazione a lungo termine. Per quanto riguarda l'attività di trascrizione, nel corso del progetto ALPHA è stato impiegato il software di dettatura vocale Dragon v.5, che tuttavia aveva evidenziato scarsa accuratezza delle trascrizioni ottenute rispetto alla specificità del corpus. Per questo motivo, nelle fasi iniziali del progetto DILIGO sono stati testati ulteriori strumenti di dettatura vocale[6] e strumenti automatici, basati su tecniche di Handwritten Text Recognition.[7] Questi ultimi, tuttavia, hanno rivelato

[6]Tra i software testati rientrano Web Speech API, Microsoft Dictation, Dictation.io.
[7]Transkribus ed e-Scriptorium.

forti limiti nel riconoscimento di grafie complesse ed eterogenee presenti nei documenti, generando un elevato numero di errori che avrebbero, di conseguenza, richiesto considerevoli operazioni di revisione, compromettendo efficienza e sostenibilità del lavoro. Il protocollo operativo è stato quindi ridefinito privilegiando la trascrizione manuale. Considerati i tempi necessari per la realizzazione delle trascrizioni e il progressivo aumento del numero e della consistenza delle cartelle cliniche per il periodo 1917-1944, è stato adottato un campionamento mirato sul corpus, prevedendo una trascrizione parziale ma rappresentativa dei periodi in analisi secondo gli obiettivi scientifici di progetto. In particolare, sono state trascritte 50 cartelle cliniche (25 uomini e 25 donne) per anno nel biennio 1917-1918, e poi è stato effettuato un campionamento quinquennale per il periodo della dittatura fascista trascrivendo 50 cartelle cliniche per gli anni 1922, 1927, 1932, 1937 e 1942, in modo da rispettare le tempistiche di progetto considerata la vasta mole documentale, consistente in oltre 7.700 unità. La necessità di effettuare le operazioni di campionamento sul corpus dei documenti è stata motivata non solo da vincoli temporali ma da una componente di intrinseca imprevedibilità: sebbene si conoscesse il numero complessivo delle cartelle cliniche da digitalizzare, non è stato possibile determinare a priori la quantità di fogli contenuti in ciascuna di esse, con il risultato che la pianificazione delle risorse e delle attività di progetto è stata definita solo in modo probabilistico. È stato, quindi, necessario ridefinire gli obiettivi di progetto al fine di garantire il raggiungimento dei risultati previsti, pur riducendo le dimensioni del campione di indagine inizialmente preventivato. Nonostante le acquisizioni realizzate congiuntamente dai due progetti ALPHA e DILIGO abbiano portato alla digitalizzazione di metà dell'archivio, il tempo intercorso tra i due progetti impone delle riflessioni circa la criticità connessa ai periodi di interruzione operativa prolungata, con delle inevitabili conseguenze sulla corretta conservazione dei documenti non ancora digitalizzati. Il rischio che ne deriva è quello di subire, nel frattempo, interventi di riordino e conservazione non idonei. L'auspicata continuità operativa risulta, tuttavia, ostacolata da barriere strutturali di vario genere come «finanziamenti inadeguati, mancanza di politiche nazionali per la conservazione digitale e assenza di infrastrutture tecniche» (Pandey & Kumar, 2020: 26). Pertanto, i finanziamenti dovrebbero prevedere risorse non solo per l'avvio delle attività ma per sostenerle fino alla loro conclusione (Pandey & Kumar, 2020). Ad oggi, la disponibilità di un corpus digitalizzato e trascritto ha permesso di condurre diverse ricerche multidisciplinari e multiprospettiche come studi di neurogenetica (Borrello et al., 2016), analisi sistematiche sul lessico e sulla terminologia psichiatrica dell'epoca (Dovetto, 2019; Chiaravalloti et al., 2020) e ricerche sociolinguistiche sull'eloquio dei pazienti (Taverniti et al., 2023). La schedatura delle cartelle cliniche all'interno del portale *Carte da Legare* e la possibilità di far riferimento al relativo schema di metadatazione rappresentano una garanzia di coerenza anche per attività di digitalizzazione non continuativa. Essendo il progetto DILIGO in fase conclusiva, emerge, quindi, la necessità di programmare nuovi interventi per la valorizzazione dell'archivio storico, tenendo conto delle circostanze locali, dei mezzi disponibili, e delle prospettive di ricerca, in linea con i principi di salvaguardia, conservazione e accessibilità della Carta UNESCO sulla Conservazione del Patrimonio digitale.

3. IL MANICOMIO DI REGGIO CALABRIA

L'ex OP di Reggio Calabria, la cui istituzione era stata sollecitata sin dal 1906, fu inaugurato nel 1932 al termine di un articolato iter istituzionale (Ministero per i beni e le attività culturali, 2010: 25). A seguito dell'entrata in vigore della Legge Basaglia, la struttura venne progressivamente dismessa fino alla definitiva chiusura nel 1992, mentre la gestione dell'archivio fu trasferita all'Unità Sanitaria Locale di Reggio Calabria. L'archivio del manicomio offre testimonianze biografiche di particolare rilevanza storica, tra cui si annoverano quella del brigante Giuseppe Musolino, noto alla cronaca giudiziaria del tempo, e di Angelo Callea, la cui biografia compare nel Dizionario biografico degli anarchici italiani.[8] Nel 2015 la Soprintendenza archivistica della Calabria ha autorizzato le attività di trasferimento, disinfestazione e messa in sicurezza della documentazione sanitaria (Rinaldi et al., 2025). Il trasferimento del materiale documentale presso i locali dell'IIT-CNR di Rende, consentito ai sensi del D.Lgs. 42/2004 - in particolare, dell'art. 21, comma 1, lettera e, relativo all'affidamento a terzi di complessi organici di documentazione appartenenti ad archivi pubblici - è stato reso possibile grazie ad una Convenzione per la temporanea custodia, siglata con il Dipartimento di Salute Mentale dell'ASP di Reggio Calabria, sede originaria dell'archivio. Il trasferimento si è reso necessario al fine di salvaguardare il fondo che, al momento del

[8]https://www.bfscollezionidigitali.org/collezioni/6-dizionario-biografico-online-degli-anarchici-italiani (cons. 28/01/2026)

recupero, versava in condizioni conservative particolarmente critiche, evidenziando la necessità di urgenti misure di tutela. Infatti, sin dal primo censimento effettuato nel 2000 dalla Soprintendenza archivistica, la documentazione risultava custodita in locali "carenti di attrezzature di controllo del clima, di rilevatori di fumo e di attrezzature antifurto".[9] Nell'ambito degli interventi di tutela del fondo, l'IIT-CNR ha inoltre investito ingenti risorse per dotarsi di attrezzature idonee alla salvaguardia dei supporti, provvedendo all'acquisto di un generatore di azoto, appositamente progettato per la disinfestazione dei documenti. Il trattamento in anossia è stato eseguito isolando i faldoni in una tenda a tenuta stagna satura di azoto, al fine di debellare le infestazioni da insetti presenti. I cicli di disinfestazione, dalla durata di 21 giorni ciascuno, hanno richiesto un tempo di lavorazione di sei mesi, comportando un significativo impegno di risorse umane, nonché la formazione specifica del personale per l'utilizzo delle attrezzature. Gli interventi di recupero condotti dall'IIT-CNR rappresentano pertanto un esempio significativo di come criticità strutturali derivanti da locali inadeguati alla conservazione a lungo termine possano essere affrontate mediante operazioni di messa in sicurezza della documentazione. Al termine di tali operazioni, si è proceduto all'avvio dell'attività di digitalizzazione, resa possibile a partire dal 2023 grazie al finanziamento da parte del Ministero della Cultura (MIC) – DGA del progetto di ricerca scientifica *D'AMORE* (Digitalizzazione dell'Archivio del ManicOmio di REggio Calabria). Nella loro fase iniziale, le operazioni di scansione – condotte mediante lo scanner planetario METIS EDS e lo scanner ad inserimento Fujitsu fi-6770[10] – hanno interessato i fascicoli dei pazienti trasferiti o provenienti dal manicomio di Girifalco, allo scopo di dare evidenza di come le storie dei due manicomi calabresi risultino intrecciate non soltanto per ragioni geografiche, ma anche in virtù delle dinamiche di ricovero, cura e gestione dei pazienti (Chiaravalloti et al., 2024). L'impiego di strumenti per scansioni ad alta definizione è stato reso possibile dallo svolgimento delle attività di digitalizzazione in un ambiente dedicato, a differenza del caso di Girifalco, in cui le operazioni sono state condotte in loco. Il progetto, conclusosi formalmente nel settembre 2025, ha reso possibile la conservazione in formato PDF di 100 registri e 570 cartelle cliniche, oltre che di 3.314 schedoni individuali, che si configurano come documenti contenenti dati riepilogativi riferiti ai pazienti ricoverati. I file digitalizzati sono conservati all'interno di volumi crittografati, al fine di garantire adeguati livelli di sicurezza e protezione dei dati sensibili. Tale soluzione si inserisce in una prospettiva di conservazione a lungo termine, assicurando l'integrità e la riservatezza della documentazione nel tempo. I criteri metodologici adottati per la digitalizzazione hanno incluso il riordino cronologico dei documenti all'interno di ciascuna cartella clinica. Questa scelta progettuale, pur garantendo una fruizione coerente e sistematica del materiale archivistico digitalizzato e rappresentando una buona pratica da adottare nell'ambito di progetti di digitalizzazione di archivi storici, ha dilatato sensibilmente i tempi impiegati nelle operazioni di scansione. Tra gli obiettivi previsti nell'ambito del progetto *D'AMORE* si segnala l'attività di metadatazione dei dati presenti all'interno delle cartelle cliniche dell'ex OP reggino, portata avanti grazie al più ampio progetto nazionale *Carte da Legare*. Per ottimizzare l'accesso e la fruizione dell'archivio e aprire la strada ad analisi automatizzate tramite tecniche di text mining, similmente a quanto fatto per DILIGO, pertanto, sarebbe auspicabile avviare la trascrizione delle cartelle cliniche, necessaria per la costruzione dell'edizione digitale dei documenti (Spina, 2023). Inoltre, la definizione di una espansione al set di metadati previsti da *Carte da Legare* potrebbe favorire il recupero di informazioni riferite a specifiche tipologie documentali, ad esempio i referti clinici. Attualmente solo una parte esigua dell'archivio risulta digitalizzata; pertanto, è indispensabile da un lato proseguire con le attività di digitalizzazione, e dall'altro procedere con la trascrizione e la strutturazione dei documenti già digitalizzati, al fine di supportare ricerche, interrogazioni semantiche e riuso scientifico dei contenuti (Barbuti 2019). Tuttavia, la realizzazione di tali interventi presuppone l'accesso a nuovi finanziamenti, attualmente necessari per sostenere futuri sviluppi relativi alla valorizzazione dei documenti dell'archivio storico dell'ex OP.

4. SFIDE DI SOSTENIBILITÀ

L'analisi dei casi di studio considerati ha permesso di sviluppare una riflessione sulle principali sfide alla sostenibilità dei progetti di digitalizzazione. I fattori di criticità emersi sono stati contestualmente

[9] https://siusa-archivi.cultura.gov.it/cgi-bin/siusa/pagina.pl?TipoPag=comparc&Chiave=157845 (cons. 05/02/2026)
[10] La scelta di differenti strumenti di acquisizione è legata alla variabilità dei supporti documentali.

ricondotti a specifiche aree di intervento, al fine di proporre potenziali direttrici d'azione utili a delineare un modello ideale di sostenibilità come illustrato nella Tabella 1.

AREA DI INTERVENTO	CRITICITÀ	IMPLICAZIONI OPERATIVE
Finanziamenti	Finanziamenti discontinui o insufficienti	• Ridotta capacità di pianificazione progettuale • Difficoltà nel garantire continuità dei processi
AREA DI INTERVENTO	**CRITICITÀ**	**IMPLICAZIONI OPERATIVE**
Governance	Frammentazione e/o assenza di politiche nazionali coordinate	• Difficoltà di standardizzazione di processi • Debole coordinamento e frammentazione decisionale
Infrastrutture tecnologiche	Carenza di strumenti e sistemi adeguati	• Ridotta efficienza dei processi • Ridotta qualità degli output
Accessibilità e fruizione	Accesso limitato o non strutturato ai dati digitalizzati	• Ridotto impatto scientifico e culturale dei risultati di progetto • Scarsa possibilità di riuso dei dati
Conservazione a lungo termine	Obsolescenza tecnologica	• Rischio di perdita o alterazione dei dati nel tempo
Sicurezza	Inadeguatezza delle misure di sicurezza e protezione dei dati	• Rischio di violazione dei dati • Vulnerabilità dei sistemi

Tabella 1. Implicazioni operative associate alle criticità rilevate, articolate per ciascuna area di intervento

Le criticità individuate non si configurano come elementi isolati, ma risultano strettamente interconnesse ed interdipendenti. Ad esempio, la discontinuità dei finanziamenti - legati a singoli bandi o iniziative progettuali - è spesso causa di interruzioni operative, determinando una digitalizzazione solo parziale degli archivi. In assenza di condizioni di conservazione adeguate, i materiali originali non digitalizzati rimangono esposti al rischio di progressivo degrado. Allo stesso tempo, il materiale acquisito richiede costanti azioni di mantenimento per ovviare al rischio di obsolescenza tecnologica, come migrazione dei formati, controllo di integrità dei file e gestione dei metadati. Ad oggi, molte criticità legate alla sostenibilità di questi progetti sembrano derivare da una mancata integrazione tra le diverse fasi del ciclo di vita dei dati. La pianificazione progettuale, pertanto, non dovrebbe essere circoscritta alla sola definizione degli obiettivi di progetto e dei workflow operativi ma dovrebbe includere, a monte, una valutazione strutturata del mantenimento dei dati in fase post-progettuale. Ciò implica la necessità di interrogarsi preventivamente su aspetti quali la manutenzione dei repository digitali e l'individuazione delle infrastrutture deputate alla conservazione dei dati digitali nel lungo termine. La long-term preservation richiede, infatti, l'abbandono di soluzioni di archiviazione locali e isolate in favore di infrastrutture cloud pubbliche, sicure e scalabili, conformi ai principi FAIR (Findable, Accessible, Interoperable, Reusable). L'integrazione degli archivi in ecosistemi digitali più ampi contribuisce a superare la frammentarietà che ancora, ad oggi, caratterizza molti progetti di digitalizzazione, evitando che iniziative metodologicamente solide rimangano isolate e difficilmente interoperabili nel lungo periodo. In parallelo, è auspicabile che sul piano della governance vengano sviluppate strategie orientate al coordinamento interistituzionale, alla puntuale definizione delle responsabilità tra i diversi livelli decisionali e all'adozione di standard procedurali condivisi. I processi di digitalizzazione, infatti, producono valore scientifico e culturale solo se gli output prodotti risultano effettivamente interrogabili e riusabili.

5. CONCLUSIONI

I progetti di digitalizzazione rappresentano il primo passo nel percorso di valorizzazione di un archivio storico. Tuttavia, i due casi di studio analizzati rivelano come strategie ed esiti dipendano in gran parte dai contesti istituzionali e dalle risorse disponibili, che ne condizionano anche la sostenibilità nel tempo. La definizione di linee guida realmente efficaci richiede di affiancare alle esperienze dei casi studio una revisione sistematica di un numero più ampio di progetti di digitalizzazione, anche relativi ad archivi

storici di diversa natura: solo attraverso una visione d'insieme più articolata si potrebbero elaborare indicazioni metodologiche applicabili in diversi contesti. Allo stesso tempo, la redazione di una documentazione delle scelte metodologiche e dei risultati ottenuti potrebbe garantire la replicabilità dei processi e l'adozione di buone pratiche comuni. L'esperienza maturata finora mostra come la disponibilità parziale di un archivio in formato digitale possa contribuire a ridurre le barriere che limitano l'accesso al patrimonio documentale, aprendo nuove prospettive di ricerca. È necessario, però, prestare attenzione al rischio di «selettività digitale» (Coburn, 2021: 398), per cui un archivio digitalizzato parzialmente può orientare gli studi verso i soli materiali disponibili. In conclusione, la digitalizzazione non può essere considerata un obiettivo finale, ma deve essere inserita all'interno di un modello più ampio che includa fin dall'inizio strategie di conservazione, accesso e riuso dei dati nel lungo periodo.

BIBLIOGRAFIA

Barbuti, N. (2019). Ripensare i formati, ripensare i metadati: Prove "tecniche" di conservazione digitale. *Umanistica Digitale, 3*(5), 121–138. https://doi.org/10.6092/ISSN.2532-8816/9055

Borrello, L., Cupidi, C., Laganà, V., Anfossi, M., Conidi, M. E., Smirne, N., Taverniti, M., Guarasci, R., & Bruni, A. C. (2016). Angela R.: A familial Alzheimer's disease case in the days of Auguste D. *Journal of Neurology*, *263*(12), 2494–2498. https://doi.org/10.1007/s00415-016-8294-x

Campisi, P., Maiorana, E., Teri, E., & Neri, A. (2009). Challenges to long term digital preservation: A glimpse of Italian experience. In *DSP 2009: 16th International Conference on Digital Signal Processing* (pp. 1-8). https://doi.org/10.1109/ICDSP.2009.5201056

Chiaravalloti, M. T., Taverniti, M., & Dovetto, F. M. (2020). Le cartelle dell'ex ospedale psichiatrico di Girifalco. Lessico, strumenti e terapie. In F. M Dovetto (A cura di). *Lingua e Patologia. I sistemi instabili* (pp. 235-268). Aracne. https://www.iris.unina.it/handle/11588/817038

Chiaravalloti, M. T., & Taverniti, M. (2021). Sanus egredieris. Storie di minori nel manicomio di Girifalco (1881–1894). *Mélanges de l'École française de Rome – Italie et Méditerranée modernes et contemporaines, 133*(1), 159–172. https://doi.org/10.4000/mefrim.10674

Chiaravalloti, M. T., Taverniti, M., & Dovetto, F. M. (2024). History and stories: Valorization of the documentary heritage of a former psychiatric hospital. *International Journal of Information Science and Technology, 8*(2), 11–20. http://dx.doi.org/10.57675/IMIST.PRSM/ijist-v8i2.290

Coburn, J. (2021). Defending the digital: Awareness of digital selectivity in historical research practice. *Journal of Librarianship and Information Science, 53*(3), 398–410. https://doi.org/10.1177/0961000620918647

Cucurullo, S., Montemagni, S., Saccoccio, R., & Sassolini, E. (2025). The Challenge of Obsolescence of Digital Archives in Cultural Heritage. A Case Study. In *2025 IEEE 8th Congress on Information Science and Technology (CiSt)* (pp. 506–512). https://doi.org/10.1109/CiSt65886.2025.11224142

Dovetto, F. M. (2019). I marginali dell'ex Ospedale psichiatrico di Girifalco e il lessico delle malattie di nerve alla testa. In D. Poli (A cura di), In limine: frontiere e integrazioni (pp. 137-161). Il Calamo.

Marcello, D. (1995). *Un secolo di manicomio: Storia del manicomio di Girifalco*. Vincenzo Ursini Editore. Ministero per i Beni Culturali e Ambientali (2010). *Primo rapporto sugli archivi degli ex ospedali psichiatrici*.

Gruppo di coordinamento del Progetto Nazionale "Carte da Legare". https://happylibnet.com/doc/473766/primo-rapporto-sugli-archivi-degli-ex-ospedali-psichiatrici

Pandey, R., & Kumar, V. (2020). Exploring the impediments to digitization and digital preservation of cultural heritage resources: A selective review. *Preservation, Digital Technology & Culture, 49*(1), 26–37. https://doi.org/10.1515/pdtc-2020-0006

Reatti, C. (2023). Archivi e spazi della follia fra reale e virtuale. I manicomi di Bologna e di Imola nel mosaico italiano. *Archivi*, 18(1), 145-176. https://www.anai.org/wp-content/uploads/2025/12/Anai_Archivi_1-2023.pdf

Rinaldi, M., Ranieri, L.M., Pani, E.M, Megna, A. & Pernati B. (2025). *Storie dalla rete scientifica*. CNR Edizioni.

Spina, S. (2023). Handwritten text recognition as a digital perspective of archival science. *AIDAinformazioni, 41*(1–2), 115–132. https://doi.org/10.57574/596529286

Taverniti, M., Chiaravalloti, M. T., & Dovetto, F. M. (2023). Uno sguardo sociolinguistico sui pazienti dell'Ospedale Psichiatrico di Girifalco. In F. Dovetto (A cura di), *Lingua e patologia. Parole dentro parole fuori* (pp. 227-265). Aracne. https://www.iris.unina.it/handle/11588/963427

A technical solution for a digital cultural heritage corpus: Notes on collaboration, requirements and sustainability

ARIANNA CIULA[1], MIGUEL VIEIRA[1], GEOFFROY NOEL[1], NEIL JAKEMAN[1], ZIHAO LU[1], TIFFANY ONG[1], SIMONA STOYANOVA[1], JONATHAN PRAG[2], ALESSIA COCCATO[2], RYAN HEUSER[3]

[1] KING'S COLLEGE LONDON
[2] UNIVERSITY OF OXFORD
[3] UNIVERSITY OF CAMBRIDGE

ABSTRACT (ENGLISH)

This submission presents the publication of a Digital Humanities project, *I.Sicily: Inscriptions of Ancient Sicily* (Prag ed., 2017-2025), an online corpus of cultural heritage inscriptions strongly influenced in its design and features by a previous web development effort (see acknowledgements). We provide an overview of the technical solution with associated public digital resource, while offering reflections on the collaboration which enabled the design, development of its technical solution, and public release. While the project is currently in its final stages of incremental development, this contribution focuses on its sustainable, modular, and open technical architecture. We demonstrate how Research Software Engineering (RSE) expertise – inclusive of design, analysis, development and maintenance – integrates with the project research questions. In this case, the RSE workflow supports epigraphy, palaeography, petrography and linguistic analysis of the digital corpus of inscriptions of ancient Sicily (from ca. 7th century BCE to ca. 7th century CE).
Keywords: research software engineering; digital scholarly annotations; technical requirements; open research.

ABSTRACT (ITALIANO)

Una soluzione tecnica per un corpus digitale del patrimonio culturale: note sulla collaborazione, le specifiche e la sostenibilità
Il contributo presenta la pubblicazione di un progetto di Digital Humanities, *I.Sicily: Inscriptions of Ancient Sicily* (Prag ed., 2017-2025), un corpus online di iscrizioni del patrimonio culturale fortemente influenzato, nel suo design e caratteristiche da un lavoro di sviluppo web precedente (*acknowledgements*). Offriamo una panoramica della soluzione tecnica con associata risorsa pubblica digitale e allo stesso tempo proponiamo delle riflessioni sulla collaborazione che ha reso possibile il design, lo sviluppo della soluzione tecnica, e la pubblicazione. Essendo il progetto è nelle sue fasi finali di sviluppo incrementale, questo contributo si focalizza sulla sua architettura tecnica sostenibile, modulare e aperta. Dimostriamo come la competenza di ingegneria software per la ricerca (RSE) – che comprende design, analisi, sviluppo e manutenzione – sia integrata con le domande di ricerca del progetto. In questo caso, il flusso di lavoro RSE supporta epoigrafia, paleografia, petrografia e analisi linguistica del corpus digitale delle inscrizioni della Sicilia antica (tra il VII secolo a.C. e il VII secolo d.C. circa).
Parole chiave: ingegneria software per la ricerca; annotazioni scientifiche digitali; specifiche tecniche; ricerca aperta.

1. STATE OF THE ART

While unfeasible to offer an in-depth state of the art here, we refer to relevant literature and developments across multiple axes of research in order of importance for this paper.
The international community is converging around the recognition of what in the United Kingdom have been defined as digital Research Technical Professionals (dRTP) – a plethora of software, data and compute infrastructure roles contributing to the creation of robust technical solutions for digital (open) research across disciplines and sectors. The consolidation of dRTP roles and practices is also valid for DH. The contribution of research software engineering design and development expertise to the scaling up of DH workflows and technical solutions is increasingly recognised (DHTech Group, 2019; Ermolaev et al., 2024). Core technical partner in the project presented here, King's Digital Lab (KDL) identifies itself as an RSE team and contributed extensively to both the systematisation of RSE practices in DH as well as the

professionalisation of roles and career pathways (Smithies & Ciula, 2021; Ciula et al. 2023; Ciula, 2025a; Ciula et al., 2025). This paper aligns with this development contextualising the RSE contribution within a digital cultural heritage project.
While complex and sophisticated in its data processing and workflows, the technical solution presented here around a static site generator (SSG) is not per se innovative; the trend of using SSG in web development has risen over the last ten years at least. However, a mirrored trend in DH stems from the discipline reaching a state of maturity where experimentation is counterbalanced by a duty of care and a holistic understanding of sustainability transversal to human and technical elements (Nowviskie, 2015; Smithies & Ciula, 2021; Tucker, 2022). The solution presented here aligns therefore both with DH and wider trends in the development of static websites following web development sustainability strategies and minimal computing philosophy. KDL evolving static-first approach is informed by its internationally recognised archiving and sustainability programme (Smithies et al., 2019; Ciula & Smithies, 2023; Ciula, 2025b) and inspired by comparable international efforts (Team T.E.P.).
As demonstrated by recent community efforts to document and critique these practices "static websites are increasingly recognized as a sustainable solution for digital editions and other scholarly publications in the Humanities" (CFP: Code4Lib Special Issue). With respect to digital editions building on TEI XML EpiDoc standard (*EpiDoc: Epigraphic Documents in TEI XML*), as outlined further in the paper, the solution presented here not only builds on the wider EpiDoc standard, but also uses the community stylesheets (Elliott et al., 2008-2021) with versioned customisation to suit the complexity of the I.Sicily multiple annotations and sophisticated user interface. While there exist tools in DH for annotating XML files and publishing TEI-based editions (most notably *Recogito studio* and *TEI Publisher*), these were not deemed suitable to the dense EpiDoc markup of the I.Sicily project. A more specific publication solution built around EpiDoc-encoded texts is EFES (*EpiDoc/EFES*) used in other projects hosted in KDL infrastructure (e.g. Reynolds et al., 2021). Widely adopted by the EpiDoc development community and used by several EpiDoc corpora, EFES processes EpiDoc XML for display and generates epigraphic indices. It is based on the Kiln platform developed for publication of more generic TEI XML projects at the Department of Digital Humanities and KDL. While the solution presented here forked EFES modules, it strives for a much nimbler architectural infrastructure while at the same time also avoiding the dependencies embedded in short-life *ad hoc* solutions with very specialised user interfaces such as EDEp (*Editionstools für eine digitale Epigraphik*) or Patrimonium project (Dalla Rosa et al., n.d.). The EpiDoc standard comprises of the schema, guidelines and transformation stylesheets. Epigraphic corpora published or under development before the creation of the EFES platform tended to use Kiln or ExistDB, or similar, and not all existing projects had the capability to transition to EFES or any other particular platform, as publishing and maintenance still depend largely on institutional support. The EpiDoc community recognises the state of infrastructure readiness and support, and so allows for flexibility in data management and publishing. The EFES platform offers lower-effort publishing of smaller, student-lead, research projects, as well as larger epigraphic corpora with steady maintenance and institutional support. However, already established publications are not always ready or prepared to move their existing infrastructure to a different platform, or indeed may have needs which EFES does not currently support.

2. DATASET, TECHNICAL REQUIREMENTS AND INTEGRATED WORKFLOWS

The core datasets for the project relies on the EpiDoc TEI-XML standard. This is used to encode all the information about the inscriptions (just under 6,000 items on project completion) and the inscribed objects, as well as the actual text itself.
The elicitation of technical requirements for the project was a collaborative effort led by a research software engineering team with decades of experience in collaborative DH projects (Smithies & Ciula, 2021; Ciula & Smithies 2023) and the researchers who conceived the idea for the European Research Council Advanced Grant CROSSREADS (grant agreement no.885040) funding application building on previous work (Prag, 2021; Prag & Chartrand, 2019).
To sum up these were the high-level functional requirements (see snapshot in time in Fig. 1) agreed at the beginning of the collaboration in order of delivery rather than priority:

1. Digital petrographic analysis environment;

2. Static website (inclusive of visual design and in compliance with agreed digital accessibility requirements);
3. Inscription webpage (inclusive of aggregated data modelling across a diversity of annotations – archival, linguistic, palaeographic and petrographic);
4. Aggregated search page (inclusive of map interaction);
5. Secondary dynamic pages (around bibliographic and museum data);
6. Web visualisations;
7. Open data endpoints.

This project solution processes EpiDoc TEI XML files, enriches the original input corpus (I.Sicily, 2026) and presents it on a static website (requirement 2 above and discussed further below) to which the project team can add contextual pages by using markdown (*Markdown Guide*).

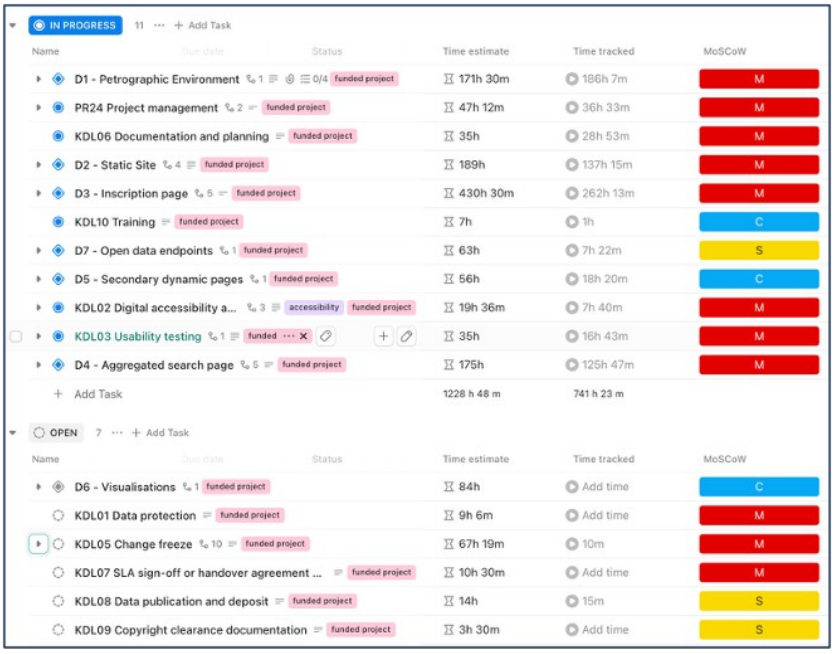


Figure 1. Requirements status and priorities at a given date

The solution uses a so called "monorepo" (Monorepo, 2025) structure with two main components: an ETL (Extract, Transform, Load) process for handling XML files, and a web application for presenting the processed data (King's Digital Lab, 2025b). Therefore, requirements 3-6 make use of the same dataset e.g. to pull a full edition of the inscription, directly encoded in EpiDoc, as well as to process other information on the inscriptions. These individual editions are published as HTML pages on the website but can also be searched and filtered through the website, as well as being freely available for download.
All editing, which takes place in an XML editor (*Oxygen*), is merged in GitHub (which provides a full version history) and the data can also be accessed through the public I.Sicily repository (Prag et al., 2024). High resolution images, where available, are also presented via a IIIF server (updates ongoing at the time of writing). The inscriptions bibliography is published in Zotero (*Zotero | Groups > ISicily*), but also in the FAIR Epigraphy bibliography (*Fair Bibliography*). Collection data (museums, archaeological sites, etc.) was originally gathered in a database of Sicilian museums and imported into the current solution. Data is standardised and made potentially interoperable by the use of recognised vocabularies, such as the Pleiades gazetteer of ancient places (Bagnall et al., 2006), the EAGLE epigraphic (*Vocabularies | EAGLE Portal*) and gradually the FAIR Epigraphy vocabularies (Heřmánková, Horster and Prag, 2025).
Linguistic annotation has been undertaken on a Greek and Latin subset of the corpus using Sketch Engine (Fendel, 2024), alongside tokenisation and lemmatisation of the complete corpus.
Palaeographic annotation has been undertaken through the development of a digital palaeographic environment with its own web application and interface (*Annotator*) also served and documented via Github (King's Digital Lab, 2026), which enables the assignment of letter typologies to linked images and texts. The environment lets the user define the palaeographic structure of allographs and then create thousands of Web Annotations (*Web Annotation Data Model*, 2017) that bind graphs as they appear on

the inscription images (using IIIF Image API region format) with their occurrences (using XPATH) in the EpiDoc edition (fetched from a DTS collection; *Distributed Text Services*) and the formal description of their structure. A faceted search interface allows the researcher to filter graph annotations by their structural patterns and define a high-level typology of allographs. The list of letter types identified for each annotated inscription are then added to the TEI file and so visible on the corpus website. The types list links back to the palaeography environment, allowing users to go back and forth between the main corpus pages and the more specialised palaeographic study. These types are also searchable from the 'Lettering' filter in the list of facet options, thus integrating them with the rest of the exploration possibilities within the entire corpus, even though only a selection of inscriptions underwent palaeographic analysis.

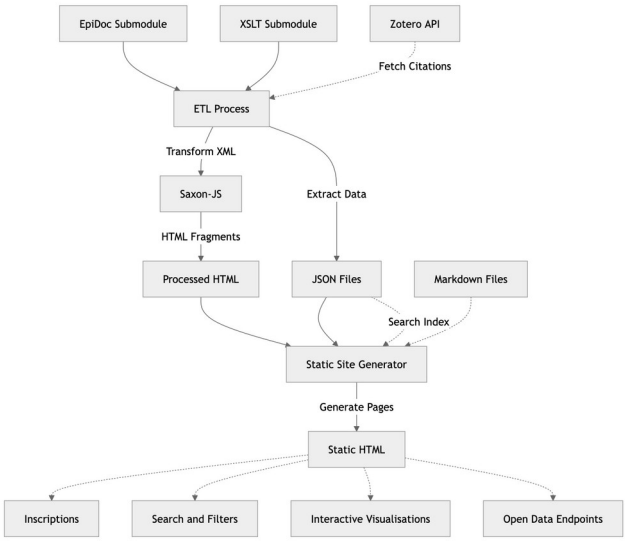


Figure 2. Corpus Building Workflow

Petrographic analysis has been undertaken on hundreds of items across the corpus. The petrographic dataset includes raw and processed data from (geo)chemical and minero-petrographic analyses on epigraphic supports and has been developed to support the identification of rock types. This latter is a fundamental achievement of the overall project, and geology-specific vocabularies (The BGS Rock Classification Scheme) are being taken into account to update the XML files with material and material provenance data. Materials provenance visualization options are being integrated in the public site, while the materials description has been significantly improved thanks to a dedicated and also KDL-supported sub-project.

The data from these three subprojects is being steadily integrated into the online corpus (see Fig. 2) and is being gradually released on the public site.

Each annotation layer – albeit via different data processing pipelines adapted to the relevant research workflows - feeds metadata into the inscription page (requirement 3), the faceted search (requirement 4) and other dynamic pages (in particular requirements 5-6).

The design of the site was informed by initial discussions around the information architecture and review of static mock-ups. A usability workshop with a group of representative prospective users conducted last year identified bugs as well as user interface and user experience refinements which have been prioritized collaboratively and integrated into the current interface which includes data visualisations (see Fig. 3).

The proposed contribution will provide an overview of the solution rather than dwell on the research, data and software workflow for each requirement. However, as an example of how the collaboration was undertaken integrating RSE expertise (DHTech Group, 2019) into the research objectives and in line with the conference thematic focus on interdisciplinary cultural heritage, requirement 1 would be examined in more detail. A core element of RSE support for this sub-project was indeed to aggregate the multi-

analytical data collected by the project's material scientist, to support (pre)processing and analysis of different data by creating tools to streamline repetitive tasks and simplify data interpretation, eventually allowing identification of the rocks where the texts (in Latin, Greek and other languages) are inscribed, and therefore their provenance. Ultimately the collaborative analysis process led to agreed RSE interventions in the research workflow formalised both by technical documentation (such as the diagram above), research dissemination (e.g. Coccato et al., 2025a; Coccato et al., 2025b; Ciula et. al., 2025; Ciula, 2025a) and, crucially from an RSE perspective, the actual code used to implement the petrographic research environment (King's Digital Lab, 2025a).

3. MORE ON TECHNICAL OVERVIEW AND SUSTANABILITY

The online corpus of inscription site is built with a Static Site Generator (SSG) called SvelteKit. This means it generates plain HTML, CSS, and JavaScript files that can be served by any standard web server (like Nginx or Apache). Therefore, it does not require a Node.js or any other application server to be running as it contains purely static assets.
Following RSE good practices, the site is available both as a so called staging instance (hosted in the KDL institutional infrastructure) to deploy updates and tests conducted by the project team in between solution increments, as well as a public production site (hosted by Oxford University infrastructure where ultimately the site will be maintained following the end of the funded period) also using Github actions and served via GitHub pages. The set of granular standard-compliant data files (DTS, EpiDoc, Web Annotations, etc.) versioned on and served from public code repositories is independent from and therefore can "live" without the software around it. To a certain extent it can also works as substitute to APIs. This portability and sustainability requirement was a consideration in all the design and development work including the other RSE components for the annotation subprojects (e.g. the palaeographic environment and the petrographic data processing pipeline).

4. COMMUNITY VALUE

By integrating different layers of annotations in the same publication, the site converges benefits from multiple communities, starting from the multidisciplinary researchers involved in the project (who are empowered to conduct data quality and analysis on the integrated corpus) to the museums and other repositories where the actual inscriptions objects are held (offering visibility to and enrichment to their collections under the same integrated digital space), and crucially including other collaborative DH teams who might want to adopt similar solutions for cultural heritage online corpus which aim to rely on open standards and open architectures.

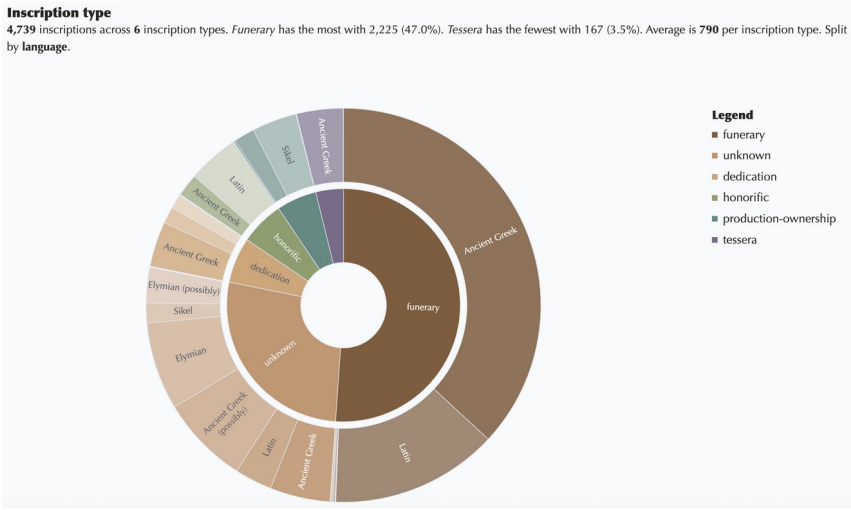


Figure 3. Visualisation of I.Sicily inscription by type and language

ACKNOWLEDGEMENTS

A previous instance of I.Sicily on which the requirements of the current solution is based was created and developed by Jonathan Prag, with the technical support of James Cummings and James Chartrand of OpenSky Solutions. In addition to this submission co-authors, other members who contributed to the

project work under its multiple subprojects and threads include: Tuuli Aalholm, Alex Antoniou, Imran Asif, Germana Barone, Chloe Colchester, Robert Crellin, Dmitry Dundua, Victoria Fendel, Ginestra Ferraro, Marta Fogagnolo, Rosemary Fox, Ilenia Gradante, James Hua, Estella Kessler, Andoni Llamazares, Martin Lopez Howe, Paolo Mazzoleni, Pam Mellen, Michael Metcalfe, Valentina Mignosa, Francesca Prado, Flavio Santini, Sophia Topf Aguiar De Medeiros, Alfredo Tosques, Claudio Vacanti, Valeria Vitale, students of the Universities of Palermo and Catania, and of the Liceo Lazzaro, Catania.

REFERENCES

Annotator. (n.d.). https://kingsdigitallab.github.io/crossreads/annotator.html

Bagnall, R., Talbert, R. J. A., Bond, S., Becker, J., Elliott, T., Gillies, S., Horne, R., McCormick, M., Rabinowitz, A., Robinson, E., & Turner, B. (2006). *Pleiades: A community-built gazetteer and graph of ancient places*. Institute for the Study of the Ancient World, New York University. http://pleiades.stoa.org

CFP: Code4Lib Special Issue: Static Websites for Scholarly Editions and Other Publications in the Humanities ← dh+lib. (n.d.). https://dhandlib.org/cfp-code4lib-special-issue-static-websites-for-scholarly-editions-and-other-publications-in-the-humanities/

Ciula, A. (2025a, October 31). *Role and responsibilities of the Research Software Analyst* | King's Digital Lab. https://kdl.kcl.ac.uk/blog/role-and-responsibilities-rsa/

Ciula, A. (2025b). Living well with brokenness in an inclusive research culture: What we can learn from failures and processes in a Digital Humanities lab. In A.-M. Sichani & M. Donnay (Eds), *Reframing Failure in Digital Scholarship* (pp. 71–85). University of London Press. https://doi.org/10.63674/yrab7358

Ciula, A., Callaghan, S., Caton, P., & Jakeman, N. (2025, December 17). *Aligning feasible technical requirements to research objectives and steering collaboration: Role and responsibilities of the Research Software Analyst in RSE teams*. RSE Conference 2025 (RSECon25). https://doi.org/10.5281/zenodo.17967411

Ciula, A., Noël, G., Caton, P., Ferraro, G., Ong, T., Smithies, J., & Vieira, M. (2023). The place of models and modelling in Digital Humanities: Some reflections from a Research Software Engineering perspective. In M. Annisius, E. Arestau, J. Burkhardt, N. Herold, & R. Sierig (Eds), *Vielfalt und Integration—Diversitá ed integrazione—Diversité et intégration* (pp. 261–281). Universitätsbibliothek.

Ciula, A., & Smithies, J. (2023). Sustainability and modelling at King's Digital Lab: Between tradition and innovation. In J. Nyhan, G. Rockwell, S. Sinclair, & A. Ortolja-Baird (Eds), *On Making in the Digital Humanities: Essays on the Scholarship of Digital Humanities Development in Honour of John Bradley* (pp. 78–104). University College London Press. https://doi.org/10.14324/111.9781800084209

Coccato, A., Hauser, R., Barone, G., Ciula, A., Mazzoleni, P., & Prag, J. (2025a). Petrographic data collection, processing and analysis: Minimal computational architecture for ancient Sicilian inscriptions. *Archéométrie 2025 XXVe Colloque Du GMPCA*, 313–314. https://gmpca2025.sciencesconf.org/

Coccato, A., Hauser, R., Monteiro Vieira, J. M., Barone, G., Ciula, A., Mazzoleni, P., & Prag, J. (2025b). Provenancing rocks used in ancient Sicilian epigraphy through a multi-analytical approach: Data collection, processing and analysis. In B. Carmina, L. Fascio, G. Innamorati, V. Marchionni, M. Pasero, & F. M. Petti (Eds), *Geosciences and the Challanges of the 21st Century* (p. 1128). Società Geologica Italiana ETS. https://doi.org/10.3301/ABSGI.2025.03

Dalla Rosa, A., Razanajao, V., & Baisson, C. (n.d.). *Patrimonium. Geography and Economy of the Imperial Properties in the Roman World*. https://patrimonium.huma-num.fr/atlas/editor/

DHTech Group. (2019). *DH Research Software Engineers—For We Are Many*. DHTech. https://dh-tech.github.io/dhrse-whitepaper/

Distributed Text Services (DTS). (n.d.). Distributed Text Services. https://distributed-text-services.github.io/specifications/

Editionstools für eine digitale Epigraphik. (n.d.). https://edep.adw.uni-heidelberg.de/index.html

Elliott, T., Au, Z., Bodard, G., Cayless, H., Lanz, C., Lawrence, F., Vandebilt, S., & Viglianti. (2008-2021). *EpiDoc/Stylesheets* (Version 9) [XSLT]. EpiDoc: Epigraphic Documents in TEI XML. https://github.com/EpiDoc/Stylesheets

EpiDoc/EFES. https://github.com/EpiDoc/EFES (Original work published 2015)

EpiDoc: Epigraphic Documents in TEI XML. (n.d.). https://epidoc.stoa.org/

Ermolaev, N., Naydan, M., Koeser, R., Wiering, J., Ciula, A., Mellen, P., Vieira, M., Czmiel, A., Maryl, Maciej, & Thomspon, L. (2024). The RSE Turn in Digital Humanities. In K. Ferguson (Ed.), *Book of Abstracts | Workflows: Digital Methods for Reproducible Research Practices in the Arts and Humanities* (p. 20). Zenodo. https://doi.org/10.5281/zenodo.13655424

Fair Bibliography. (n.d.). https://biblio.inscriptiones.org/

Fendel, V. B. (2024). The I.Sicily Sketch Engine Corpus. *Journal of Open Humanities Data*, 10(1). https://doi.org/10.5334/johd.258

Heřmánková, P., Asif, I., Horster, M., & Prag, J. R. W. (2025). From Fragmented Data to Linked History: Developing the FAIR Epigraphic Vocabularies. *Journal of Open Humanities Data*, 11, 74. https://doi.org/10.5334/johd.428

I.Sicily. (2026). *I.Sicily: the inscriptions of ancient Sicily*. https://github.com/ISicily/ISicily (Original work published 2017)

King's Digital Lab. (2025a). *Kingsdigitallab/crossreads-petrography*. https://github.com/kingsdigitallab/crossreads-petrography

King's Digital Lab. (2025b). *Kingsdigitallab/corpus-building*. https://github.com/kingsdigitallab/corpus-building

King's Digital Lab. (2026). *Kingsdigitallab/crossreads*. https://github.com/kingsdigitallab/crossreads

Markdown Guide. (n.d.). https://www.markdownguide.org/

Monorepo. (2025). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Monorepo&oldid=1318524755

Nowviskie, B. (2015). Digital Humanities in the Anthropocene. *Digital Scholarship in the Humanities*, 30(suppl_1), i4–i15. https://doi.org/10.1093/llc/fqv015

Oxygen XML Editor. (n.d.). https://www.oxygenxml.com/

Prag, Jonathan R.W. (ed.) (2017-2025), with A. Antoniou, A. Coccato, R. Crellin, V. Fendel, Marta Fogagnolo, I. Gradante, M. Metcalfe, V. Mignosa, F. Prado, F. Santini, S. Stoyanova, A. Tosques, King's Digital Lab (tech ed.), with A. Ciula, N. Jakeman, R. Heuser, Z. Lu, G. Noël, T. Ong, M. Vieira. *I.Sicily: Inscriptions of Ancient Sicily*. https://sicily.classics.ox.ac.uk. Stable deposit: https://doi.org/10.5281/zenodo.2556743.

Prag, J. R. W. (2021). I.Sicily and Crossreads: A digital epigraphic corpus for ancient Sicily. In C. Prescott, A. Karivieri, P. Campbell, K. Göransson, & S. Tusa (Eds), *Trinacria, 'An Island Outside Time'* (pp. 181–192). Oxbow Books. https://doi.org/10.2307/j.ctv24q4z4h.24

Prag, J. R. W., & Chartrand, J. (2019). 19 I. Sicily: Building a Digital Corpus of the Inscriptions of Ancient Sicily. In A. De Santis & I. Rossi (Eds), *Crossing Experiences in Digital Epigraphy: From Practice to Discipline* (pp. 240–252). De Gruyter Open Poland. https://www.degruyterbrill.com/document/doi/10.1515/9783110607208-020/html

Prag, J., Stoyanova, S., Mignosa, V., Antoniou, A. A., Chartrand, J., Neibart, S., Rocchetti, A., Llamazares, A., Gradante, ilenia, MhewittOx, Capra, G., Ahlholm, T., Barbieri, M., Prado, F., et al. (2024). *ISicily/ISicily: Work-in-progress*. Zenodo. https://doi.org/10.5281/zenodo.10966077

Recogito Studio. (n.d.). https://recogitostudio.org/

Reynolds, M., Roueché, C. M., Bodard, G., & Barron. (2021). *IRT. Inscriptions of Roman Tripolitania*. https://irt2021.inslib.kcl.ac.uk/en/

Smithies, J., & Ciula, A. (2021). Humans in the Loop: Epistemology & Method in King's Digital Lab. In K. Schuster & S. Dunn (Eds), *Handbook of Digital Humanities*. Routledge.

Smithies, J., Sichani, A. M., Westling, C., Mellen, P., & Ciula, A. (2019). Managing 100 Digital Humanities Projects: Digital Scholarship & Archiving in King's Digital Lab. *Digital Humanities Quarterly*, 13(1). http://www.digitalhumanities.org/dhq/vol/13/1/000411/000411.html

Team, T. E. P. (n.d.). *The Endings Project: The Endings Project | Le projet Endings*. https://endings.uvic.ca/

TEI Publisher Documentation and Demo. (n.d.). https://teipublisher.com/index.html

The BGS Rock Classification Scheme. (n.d.). *British Geological Survey*. https://www.bgs.ac.uk/technologies/bgs-rock-classification-scheme/

Tucker, J. (2022). Facing the challenge of digital sustainability as humanities researchers. *Journal of the British Academy*, 10, 93–120. https://doi.org/10.5871/jba/010.093

Vocabularies | EAGLE Portal. (n.d.). https://www.eagle-network.eu/resources/vocabularies/

Web Annotation Data Model. (2017, February 23). https://www.w3.org/TR/annotation-model/

Zotero | Groups > ISicily. (n.d.). https://www.zotero.org/groups/382445/isicily/

[1/C] Co-produzione di archivi e collezioni di comunità

L’Archivio Digitale di Vigne Nuove. Costruire un archivio di comunità, tra teoria e pratica

MANFREDI SCANAGATTA
OZ-OFFICINE ZERO, SEGRETERIA@OZOFFICINEZERO.ORG - UNIMORE, MANFREDI.SCANAGATTA@UNIMORE.IT

ABSTRACT (ENGLISH)

The Vigne Nuove Digital Archive. Building a Community Archive Between Theory and Practice
This paper reflects on the epistemic role of metadata in the construction of digital archives, with particular attention to archives developed within participatory and community-based contexts. Drawing on the case study of the digital archive on the history and memory of Vigne Nuove, developed as part of the We-Z project, the paper examines how archival practices inspired by the OAIS reference model can be adapted to manage heterogeneous sources, including born-digital materials, resources collected through crowdsourcing processes, and digital copies imported from other archives.
The contribution focuses on metadata as instruments of digital mediation, capable of supporting archival coherence and activating meaningful relationships among sources. In this framework, metadata creation is interpreted not merely as a descriptive practice, but as a process that shapes access, contextualization, and interpretation.
Finally, the paper proposes understanding semantic and conceptual metadata as epistemic devices through which the digital archive can be configured as an active space for the production of historical knowledge within Public Digital History. From this perspective, the paper aims to stimulate reflection on the development of flexible and replicable metadata models that may be adopted also in non-specialist contexts.
Keywords: Digital archives; Community archives; Conceptual metadata; Public Digital History

ABSTRACT (ITALIANO)

Il contributo propone una riflessione sul ruolo epistemico dei metadati nella costruzione degli archivi digitali, con particolare riferimento ad archivi sviluppati in contesti partecipativi e di comunità. A partire dal caso di studio dell’archivio digitale sulla storia e la memoria di Vigne Nuove, realizzato nell’ambito del progetto We-Z, il paper analizza come pratiche archivistiche ispirate al modello di riferimento OAIS possano essere adattate per gestire fonti eterogenee, native digitali, raccolte attraverso processi di crowdsourcing o importate in copia digitale da altri archivi.
L’attenzione è rivolta ai metadati come strumenti di mediazione digitale, capaci di sostenere la coerenza dell’archivio e di attivare relazioni significative tra le fonti. In questo quadro, la metadatazione viene interpretata non solo come pratica descrittiva, ma come processo in grado di orientare l’accesso, la contestualizzazione e l’interpretazione delle risorse.
Il contributo propone infine di considerare i metadati semantici e concettuali come dispositivi epistemici, attraverso i quali l’archivio digitale può configurarsi come uno spazio attivo di produzione di conoscenza storica nell’ambito della Public Digital History. In questa prospettiva, il paper intende stimolare una riflessione sulla costruzione di modelli di metadatazione flessibili e replicabili, potenzialmente adottabili anche in contesti non specialistici.
Parole chiave: Archivi digitali; Archivi di comunità; Metadati concettuali; Public Digital History;

1. IL PROGETTO

Ogni comunità custodisce archivi invisibili: storie, voci, emozioni che rischiano di dissolversi nel tempo. Costruire un archivio digitale di comunità significa trasformare queste tracce effimere in un patrimonio collettivo vivo e accessibile. Significa offrire strumenti per rigenerare non solo la memoria, ma anche il tessuto sociale di una comunità.

L'archivio sulla storia e la memoria del quartiere di Vigne Nuove[1] di Roma si inserisce nel quadro più ampio del progetto *We-Z emotional WEllbeing of generation Z: reconnecting communities and spaces through imperfect health*[2], cofinanziato con i fondi dell'European Urban Initiative.
Tra i partner figurano il III Municipio di Roma, Ater, Soprintendenza Speciale Archeologia Belle Arti e Paesaggio di Roma, ASL Roma1, il Dipartimento di Architettura dell'Università degli Studi di Roma 3, Avanzi, Solidarius, Parsec Coop., Parsec Agricultura, Oz-Officine Zero, Future Ecologies, Arte Studio.
Nelle linee guida della New European Bauhaus la valorizzazione del patrimonio culturale immateriale è indicata come azione strategica per attivare processi di rigenerazione di aree urbane[3].
Il quartiere di Vigne Nuove situato nella periferia nord di Roma, fu affidato nel 1972 dallo IACP di Roma a un gruppo di progettisti coordinato da Lucio Passarelli. Il quartiere fa parte di un piano di finanziamenti della GESCAL del 1969 che prevedeva la costruzione di alloggi e attrezzature collettive[4]. Il complesso, costituito da 524 alloggi per 3300 abitanti è oggi ancora segnato da anni di isolamento istituzionale, abbandono e degrado.
Le memorie degli abitanti rappresentano un patrimonio intangibile di grandissimo valore, capace tra le altre cose, di creare un nuovo significato sui quei luoghi.
La costruzione di un archivio che possa contenere le video interviste degli abitanti dell'area, realizzate con il metodo della storia orale, oltre a materiali raccolti attraverso pratiche di crowdsourcing (foto, diari, lettere) e documenti conservati in altri archivi, consente di dare forma ad un nuovo processo narrativo dell'area stessa, un racconto costruito dal basso, capace di discostarsi dalle narrazioni esterne, spesso pregne di preconcetti e semplificazioni. L'archivio vuole essere anche uno strumento di ricerca per tutti coloro che avranno interesse a scoprire e studiare la storia di Vigne Nuove.
Per la realizzazione è stato utilizzato Omeka S, una piattaforma open-source di web-publishing per la creazione e fruizione di collezioni digitali, sviluppata dal Roy Rosenzweig Center for History and New Media presso la George Mason University.[5].
In relazione alle prassi delle Digital Humanities e dell'archivistica digitale, da un punto di vista metodologico con questo lavoro vi è l'interesse ad aprire un focus sulle implicazioni date da un processo di metadatazione ricorsiva, così come descritto nella definizione di Oggetto Informativo (Michetti 2008) data per OAIS (Open Archival Information System)[6] e realizzabile tramite l'utilizzo dello standard Dublin Core Metadata[7], che consenta di creare una rete di relazioni dinamiche tra diverse tipologie di fonti primarie e secondarie.
Senza mai perdere di vista la necessità di una struttura solida per la conservazione di lungo periodo dei documenti, è nostro interesse creare delle connessioni di tipo concettuale tra tutti i diversi tipi di documenti, così da aumentare il portato informativo delle fonti stesse.
Stiamo sviluppando un modello che risponda alle stimolazioni della Public Digital History (Ravveduto, Salvatori, 2023) capace di favorire il lavoro interdisciplinare, che coinvolga una comunità di riferimento e che sia un modello logico e tecnologico facilmente replicabile.

2. METADATI. TRA STRUTTURA E PERFORMANCE

L'archivio si compone di cinque macro-insiemi di fonti, ciascuno caratterizzato da differenti modalità di produzione, conservazione e relazione archivistica:

- Fonti prodotte per l'archivio: video-interviste, fotografie, liberatorie, trascrizioni
- Fonti provenienti da archivi personali (crowdsourcing): fotografie, lettere, diari, contratti di locazione, ecc.
- Fonti conservate in altri archivi: Archivio ATER, Archivio Storico Capitolino, Fondo Passarelli – Archivio di Architettura MAXXI, Archivio del Dipartimento P.A.U.
- Fonti secondarie: bibliografiche, filmiche, musicali
- Fonti prodotte durante lo sviluppo del progetto con l'intenzione esplicita di generare nuovo patrimonio culturale

[1] https://portico.urban-initiative.eu/urban-stories/european-urban-initiative/fostering-community-cohesion-through-collective-memory-7159
[2] https://www.comune.roma.it/web/it/attivita-progetto.page?contentId=PRG1188650
[3] https://eur-lex.europa.eu/legal-content/IT/TXT/HTML/?uri=CELEX:52021XG1213(03)
[4] https://archidiap.com/opera/quartiere-di-vigne-nuove-piano-di-zona-n-7/
[5] https://omeka.org/s/
[6] http://www.oais.info/
[7] https://www.dublincore.org/

Questa eterogeneità rende necessaria una progettazione della metadatazione che non sia soltanto descrittiva, ma capace di sostenere strutturalmente e performativamente la relazione tra fonti, evitando sia la decontestualizzazione dei documenti sia l'appiattimento su un'unica logica archivistica.
Per elaborare la maschera di metadatazione da un punto di vista logico, i metadati sono stati clusterizzati in base alla loro funzione, assumendo come riferimento il modello OAIS (Open Archival Information System). Tale modello non viene applicato in senso letterale, ma utilizzato come infrastruttura concettuale certificata entro cui sviluppare forme di sperimentazione archivistica coerenti con archivi digitali nativi e ibridi.
In questa prospettiva, i cluster di metadati si ispirano alle logiche degli Information Package OAIS:

- Metadati descrittivi fondamentali – in relazione al livello di Submission Information Package (SIP)
- Metadati relativi all'accessibilità, all'uso e alla gestione – in relazione al livello di Archival Information Package (AIP)
- Metadati per la costruzione della struttura relazionale tra le fonti – ancora in relazione al livello AIP
- Metadati per l'organizzazione e la restituzione di relazioni di ordine concettuale – in relazione al livello di Dissemination Information Package (DIP)

L'adozione dello standard Dublin Core Metadata Initiative consente di mantenere la struttura dell'archivio flessibile e adattabile, permettendo l'aggiunta o la rimozione di campi in funzione delle esigenze descrittive e della fruizione anche da parte di pubblici non specialistici. Allo stesso tempo, Dublin Core garantisce una base comune per l'interoperabilità tra tipologie di fonti profondamente diverse.
Per questa ragione, non è stato definito un unico schema di metadati valido per tutte le risorse, ma sono stati elaborati schemi *ad hoc* per ciascuna tipologia di oggetto digitale. Tali schemi mantengono la distinzione logica tra SIP, AIP e DIP, ma modulano di volta in volta i campi rappresentati, così da valorizzare la natura specifica dell'oggetto senza appesantire il sistema informativo.
Metadatare una video-intervista, fonte prodotta sincronicamente alla costruzione dell'archivio, è strutturalmente e concettualmente diverso dal metadatare un documento conservato in un altro archivio. Nel primo caso, la fonte nasce già all'interno di un sistema digitale progettato; nel secondo, il documento è parte di un complesso archivistico preesistente, caratterizzato da un vincolo formatosi storicamente e indipendentemente dall'archivio digitale che lo intercetta.
Questa distinzione incide direttamente sui metadati riconducibili al livello SIP.
Nel caso di documenti provenienti da altri archivi, i campi *dcterms:creator* e *dcterms:publisher* fanno riferimento rispettivamente al soggetto produttore originario e all'archivio di conservazione. Nel caso delle fonti prodotte per l'archivio (video-interviste, liberatorie), tali campi rimandano invece al progetto stesso, in questo caso We-Z. Una situazione intermedia si riscontra nelle fonti raccolte tramite crowdsourcing, dove *creator* e *publisher* rappresentano rispettivamente il produttore originario del documento e il progetto We-Z come soggetto che ne cura l'integrazione archivistica.
Nei cluster riconducibili all'AIP assumono un ruolo centrale i metadati deputati alla costruzione delle relazioni tra le fonti, in particolare i campi *dcterms:IsFormatOf*, *dcterms:HasPart*, *dcterms:IsPartOf* e *dcterms:Source*. Per le fonti prodotte per l'archivio, tali campi sono indispensabili per stabilire relazioni tra video-interviste, trascrizioni, liberatorie (queste presenti ma non accessibili al pubblico), eventuali fonti di crowdsourcing riconducibili all'intervistato e materiali di contesto.
In questo caso il vincolo tra le fonti non si struttura in modo naturale come indicato dalla tradizione archivistica (De Felice, 1988), ma viene strutturato post-facto attraverso la metadatazione. Questo passaggio sappiamo non essere privo di criticità, ma è forse uno degli elementi che maggiormente definisce il concetto di archivio inventato. Assumendo che nei contesti digitali e comunitari il vincolo non possa essere originario, il modello sviluppato non intende sostituire né simulare il vincolo archivistico classico, ma attivarne una funzione contestuale e relazionale, capace di garantire coerenza senza compromettere l'essenza dell'archivio né negare i vincoli originari laddove esistenti.
Per metadatare fonti conservate in altri archivi si sono dunque usati quei campi per riportare e rendere leggibili i vincoli archivistici già descritti negli archivi di provenienza, nonché, ove necessario, a connettere tali fonti a materiali prodotti o raccolti dall'archivio digitale stesso. In questo modo, l'archivio non dissolve i contesti originari, ma li mette in relazione con nuove costellazioni documentarie.
Sempre nell'ambito dell'AIP, è stato necessario definire un Data Management Plan volto a chiarire la proprietà e la titolarità dei dati. Nel caso del progetto We-Z, le fonti prodotte durante il progetto sono caricate nell'archivio sulla base di un accordo di contitolarità tra OZ – Officine Zero, il Dipartimento di

Architettura di Roma Tre e il Terzo Municipio di Roma Capitale. Questa regola non si applica alle fonti preesistenti conservate in altri archivi, per le quali nei campi *dcterms:rights* e *dcterms:RightsHolder* sono state riportate le indicazioni fornite dagli enti di conservazione.
Un Data Management Plan non risponde esclusivamente alla necessità di chiarire la gestione dei diritti, ma assume un ruolo centrale in relazione alla conservazione nel medio e lungo periodo. La scelta della contitolarità tra più enti non è quindi soltanto giuridica, ma archivistica: affidare la titolarità dei dati al solo progetto o agli accordi di partenariato avrebbe infatti comportato il rischio che, una volta concluso il progetto stesso, le risorse digitali risultassero prive di un soggetto responsabile della loro conservazione. La contitolarità consente invece di distribuire nel tempo la responsabilità di conservazione, aumentando le possibilità di mantenimento, migrazione e accessibilità delle fonti anche oltre la durata formale del progetto.
L'ultimo insieme di metadati, quello relativo ad una descrizione semantica e concettuale degli Oggetti Digitali si ispira alle funzionalità del DIP, dove però, a livello sperimentale, abbiamo voluto forzare il modello cercando di valorizzare il portato epistemologico di DCMI.
È opinione comune che il Web rappresenti uno spazio di disintermediazione, riteniamo invece che questa affermazione sia non solo inesatta, ma fuorviante.
Il Web è fortemente mediato, ogni algoritmo incorpora decisioni umane, definisce cosa e come vediamo e riformula le possibilità di accesso all'informazione. I metadati stessi operano come strumenti di intermediazione digitale.
Le teorie e le pratiche del Web Semantico (Web 3.0), definiscono in modo chiaro il portato di mediazione semantica dei metadati. La costruzione di triple in RDF composte da soggetto, predicato e oggetto, e l'utilizzo di linguaggi ontologici come OWL, SPARQL, RIC-O, consentono alla macchina di acquisire capacità semantiche, cioè di interpretare le richieste in termini di significato piuttosto che di forma.
In questa prospettiva, risultano fondamentali gli studi e i progetti di Francesca Tomasi (2022), che, oltre a mettere in luce il portato rivoluzionario del Web Semantico, evidenziano la complessità logica e tecnica di un processo innovativo che richiede competenze di alto profilo.
Dal punto di vista della Public Digital History, ritengo utile la possibilità di strutturare processi meno complessi e più facilmente replicabili, capaci di costruire reti di relazioni concettuali tra le fonti, che seppur non raggiungono la sofisticazione semantica del Web 3.0 e non consentono alle macchine di comprendere le richieste, attivano una profonda significazione della fonte, generando una interoperabilità semantica capace di guidare il fruitore in una narrazione storica già all'interno dell'archivio. A partire dalla creazione del *Digital Object*, OAIS struttura un processo di metadatazione ricorsiva che genera *Representation Information*, ossia l'insieme di informazioni e metadati necessari perché l'unità informativa possa essere interpretata dalla *Designated Community* in relazione ad una *Knowledge Base* preliminarmente condivisa. uesto elemento consente di utilizzare dei termini descrittivi estrapolati dal contenuto del documento stesso e capaci di andare a costruire significazione sull'oggetto digitale così da renderlo un oggetto informativo in una relazione semantica con altri oggetti informativi.
In questa prospettiva, la metadatazione non rappresenta più soltanto una forma di descrizione archivistica, ma un vero e proprio atto interpretativo e si mostra ai public historians come una possibile innovativa forma di produzione storiografica, dove la fonte, per la prima volta emerge come elemento intellegibile del discorso storico.
Paul Weston e Lucia Sardo (2017) individuano i metadati come elementi con i quali poter anche attribuire significati estrapolandoli con attenzione dal contenuto o dal contesto delle fonti stesse. A questa riflessione si unisce anche Federico Valacchi che, pur sottolineandone alcune criticità fa emergere l'utilità di un processo di concettualizzazione della conoscenza che precede e orienta le pratiche di classificazione e diffusione. (Valacchi, 2024)
I processi di concettualizzazione della conoscenza attraverso l'analisi ermeneutica delle fonti sembrano poter avere un impatto determinante sulla costruzione e narrazione del discorso storico. Questa non è, tuttavia, una novità introdotta dal digitale, già Reinhart Koselleck (2007) riconosceva l'importanza dei concetti per comprendere e raccontare la storia.
I campi dcterms:spatial, dcterms:temporal, dcterms:description, nel nostro modello non hanno come unico scopo quello di descrivere i documenti, ma diventano vere e proprie chiavi di accesso alle fonti. Parole chiave, date e luoghi vengono estrapolate dalle fonti attraverso un'analisi critica e comparativa svolta con il coinvolgimento di più soggetti/curatori, e riattribuite ad esse in qualità di metadati attraverso un processo circolare. La ricorsività consentita da DCMI si mostra quindi non come elemento tecnico ma propriamente epistemologico e può rendere l'archivio un atipico oggetto storiografico che per essere

fruito richiede un atto di scelta da parte del fruitore. Un'azione performativa che determinerà la risposta del sistema e che mostrerà all'utente sempre diverse relazioni tra gli Oggetti Digitali.
Come osserva William Hersh (2024), i sistemi di intelligenza artificiale generativa hanno spostato l'attenzione dell'utente dalla ricerca delle fonti alla ricezione di risposte sintetiche. In contrapposizione a ciò che oggi è stato definito come "motori di risposta", con il nostro archivio e la sua funzione performativa si vuole creare un modello di funzionamento dove la fonte e l'azione di ricerca tornano ad essere centrali, ciò non di meno però l'AI ha avuto un ruolo nell'estrapolazione di concetti.
Nel lavoro di Wenli Yang et al (2025) si porta l'attenzione sull'impatto dell'AI nei processi di metadatazione, introducendo il concetto di *metadata intelligence*, in cui il metadato non è più solo il prodotto della catalogazione umana, ma un elemento co-costruito fra umani e macchine. I sistemi di intelligenza artificiale si collocano oggi simultaneamente su entrambi i versanti della metadatazione sia come produttori, nella misura in cui contribuiscono a generare o arricchire metadati, sia come fruitori, poiché la loro capacità di ricerca, correlazione e citazione delle fonti dipende proprio dai metadati stessi.
Nel nostro lavoro l'AI non sostituisce l'analisi ermeneutica, ma ne diventa un'estensione computazionale. Tale ibridazione tra intelligenza umana e macchina nella produzione e organizzazione dell'informazione, corrisponde a ciò che Luciano Floridi ha descritto come società *iperstoriche (*2017*),* società in cui la conoscenza viene prodotta, catalogata e archiviata attraverso l'interazione continua fra soggetti umani e sistemi computazionali automatizzati.
Ancora da implementare è un lemmario che funzioni come una mappa concettuale per navigare l'archivio. I metadati estrapolati per essere usati nei campi *spatial*, *temporal* e *description* diventeranno esplicite chiavi d'accesso che condurranno il fruitore all'interno di una rete di relazioni tra fonti molto eterogenee tra di loro, ma capaci di raccontare la storia di un luogo urbano attraverso un narrazione costruita dal basso e collaborativamente, con il coinvolgimento diretto degli abitanti.

3. CONCLUSIONE

L'archivio digitale sulla storia e la memoria di Vigne Nuove è online e liberamente accessibile al seguente link: https://oz.onebox.it/s/archivio-digitale-vigne-nuove/page/ricerca. Attualmente ancora in fase di implementazione, l'archivio conserva ad oggi 81 oggetti digitali, configurandosi già come un nucleo archivistico strutturato e operativo, ma al tempo stesso aperto a ulteriori sviluppi.
Tra questi materiali, 15 sono video-interviste non strutturate realizzate con il metodo della storia orale, corredate di trascrizione, che costituiscono una componente particolarmente rilevante del corpus, in quanto restituiscono in forma diretta e situata le memorie, le percezioni e le narrazioni degli abitanti. Significativa è anche la presenza di documenti recuperati attraverso pratiche di crowdsourcing, prevalentemente fotografie. Ai documenti prodotti e raccolti dalla comunità si affiancano, come visto, materiali conservati in archivi istituzionali.
Una delle sfide principali nella realizzazione di questo archivio riguarda la replicabilità del modello. La scelta di Omeka S come piattaforma di implementazione è motivata non solo dalla semplicità d'uso, ma anche dal fatto che si tratti di un software libero e open source. L'adozione dello standard DCMI risponde alla necessità di utilizzare uno schema di metadatazione flessibile, comprensibile e applicabile anche in contesti non altamente specializzati. Il modello proposto è inoltre riusabile poiché i campi di metadati sono resi visibili in front-end per ogni oggetto digitale, consentendo una lettura trasparente delle logiche descrittive adottate. Lo schema di metadati adottato garantisce la salvaguardia del principio di provenienza, del vincolo archivistico e del soggetto produttore, oltre a forme di interoperabilità descrittiva e semantica.
In relazione all'individuazione di metadati concettuali, nel lavoro è stato utilizzato anche un sistema di intelligenza artificiale (ChatGPT). L'interazione con il sistema si basa sulla condivisione delle trascrizioni delle interviste e sull'attivazione di un processo dialogico tra operatore umano e macchina, finalizzato all'emersione e alla comparazione di concetti utilizzabili come metadati.
Questo approccio consente di costruire un sistema informativo nel quale, ad esempio, l'utilizzo della parola chiave "trasformazione" restituisce come risultato 4 video-interviste, 4 fotografie raccolte tramite crowdsourcing, 2 documenti conservati presso l'Archivio Capitolino e 2 fonti secondarie (un volume e un film). Ciascuna di queste risorse è a sua volta descritta mediante ulteriori metadati concettuali, generando una rete di relazioni tra fonti eterogenee. In un caso, una fotografia raccolta tramite crowdsourcing consente di riconnettere l'utente sia all'archivio online di un fotografo locale sia alla

versione su YouTube del film *My Name is Joe* di Ken Loach che lo stesso regista presentò a Vigne Nuove, diventando membro onorario del comitato di quartiere.
L'archivio si mostra come uno strumento narrativo polifonico, dove la comunità ha un ruolo centrale nella condivisione di contenuti e nella co-costruzione di significati. In questa prospettiva, l'archivio digitale non si configura soltanto come uno strumento di conservazione e accesso, ma come un dispositivo di costruzione e mediazione del sapere storico, in cui i metadati svolgono una funzione attiva nell'organizzazione e nell'emersione dei significati.
Da un punto di vista strettamente sociale l'interoperabilità di fonti private e istituzionali rende possibile agli abitanti prendere conoscenza dei dibattiti politici che hanno portato allo sviluppo di un progetto architettonico che ha avuto un impatto determinante sulle loro vite. Comprendere che dietro la realizzazione del progetto ci sono stati anni di studi e dibattiti oltre che a importanti investimenti, può magari aiutare a sentirsi parte di un qualcosa di più grande, che supera i limiti urbanistici dell'area e si connette con tutta la città, con le sue aspettative e i suoi limiti.
Il modello proposto intende mostrare come sia possibile sviluppare archivi digitali di comunità capaci di coniugare accessibilità, replicabilità e profondità interpretativa per promuovere processi di innovazione sociale.

BIBLIOGRAFIA

Berners-Lee, T., & Hendler, J. (2001). The semantic web: A new form of web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American, 284*(5), 34-43

De Felice, R. (1988). *L'archivio contemporaneo. Titolario e classificazione sistematica di competenza nei moderni archivi di enti pubblici e privati*. La Nuova Italia Scientifica.

Floridi, L. (2017). *La quarta rivoluzione. Come l'infosfera sta trasformando il mondo*. Raffaello Cortina.

Hersh, W. (2024). Search still matters: Information retrieval in the era of generative AI. *Journal of the American Medical Informatics Association, 31*(9), 2159–2161. https://doi.org/10.1093/jamia/ocae014

Michetti, G. (2008). Il modello OAIS. *Digitalia, 3*(1), 153–169.

Koselleck, R. (2007). *Futuro passato. Per una semantica dei tempi storici*. Clueb.

Ravveduto, M., & Salvatori, E. (2023). Storia digitale e Digital Public History: Le novità di un antico mestiere. In F. Ciotti (Ed.), *Digital humanities. Metodi, strumenti, saperi* (pp. 229–255). Carocci.

Tomasi, F. (2022). *Organizzare la conoscenza: Digital humanities e web semantico*. Editrice Bibliografica.

Valacchi, F. (2024). *L'archivio aumentato. Tempi e modi di una digitalizzazione critica*. Editrice Bibliografica.

Weston, P. G., & Sardo, L. (2017). *Metadati*. Associazione Italiana Biblioteche.

Yang, W., Li, J., Wang, X., Zhang, Y., & Liu, X. (2025). The impact of modern AI in metadata management. *Human-Centric Intelligent Systems*. https://doi.org/10.1007/s44230-025-00106-5

Participatory Digital Archives as a Method for Co-constructing Cultural Memory: The Bunjevke Women's Platform

SANDRA IRŠEVIĆ PHD
INDIPENDENT ARTIST AND RESEARCHER, SERBIA - SANDRA.IRSEVIC@GMAIL.COM

ABSTRACT

Participatory digital practices serve not simply as dissemination tools but as genuine research methodologies for collaboratively building cultural heritage knowledge alongside marginalized communities. This study analyzes the Bunjevke digital platform—an open multimedia archive featuring community-submitted photographs, audio recordings, and narratives—to explore how such initiatives engage diverse publics, contest established cultural narratives, and cultivate inclusive digital scholarship. Grounded in memory studies, co-creation frameworks, and critical digital humanities, the analysis confronts ethical dilemmas, data stewardship challenges, and the integration of public contributions within Open Science and FAIR principles. Ultimately, the paper advances a practical model for conceptualizing participatory digital archives as dynamic research instruments rather than static repositories.
Keywords: participatory digital archives; digital humanities; co-construction of knowledge; cultural memory; Open Science

ABSTRACT (ITALIANO)

Archivi Digitali Partecipativi per la Co-costruzione della Memoria Culturale: Il Caso Bunjevke
Le pratiche digitali partecipative trascendono il ruolo di semplici strumenti di divulgazione per configurarsi come autentici metodi di ricerca nella co-costruzione della conoscenza del patrimonio culturale con comunità marginalizzate. Questo studio esamina la piattaforma digitale Bunjevke—archivio multimediale aperto con contributi comunitari di fotografie, registrazioni audio e narrazioni—per comprendere come tali iniziative coinvolgano il pubblico, sfidino narrazioni culturali dominanti e promuovano una digital scholarship inclusiva. Radicato negli studi della memoria, nelle metodologie di co-creazione e nelle digital humanities critiche, l'analisi affronta dilemmi etici, sfide nella gestione dei dati e l'integrazione dei contributi pubblici nell'Open Science e nei principi FAIR, proponendo un modello pratico per concepire gli archivi digitali partecipativi come strumenti di ricerca dinamici anziché semplici depositi.
Parole chiave: participatory digital archives; digital humanities; co-construction of knowledge; cultural memory; Open Science

1. INTRODUCTION

Digital curation has been radically transformed since 2001. The erstwhile static repository of institutional knowledge now is rendered as a participatory platform in the production, negotiation, and (re)distribution of cultural memory. It unsettles and challenges archival tradition that has prioritized authoritative curation, stable metadata and one-way preservation models. Rather, these are participatory digital archives where communities are actively co-constituting historical narratives rather than implicitly being made objects of documentation. In this context, archives stop being what they were and becoming epistemological tools that determine how the past can be thought, who can speak it and what stories find visibility. This article explores the Bunjevke digital platform through the lens of participatory archiving. The platform currently features photographic, oral and written submissions uploaded by Bunjevac community members, with a particular focus on women whose daily cultural existence has traditionally been invisible in institutional archives. Rather than a closed archive system, Bunjevke is an open and iterative cultural memory where outdoor and indoor collective memories create culturally remembered practices. The design of the platform invites users to add their personal materials, annotate existing ones, and participate in cross-generational conversation turning the archive into a space of co-creation.

2. THEORETICAL FRAMEWORK

The starting point of this study is that participatory archives need to be understood not only as technological platforms but as methodological and ethical frameworks for the production of cultural memory. They do more than simply preserve heritage; they live it. Through the lens of the Bunjevke platform, the article explores how digital participation reshapes the notion of authorship, disperses curatorial power and engenders novel methodological and ethical concerns. The sections are five. In the wake of this introduction, a theoretical framework places participatory archiving within the context of contemporary memory theory and introduces the CARE Principles as the overarching ethical framework for community data governance. Section 3 describes the methodological approach and delineates the dataset. 4 reports the analytical results in terms of interface mediation, curatorial tensions and ethical negotiation. The conclusion articulates the implications for digital heritage practice and indicates potential areas for further research. The development of cultural memory has long been shaped in part by the dualism of preservation and hermeneutics. However, records were previously regarded in early archival theory as objective traces of the past, and the historical reliability of records was secured by archival principles such as provenance and original order. The answer in contemporary memory and digital humanities scholarship has been a resounding no to this positivist position. Remembering theorists such as Jan Assmann and Astrid Erll have highlighted that memory is fundamentally selective, socially constructed and subject to renegotiation within and across generations. In digital media, this process is more explicit: archives have ceased to be neutral containers, and are now technologically mediated mechanisms that shape what is captured, how it is organized and which stories are foregrounded or backgrounded.

Participatory archives further this critique by decentering the curator. Instead of gatekeepers at institutions, these platforms allow the public to add to, annotate, and contextualize the collections. This democratization of archival work dovetails with larger trends in digital humanities methodologies toward co- production, open scholarship, and reflexive methodology. But participation does not equal equity, or accuracy. The democratization of contribution brings variation in the depth of documentation, context, and interpretative coherence. Archival integrity, then, must be preserved through the process of mediation, and not at the cost of muting community voices. This tension is not a weakness of participatory systems but rather a symmetry of their epistemological form. As a way to manage these tensions, this research is informed by the CARE Principles for Indigenous Data Governance (Carroll et al., 2021). Developed initially for Indigenous peoples, the CARE framework on the basis of Collective Benefit, Authority to Control, Responsibility, and Ethics can be considered a strong normative framework that other marginalized groups involved in self-documentation may be able to draw upon. They challenge extractive data practices by affirming that communities should maintain sovereignty over their cultural materials in how they are collected, through interpreted, and in how they are shared. Within the framework of the Bunjevke platform, CARE serves as a lens through which to scrutinize consent protocols, narrative ownership and ethical obligations in the very design of the platform. But it turns participation into more than just submitting data—it becomes an expression of cultural self-determination.

3. RESULT

In such a theoretical field cultural memory is alive process rather than static documentation. It is formed through social interactions, emotional experiences, and interactions with technology. By making memory work into a collaborative practice, participatory archives concretize this dynamic. They illustrate how digital infrastructures may reclaim disperses histories, validate vernacular knowledge, and subvert institutional narratives that marginalized peripheral communities. The Bunjevke platform illustrates this dynamic by allowing women to capture and preserve domestic traditions, language use, and intergenerational narratives that are otherwise excluded from formal heritage archives. In this way, it demonstrates how digital participation can be understood as a mode of co-construction of cultural memory and not solely as a site of its preservation. 3. Methodology and Dataset The analytic approach in this study brings together the following three interrelated methods walkthrough method, visual and discursive analysis, and computational assistance. The walkthrough method (see Light, Burgess & Duguay, 2017) offers a structured approach to analyzing platform design, users flows, and interface affordances. Its analysis tracks how design decisions affect patterns of participation, visibility of

contributions, and dynamics between communities of users as it follows journeys through the Bunjevke platform at different moments in a series of sessions from the perspective of participation. This strategy views the interface function not as a neutral container of archival practice or as a simple conduit but as an active mediator of archival practice. Based on Gillian Rose's (2016) methodological approach, the visual and discursive analysis is conducted on the user-generated content. Photographs, scanned documents, and audio transcripts are analyzed with respect to compositional decisions, narrative framing, and contextual notes.

Discursive analysis is applied to user comments, metadata submissions, and community corrections to explore the processes by which meaning is negotiated, challenged, and reified on this platform. Together, these approaches facilitate a sophisticated engagement with the material and interpretive aspects of participatory archiving. In order to facilitate analysis of vast quantities of visual submissions, computational image recognition tools were incorporated in the initial data processing stage. OpenAI's CLIP (Contrastive Language-Image Pre-training) model and Google Cloud Vision API are the two we use for generating base metadata tags that describe images, recognize repeating visual themes, and reveal possible biases within automated tagging. These instruments were not utilized as definitive categorization tools, but rather as analytical tools to detect patterns for further scrutiny using human-centered analytical Means. Community annotations were used to compare with the outputs of the AI to examine areas of agreement and disagreement as well as the limits of algorithmic interpretation in culturally bound situations.

For this study we analyzed contributions made from June 2024 until December 2025. For the input, we used the registered contributions of community members: 1,842 photos, 312 audio files, and 567 text stories. The textual corpus is written mostly in the Serbian and Bunjevian language. All the content was uploaded via the platform's moderated submission system, which involves a simplistic consent form and provides contributors with options to indicate the level of access to the material (public, community-only, or restricted). The dataset is deliberately self-selecting, and reflects the demographics of active users of the platform rather than those of the wider community. This introduces a known limitation for representational balance with respect to age group and geographical distribution. However, it offers a solid basis for doing much of what explore – how participatory mechanisms influence cultural memory production, specifically in terms of women's everyday heritage practices visibility.

4. ANALYSIS AND FINDINGS

The Bunjevke platform is presented as a dialogical archive where individual entries crisscross, accumulate and are transformed through the collective engagement of the community. Unlike traditional repositories that lock objects behind rigid catalog records, Bunjevke is all about layered access. Users often identify individuals, add clarifications on historical moments, or relate family histories in comments on photos. Audio recordings frequently include transcriptions, which are edited by multiple users. This recursive process of annotation converts the archive from a monolithic accumulation into a narratological place where memory is perpetually evolving and deepening.

The major result from the [analysis] is the competing claims for open participation and curatorial responsibility. The platform design intentionally makes it easy to contribute, but this open environment leads to variation in quality of documentation and interpretive framing. Some submissions do not have exact dates, geographical locations, or application descriptions, others including in-depth information about the origin and dossier notes or links to external sources. Invidious space is preserved without requiring town hall-style full consensus by employing a mild-moderation model. Community moderators also confirm that submissions superficially follow the content rules, they mark sensitive content, and recommend improvements on the metadata, but they definitely do not change the original stories once they have been posted. Such a methodological approach retains authenticity while pragmatically utilising aspects of archival usability. The emergent organization emphasizes that participatory archives do not need centralization to exist; rather, they are constituted through distributed responsibility and shared archival literacy.

Ethical questions surface as a structural concern of the platform, as opposed to an ethics of consequence. Users become actively engaged in acts of culture risk when they deposit family photos, oral histories, or personal papers. On issues of consent, ownership of narrative, and representation of descendants arises very quickly. The platform recognizes these issues with clear permission settings, options for contributor accountability, and a mechanism for withholding sensitive documents. More importantly, the

community's internal code of respect and contextual care acts as a de facto ethical safeguard. This is consistent with the emphasis on collective responsibility and community authority in the CARE Principles. Instead of approaching ethics as a box to tick for compliance, the platform instills ethical consciousness in its culture of participation, providing a model for how marginalized groups can manage their own digital heritage without institutional mediation.

Interface design is pivotal in determining the dynamics of these relations. The organization of the platform, its tagging system and its search features channel user conduct. For example, the prioritization of chronological and thematical browsing over keyword-based search promotes exploratory activities instead of one for result-oriented searching. Recently added or most commented on are visually privileged, which determines which stories becomes visible. These are not neutral design decisions; they are decisions on the level of what knowledge is made relevant, how memory is to be accessed, and who gets to speak. The walkthrough demonstrates that small changes to the interface—moving the annotation tool, for instance, or changing which comments are visible by default—can have a significant impact on how people participate. This underscores the fact that technological infrastructure and archival practice are tightly intertwined. The platform, in the end, illustrates the potential participatory digital archives have to reshape cultural visibility. Focusing on women's domestic work, linguistic custom, and oral/trans-generational Tradition Keeping, Bunjevke confronts historiographical paradigms that have Othered vernacular heritage. The archive does not simply recuperate missing narratives but reimagines them as plural, non-hierarchical forms of representation. Institutional archives are not obliterated in the process but rather complemented, as the possible epistemic horizons of the culturally meaningful are widened. What this research suggests is that participatory archiving—when it is thoughtfully designed and ethically informed—can also be a way of co-producing cultural memory that does not have to be an afterthought to institutional heritage preservation.

5. CONCLUSION

Participatory digital archives constitute a methodological and epistemological shift in a move from top-down preservation to bottom-up knowledge making. Our analysis of Bunjevke platform indicates that the CARE Principles and design for ethics negotiation as embedded in platform design have the potential to empower nonhegemony communities to co-author their histories and exercise data sovereignty not only within Indigenous contexts but beyond. Yet, participatory archiving is fraught with challenges on its own terms, such as varying quality of contributions, self-selection biases, and the ethics of documenting individual lives. Instead of thinking of them as limitations to be worked around, we suggest they be engaged with transparently and through the development of community literacies and distributed responsibilities. With the proper critical approach, computational image analysis can in a sense supplement (though never substitute for) Interdisciplinary analyses, While at the same time strengthening and exercising methodological reflexivity. We suggest that future research pursue (a) investigations of participatory archives over time, (b) analyses of how community norms evolve, and (c) studies of the effects of updates on platforms on narrative visibility. In order to parse participatory organization into what is transferable and what is contextually native, cross-cultural and cross-linguistic analyses will be required. Critical interrogation of the use of AI in archival processing, in particular for bias detection and cultural particularity, is sorely needed. A participatory digital archive is never a still repository but a constantly evolving mode of practice – attuned to negotiation through engagement with, care for, and technological mediation of relationship. The Bunjevke platform presents a generalizable approach for community-based initiatives and partnering institutional efforts committed to fostering wider, ethically-situated, and methodologically sustainable strategies for engaging with (digital) cultural memory.

REFERENCES

Assmann, J. (2011). Cultural Memory and Early Civilization: Writing, Remembrance, and Political Imagination. Cambridge University Press.

Bishop, J. (2018). Auditing participatory platforms: A framework for evaluating community-driven digital projects. International Journal of Heritage Studies, 24(8), 845–860.

Cameron, F., & Kenderdine, S. (Eds.). (2007). Theorizing Digital Cultural Heritage: A Critical Discourse. MIT Press.

Caswell, M. (2014). Seeing yourself in history: Community archives and the fight against symbolic annihilation. Public Historian, 36(4), 26–37.

Cook, T. (2013). Evidence, memory, identity, and community: Four shifting archival paradigms. Archival Science, 13(2–3), 95–120.

Denton, W., et al. (2020). Ethics of FAIR in practice: Critical perspectives on data stewardship. International Journal of Digital Humanities, 1(4), 387–405.

D'Ignazio, C., & Klein, L. F. (2020). Data Feminism. MIT Press.

Drucker, J. (2021). The Digital Humanities Coursebook: An Introduction to Digital Methods for Research and Design. Routledge.

Frisch, M. (1990). A Shared Authority: Essays on the Craft and Meaning of Oral and Public History. State University of New York Press.

Kilibarda, K. (2018). Minorities and citizenship in post-Yugoslav states: The case of Bunjevci. Nationalities Papers, 46(5), 841–857.

Light, B., Burgess, J., & Duguay, S. (2017). The walkthrough method: An approach to the study of archived, platform-specific social media data. New Media & Society, 19(6), 881–900.

Ridge, M. (Ed.). (2014). Crowdsourcing our Cultural Heritage. Ashgate.

Rose, G. (2016). Visual Methodologies: An Introduction to Researching with Visual Materials (4th ed.). Sage.

Schwartz, J. M., & Cook, T. (2002). Archives, records, and power: The making of modern memory. Archival Science, 2(1–2), 1–19.

Simon, N. (2010). The Participatory Museum. Museum 2.0.

Terras, M. (2016). Digital humanities and the changing ecology of scholarly communication. In A New Companion to Digital Humanities (pp. 557–572). Wiley Blackwell.

Wilkinson, M. D., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3, 160018.

Il ritorno dei partigiani sardi. Una collezione digitale co-prodotta

WALTER FALGIO[1], LAURA STOCHINO[2], MATTEO QUARANTIELLO[3]

[1]ISSASCO, WALTERFALGIO@GMAIL.COM
[2]ISSASCO, LAURA.STOCHINO@GMAIL.COM
[3]MATTEOQUARANTIELLO@GMAIL.COM

ABSTRACT (ITALIANO)

La proposta intende illustrare gli esiti attesi di un progetto di ricerca in corso inerente all'analisi e alla comparazione delle biografie di alcuni antifascisti sardi, partigiani, oppositori politici, militari, che presentano caratteri di interesse sotto il profilo della vicenda personale e della relazione con i contesti geografici, sociali e politici nei quali si sono trovati a interagire soprattutto nel post-Resistenza. L'obiettivo finale prevede la valorizzazione e la pubblicazione in formati *Open Access* delle letture e delle fonti, prevalentemente audiovisive e documentarie, attinenti all'indagine storica. In coerenza con l'idea di "digitale" come spazio di restituzione pubblica e di partecipazione, gli esiti della ricerca confluiranno in una collezione costruita in collaborazione con un ente conservatore (Cineteca Sarda, Archivio di Stato di Cagliari, Issasco, Anppia) e sviluppata con strumenti consolidati e sostenibili anche con risorse economiche contenute. Il *public engagement* si articolerà in pratiche formative e partecipative con classi della scuola secondaria di secondo grado, concepite come dispositivo di co-produzione controllata.
Parole chiave: guerra partigiana; post-Resistenza; ritorno dei partigiani; collezione digitale; co-produzione.

ABSTRACT (ENGLISH)

THE RETURN OF THE SARDINIAN PARTISANS. A CO-PRODUCED DIGITAL COLLECTION
The proposal aims to illustrate the expected outcomes of an ongoing research project concerning the analysis and comparison of the biographies of several Sardinian anti-fascists—partisans, political opponents, and military figures—whose lives are of particular interest both in terms of their personal trajectories and their relationships with the geographical, social, and political contexts in which they interacted, especially in the post-Resistance period.
The final objective is to enhance and publish, in Open Access formats, the readings and sources—mainly audiovisual and documentary—relevant to the historical investigation. In line with the idea of the "digital" as a space for public dissemination and participation, the research outcomes will be brought together in an collection developed in collaboration with a heritage institution (Cineteca Sarda, State Archives of Cagliari, Issasco, Anppia) and built using well-established and sustainable tools, also suitable for contexts with limited financial resources.
Public engagement activities will take the form of educational and participatory practices involving upper secondary school classes, conceived as a framework for supervised co-production.
Keywords: partisan war; post-Resistance; return of the partisans; digital collection; co-production.

1. INTRODUZIONE

Il post-Resistenza, le dinamiche di integrazione sovente complesse dei combattenti antifascisti e degli oppositori al regime nei contesti sociali di provenienza o in quelli di adozione, le sfaccettate biografie inscritte in percorsi molto disparati (l'impegno civile, la cospirazione sotto il fascismo e la lotta armata dopo l'8 settembre '43 assumono caratteri specifici in base alle aree geografiche nelle quali si dispiegano), hanno spinto a interrogarci sui destini di alcuni partigiani e militanti sardi scelti prevalentemente tra coloro che hanno lasciato tracce scritte delle proprie esperienze. Preso atto del significato di profonda discontinuità rappresentato dalla guerra partigiana in Italia, si assumono come tuttora efficaci le considerazioni del leader comunista Pietro Secchia in base alle quali la Resistenza «non fu un miracolo».

2. UN QUADRO FRASTAGLIATO

«La Resistenza non fu un miracolo, né un fenomeno spontaneo: dovette essere organizzata. Dura, difficile, piena di difficoltà fu sempre, fino alla fine, ma soprattutto agli inizi». Queste, dunque, le parole di Secchia pubblicate nel 1973 (110-1) che, come altre, rendono conto della complessità, dei limiti e della straordinarietà della guerra partigiana (Peli, 2025). Come peraltro *la scelta*, l'apertura al nuovo, il frangente in cui, nelle parole di Vittorio Foa (1980: 50), «tutto ci era parso possibile», non fu «un'istantanea illuminazione» ma, piuttosto, «un processo che si apre la strada a fatica, perché affaticati sono gli uomini che lo vivono» (Pavone, 1991: 39). Santo Peli, uno degli storici che ha contribuito con maggiore lucidità alla ricostruzione e alla lettura della guerra partigiana, osserva «il costante intrecciarsi di motivazioni "alte" e di scelte della più varia natura, individuali e localistiche, che portano a entrare in banda, ma a volte anche a uscirne» (2025: 146). Questa esperienza intensa, tormentata e «passionale», per richiamare un recente studio di Chiara Colombini, è stata per di più soggetta alternativamente a «polemiche roventi, denigrazioni, celebrazioni, frettolose archiviazioni e altrettanto frettolosi recuperi strumentali», tali da determinare «rappresentazioni a volte distorte e smaccatamente liquidatorie, a volte di segno opposto», ma sovente semplificatorie (Colombini, 2023: 5).
In tale quadro frastagliato e certamente non omogeneo si inscrivono testimonianze di sarde e di sardi che intendiamo selezionare e raccogliere in uno spazio digitale *Open Access* anche attraverso pratiche formative e partecipative nelle quali sono coinvolti e saranno coinvolti studentesse e studenti della scuola secondaria di secondo grado.
Sinteticamente, il progetto prevede la ricostruzione, l'analisi e la comparazione delle biografie di alcune antifasciste e di alcuni antifascisti sardi, partigiani, oppositori politici, militari, che presentano caratteri di interesse e forse specificità sotto il profilo della vicenda personale e della relazione con i contesti geografici, sociali e politici nei quali si sono trovati a interagire. Le fonti così raccolte, alcune delle quali audiovisive, saranno pertanto rese disponibili in una collezione digitale e integrate da schede esplicative e riassuntive elaborate dagli autori e dalle classi coinvolte attraverso i laboratori didattici.

2.1 L'esperienza di Nino Garau

I temi legati all'opportunità del "ritorno" e alle modalità di inserimento in vecchi o nuovi contesti sociali da parte dei reduci, oltre, naturalmente agli aspetti rilevanti delle storie individuali, sono al centro della riflessione. Sulla partecipazione dei sardi alla guerra di Liberazione negli ultimi anni si è registrata la riapertura di un dibattito, accompagnato anche dalla pubblicazione di studi scientifici (Sanna, 2018) promossi dall'Istituto sardo per la storia dell'antifascismo e della società contemporanea (Issasco). In particolare, grazie sempre all'iniziativa dell'Istituto storico, si è indagato sulla figura del comandante partigiano Antonio "Nino" Garau (Garau & Falgio, 2021). Nato a Cagliari nel 1923 e morto nella stessa città nel 2020, operativo durante la Resistenza nella pianura della provincia di Modena, aveva costituito e guidato la brigata intitolata ad "Aldo Casalgrandi". Nel 2012 è stato reso pubblico un mediometraggio (Geppe e gli altri. Storia di vita di un comandante partigiano sardo, Issasco, Lev, 2014, regia di Francesco Bachis)[1] contenente l'intervista al combattente, frutto di un lavoro di ricerca e di montaggio avviato nel 2006. Il cagliaritano, ex alto funzionario del Consiglio regionale della Sardegna, dopo un lungo oblio durato oltre sessant'anni, decise di trasferire la lucida memoria degli eventi di guerra e delle vicende immediatamente successive al 1945 che lo videro, tra l'altro, incarcerato ingiustamente a seguito di una accusa infondata e anonima relativa al coinvolgimento nell'omicidio di un fascista nel Modenese. Con queste parole Antonio Garau raccontò il suo arresto a Cagliari nel 1949:

> Una mattina, esattamente alla fine di Luglio del 1949 (io mi laureai in giugno), sentimmo alle 5,30/6,00 del mattino bussare alla porta di casa. Mio padre aprì. Era una squadra di questurini tutti in borghese che fece vedere a mio padre il mandato di perquisizione. Dissero: «È per suo figlio, indiziato per un omicidio». […] Mi portarono in Questura che al tempo era in piazza San Cosimo, mio padre mi seguì. Lì non subii nessun interrogatorio pertanto non seppi nulla del motivo del mio arresto. Ad un certo punto, dopo mezz'ora, credo che avessero parlato con mio padre, mi misero le manette (o meglio una catena con lucchetto).

[1] https://www.youtube.com/watch?v=MLMhYzwDE88 (cons. 31/01/2026)

Su Garau è stato anche messo in campo un intervento di valorizzazione della sua esperienza attraverso risorse digitali e a seguito di un progetto didattico finalizzato alla elaborazione della pagina Wikipedia dedicata[2]. L'iniziativa, coordinata dal Convitto Nazionale "Vittorio Emanuele II" di Cagliari, trae origine da un percorso formativo promosso dalla fine del 2020 dall'Istituto storico della Resistenza nel Novarese e indirizzato ai docenti e agli studenti. Il laboratorio ha consentito di approfondire la conoscenza delle enciclopedie digitali e la realizzazione di un nuovo contenuto on-line, riferito a episodi e protagonisti della storia italiana contemporanea connessi anche alla propria realtà territoriale.

2.2 I partigiani sardi e il ritorno

Alla biografia di Garau si associano altre storie di vita rilevanti, come accennato, nell'ottica della complessa o mancata integrazione nelle comunità di provenienza. Riguardo ad alcune vicende personali si possono richiamare fonti audiovisive disponibili e di indiscusso valore storico. Questo è senz'altro il caso di Fausto Cossu (Tempio Pausania, 1914 – Piacenza, 2005): allievo ufficiale carabiniere, nel '41 è inviato nei Balcani per combattere la guerra di occupazione fascista affiancata ai tedeschi. È qui, in un contesto di scontro sanguinoso, crudele e spietato nei confronti dei civili, che il tempiese, come tanti altri sardi, matura la sua profonda avversione verso il regime. Dopo l'internamento in Austria da dove riuscirà a fuggire, Fausto Cossu rientrerà in Italia e raccoglierà l'adesione alla lotta antifascista di altri carabinieri non disponibili all'arruolamento nella Repubblica di Salò sino a formare una compagnia. Arriverà a guidare la divisione "Piacenza" e ad essere ispettore generale di tutte le forze gielliste dell'Oltrepò Pavese. Dopo la guerra Cossu deciderà di non fare rientro in Sardegna: aspetto che deve essere indagato al di là delle informazioni rese finora disponibili dalle ricostruzioni storiche. «Non vi fece mai ritorno, seppure rammaricandosi, come mi scrisse in una lettera testimonianza nei primi anni Ottanta, che la Sardegna e la sua città non si fossero ricordate di lui» (Sechi, 2002: 288).
Ora, a sostegno della ricostruzione della figura del capo partigiano di Tempio, si aggiunge anche il lavoro della Società Umanitaria Cineteca Sarda di Cagliari condotto dall'operatrice culturale Giulia Mazzarelli. Si tratta di una ricerca su un documento piuttosto raro, una pellicola ritrovata e custodita nell'Archivio Superottimisti di Torino in cui compare il comandante Cossu durante la guerra di Liberazione[3].
Dai Balcani provenivano anche il samassese Giovanni Cuccu (Samassi 1914-2005) e il finanziere di Cabras Dario Porcheddu (Cagliari 1922-2009). Il primo, grande esperto nell'uso delle armi pesanti, combattente in Slovenia, si congederà con il grado di maggiore dell'esercito jugoslavo. Il suo ritorno a casa non sarà facile:

> Io, invece, che avevo fatto tanto sacrificio ero senza lavoro, senza studio, solo con quel poco che avevo imparato nei miei brevi anni di scuola elementare e che la vita e la natura mi avevano insegnato. Così per circa venticinque anni tirai avanti senza avere pace, perché controllato in continuazione in quanto sospettato di essere spia dell'Armata Rossa (Cuccu, 2000: 193).

Il secondo, operativo nel Montenegro, scampato rocambolescamente a una fucilazione, per anni presidente dell'Unione Autonoma Partigiani Sardi, al contrario vivrà un rientro nel suo paese di origine non traumatico:

> Per giorni fu festa grande, per giorni fu un andirivieni di gente, di amici, di parenti, di conoscenti, di persone che non riconoscevo più, tutti a congratularsi, tutti a voler sapere, tutti ad abbracciarmi. Quanti di loro avevano perso un padre, un amico, in quella maledetta guerra, ed io ero lì, per tutti loro, per tutti quelli che non potevano più tornare. Per ricominciare. (Porcheddu, 1994: 152).

Entrambi sono tra i pochissimi ad aver lasciato una traccia scritta della loro esperienza di guerra. A questi nomi si possono aggiungere quelli di Antonio Tedde (1905-1987)[4], Modesto Melis (1920-2017)[5], Vittore Bocchetta (1918-2021)[6], Tommaso Serra (1900-1985), Luigi Polano (1897-1984)[7], Giampiero Besson (1922-1952). Isolani che in forme diverse e in periodi non sempre coincidenti si sono opposti alle

[2] https://it.wikipedia.org/wiki/Nino_Garau (cons. 31/01/2026)
[3] https://www.facebook.com/cineteca.umanitaria/videos/997432052584972/ (cons. 31/01/2026)
[4] https://www.francoangeli.it/Libro/Un-ufficiale-scomodo.-Dall%27armistizio-alla-guerra-di-liberazione-(1943-1945)?Id=10037 (cons. 31/01/2026)
[5] https://www.sardegnadigitallibrary.it/detail/6499b9eae487374c8f804925 (cons. 31/01/2026)
[6] https://www.youtube.com/watch?v=nkr921BD8nA (cons. 31/01/2026)

dittature del Novecento e su cui esistono fonti accessibili in grado di restituire elementi utili alla definizione della

> cruciale dimensione del dopo conflitto […] con tutto il bagaglio di riconoscimenti oppure disconoscimenti dell'esperienza partigiana, nonché delle successive e spesso dolorose scelte individuali (impegno civile e politico-sindacale, ripiegamento nella sfera privata, emigrazione) (Fimiani, 2016: 12).

2.3 Il contributo del Mezzogiorno alla Resistenza

Riguardo al contributo del Mezzogiorno d'Italia alla Resistenza gli studi, in particolare di Isabella Insolvibile e di Toni Rovatti, hanno consentito di individuare alcuni aspetti centrali estendibili all'esperienza dei sardi. Per esempio, si osserva che il ritorno nella sfera privata dei combattenti nel sud dopo la guerra fu la scelta maggiormente praticata «in un contesto non disposto a un'apertura politico culturale verso i temi della Resistenza» perché assorbito dall'urgenza della ricostruzione e scarsamente propenso a rivedere criticamente il passato se non in vesti «artificialmente democratiche» (Insolvibile, 2025: 93). Di una «precoce e vivida delusione delle aspirazioni etiche-ideali maturate durante la Resistenza» parla Rovatti (2016: 179) rimarcando che gli ex combattenti del sud estromessi dalla dimensione pubblica «si rivelano promotori di uno specifico radicalismo ideale» e di un precoce e profondo senso di disillusione verso le istituzioni repubblicane (Rovatti, 2016: 179). Si ritiene pertanto ancora stimolante la domanda posta da Rovatti se esistano o meno caratteri che rendono peculiare l'adesione e la partecipazione degli uomini e delle donne del sud Italia alla lotta di Liberazione nel nord (2016: 159) e si intende estendere l'eventuale pertinenza dell'interrogativo anche alle modalità del ritorno e dell'integrazione nelle comunità originarie dei sardi, con l'introduzione della variabile necessaria legata ai caratteri della condizione insulare.

3. IL PROGETTO DIGITALE

In coerenza con l'idea di "digitale" come spazio di restituzione pubblica e di partecipazione, gli esiti della ricerca confluiranno in una collezione digitale costruita in collaborazione con un ente conservatore (Cineteca Sarda, Archivio di Stato, Issasco o altro soggetto analogo) e sviluppato in ambiente WordPress tramite il plugin Tainacan, soluzione consolidata e sostenibile anche con risorse economiche contenute. La collezione raccoglierà fonti audiovisive e documentarie e le metterà in relazione con schede biografiche e contestuali: non soltanto un insieme di materiali, dunque, ma un sistema interrogabile che consenta di ricostruire nessi tra persone, luoghi ed eventi, mantenendo tracciabili le scelte di selezione e descrizione e rendendo esplicito il modello con cui i contenuti vengono organizzati.
Dal punto di vista strutturale, la collezione sarà articolata in un numero limitato di entità ricorrenti e tra loro connesse (persona, fonte, luogo, evento/fase), così da evitare una mera giustapposizione di oggetti e rendere invece possibile una lettura comparativa. Per ciascuna figura biografica verrà predisposta una scheda "persona" che fungerà da nodo principale, associando un nucleo di informazioni essenziali (nome, ruolo, cronologia di riferimento, profilo biografico) a un carnet di materiali e rimandi: almeno una fonte audiovisiva o documentaria quando disponibile, un apparato di sintesi e contestualizzazione, e la geografia dei luoghi rilevanti (nascita, aree operative, rientro o non rientro, eventuali luoghi di detenzione o emigrazione). La piattaforma consentirà di mantenere distinti i contenuti (video, documenti, immagini, estratti) dalle informazioni descrittive, così da rendere chiaro "che cosa" viene mostrato e "come" viene descritto e classificato.
Sul piano dei metadati, la descrizione delle risorse adotterà campi standard di uso comune (Dublin Core per titolo, datazione, descrizione, soggetti, provenienza, diritti), integrati da un set di campi specifici del progetto, necessari alla comparazione storica: tipologia di fonte (intervista, pellicola, documento, memorialistica), rapporto della fonte con la biografia (testimonianza diretta, ricostruzione, citazione), luoghi normalizzati e periodizzazione, nonché alcune categorie interpretative rese operative tramite vocabolari controllati (ad esempio ruolo, contesto operativo, modalità del rientro o della mancata integrazione). Questa scelta consente di evitare descrizioni puramente narrative e di offrire invece filtri e percorsi tematici replicabili, utili sia alla consultazione pubblica sia a successive elaborazioni scientifiche.

[7] https://www.raiplay.it/video/2025/02/Sa-die-prima-de-cras-p3-f9299107-9e8e-49e1-972b-0fd511b77a9c.html (cons. 31/01/2026)

Per Open Access intendiamo, prima ancora che una scelta tecnica, un impegno di responsabilità scientifica e civile: garantire la consultazione libera e stabile dei contenuti che possono essere condivisi, evitando barriere di accesso e favorendo una fruizione informata e citabile. Ciò comporta un lavoro esplicito sulle provenienze e sulle condizioni d'uso, distinguendo con chiarezza tra materiali integralmente riproducibili, risorse consultabili tramite collegamenti o incorporamenti da piattaforme autorizzate, ed eventuali estratti pubblicabili previa autorizzazione. Quando possibile, le schede e i metadati saranno rilasciati con licenze aperte, così da rendere trasparente l'intento di riuso e di circolazione del sapere; al contempo, la gestione dei diritti sarà trattata come parte integrante della restituzione, esplicitando limiti, responsabilità e condizioni di accesso.
Il public engagement si articolerà in pratiche formative e partecipative con classi della scuola secondaria di secondo grado, concepite come dispositivo di co-produzione controllata. A partire da dossier predisposti dai ricercatori (selezione di fonti, bibliografia essenziale, nodi interpretativi), le studentesse e gli studenti contribuiranno alla redazione di sintesi, glossari e indicizzazioni guidate, secondo criteri condivisi e ripetibili; i contenuti prodotti saranno revisionati da docenti e ricercatori prima della messa online, così da coniugare partecipazione e affidabilità e rendere osservabile il processo con cui la conoscenza viene costruita e validata. A supporto della fruizione, la collezione includerà una mappa dei luoghi e una timeline (uMap e TimelineJS) alimentate dai metadati e collegate alle schede, per rendere visibile l'intreccio tra biografie, geografie e temporalità del "ritorno" e permettere percorsi di lettura tematici e comparativi. L'impianto è pensato per crescere per lotti, mantenendo la struttura coerente e replicabile, in modo da garantire continuità nel tempo senza appesantire i partner coinvolti.

RINGRAZIAMENTI

Il progetto si avvale di un finanziamento della Regione Autonoma della Sardegna.

BIBLIOGRAFIA

Colombini, C. (2023). Storia passionale della guerra partigiana. Bari-Roma, Laterza.

Cuccu, G. (2000). Le stelle ci guidano. Cagliari, Cuec.

Foa, V. (1980). Carlo Levi uomo politico (1967). In Id. Per una storia del movimento operaio. Torino, Einaudi.

Fimiani, E. (Cur). (2016). La partecipazione del Mezzogiorno alla Liberazione d'Italia (1943-1945). Firenze, Le Monnier.

Garau, A., & Falgio, W. (2021). La resistenza di Geppe. Sassari, Soter.

Insolvibile, I. (2025). Non solo vento del nord: il Meridione e i meridionali nella Resistenza. In F. Focardi & S. Peli (Cur.), Resistenza. Roma, Carocci, 81-97.

Pavone, C. (1991). Una guerra civile. Torino, Bollati Boringhieri.

Peli, S. (2025). Guerra partigiana e rifiuto della guerra. In F. Focardi & S. Peli (Cur.), Resistenza. Roma, Carocci, 139-161.

Porcheddu, D. (1994). Ho baciato la morte. Cagliari, Taim.

Rovatti, T. (2016). Combattere lontano da casa. L'esperienza dei partigiani meridionali nelle regioni del Nord. In Fimiani, E. (Cur). La partecipazione del Mezzogiorno alla Liberazione d'Italia (1943-1945). Firenze, Le Monnier, 159-199.

Sanna, D. (Cur) (2017). La Sardegna e la guerra di Liberazione. Milano, FrancoAngeli.

Secchia, P. (1973). Il partito comunista italiano e la guerra di Liberazione 1943-1945. Ricordi, documenti inediti e testimonianze. Milano, Istituto Giangiacomo Feltrinelli.

Sechi, S. (2002). Il partigiano Cossu. Almanacco Gallurese, 10, 286-8.

Dalla materialità delle fonti al dato condiviso: metodi ibridi per un archivio digitale partecipativo

TIZIANA PASCIUTO
DIPARTIMENTO DI STUDI STORICI, UNIVERSITÀ DI TORINO, ITALIA – TIZIANA.PASCIUTO@UNITO.IT

ABSTRACT (ITALIANO)

Il contributo propone una riflessione metodologica sulle pratiche delle Digital Humanities applicate a contesti storici e culturali caratterizzati da forte eterogeneità delle fonti e da una crescente esigenza di impatto pubblico. Partendo dal progetto REDMIX, volto alla creazione di un archivio digitale su persone e gruppi con ascendenza mista del Mar Rosso dall'Ottocento ad oggi, il contributo analizza come la scelta consapevole degli strumenti digitali debba essere guidata dalla materialità e dalla natura delle fonti, anziché dalla mera disponibilità tecnologica. Attraverso due casi studio - tratti dall'Archivio privato di Daniela Aleggiani, costituito da diari ibridi dattiloscritti, fotografie e disegni relativi ai campi di prigionia in Africa orientale appartenuti al Maggiore Umberto Aleggiani, e dall'Archivio privato della Famiglia Vento Chiaffredo, comprendenti poesie e canzoni del giovane soldato Chiaffredo Vento in Libia - il contributo mostra come fonti diverse richiedano workflow differenziati. L'analisi illustra l'uso combinato di tecnologie quali OCR, HTR, analisi computazionale del testo, codifica semantica in XML-TEI e visualizzazioni spazio-temporali, evidenziandone limiti, potenzialità e implicazioni metodologiche. I casi studio fungono da laboratorio per la progettazione dell'infrastruttura del progetto REDMIX, concepita come archivio digitale partecipativo e piattaforma di Digital Public History. Il contributo sostiene che l'adozione di workflow ibridi, aperti e trasparenti, consenta non solo una valorizzazione scientificamente rigorosa delle fonti, ma anche la loro trasformazione in strumenti di public engagement e co-costruzione della memoria. In questa prospettiva, la riflessione metodologica diventa il presupposto tecnico ed etico per la creazione di archivi digitali inclusivi, sostenibili e capaci di dialogare con pubblici non specialistici.
Parole chiave: Archivi digitali; Workflow metodologici; Co-costruzione della memoria; Public Engagement; Digital Humanities

ABSTRACT (ENGLISH)

From the Materiality of Sources to Shared Data: Hybrid Methods for a Participatory Digital Archive.
This paper offers a methodological reflection on the practices of Digital Humanities applied to historical and cultural contexts characterised by a high degree of heterogeneity of sources and a growing need for public impact. Starting from the REDMIX project, aimed at creating a digital archive on mixed-ancestry people and groups in the Red Sea region from the 19th century to the present day, the contribution analyses how the conscious choice of digital tools should be guided by the materiality and nature of the sources, rather than by mere technological availability. Through two case studies - drawn from the private archive of Daniela Aleggiani, consisting of hybrid typewritten diaries, photographs, and drawings related to the prisoner-of-war camps in East Africa belonging to Major Umberto Aleggiani, and from the private archive of the Vento Chiaffredo family, comprising poems and songs by the young soldier Chiaffredo Vento in Libya - the paper shows how different sources require differentiated workflows. The analysis illustrates the combined use of technologies such as OCR, HTR, computational text analysis, semantic encoding in XML-TEI and space-time visualisations, highlighting their limitations, potential and methodological implications. The case studies serve as a laboratory for the design of the REDMIX project infrastructure, conceived as a participatory digital archive and Digital Public History platform. The contribution argues that the adoption of hybrid, open and transparent workflows allow not only for a scientifically rigorous enhancement of sources, but also for their transformation into tools for public engagement and the co-construction of memory. In this perspective, methodological reflection becomes the technical and ethical prerequisite for the creation of inclusive, sustainable digital archives capable of communicating with non-specialist audiences.
Keywords: Digital archives; Methodological workflows; Co-construction of memory; Public Engagement; Digital Humanities

1. INTRODUZIONE

Le Digital Humanities si trovano oggi ad un crocevia tra la sofisticazione tecnologica e la pressante domanda di impatto pubblico. In questo contesto, per trasmettere i propri risultati scientifici, si rischia di adottare strumenti digitali in modo acritico, guidati dalla loro disponibilità più che da una riflessione metodologica adeguata. Al centro di un progetto digitale, specialmente in ambito storico e culturale, deve esserci invece un'attenta analisi delle fonti primarie e una conseguente progettazione di workflow metodologici ibridi e flessibili. Il progetto REDMIX[1], che mira a creare un archivio digitale su persone e gruppi ad ascendenza mista del Mar Rosso dall'Ottocento ad oggi, costituisce una cornice ideale per questa riflessione. Il progetto nasce con il duplice obiettivo di colmare una lacuna storiografica e costruire, attraverso un processo di co-design, un archivio digitale che favorisca e stimoli la memoria condivisa. Per realizzarlo, si dovrà affrontare una sfida centrale: l'eterogeneità delle fonti raccolte, le quali spazieranno da documenti d'archivio multilingue a fotografie, registrazioni orali e materiali epistolari. Inoltre, per comunicare i risultati del progetto oltre l'ambito accademico, la voce delle fonti dovrà essere ulteriormente amplificata attraverso operazioni e strategie di public engagement (Dal Zotto & Fusari, 2022). Per esplorare queste tematiche concretamente, il contributo si avvale di due casi studio che fungono da laboratorio metodologico. Attraverso di essi, si mostrerà come a fonti diverse corrispondano strumenti digitali diversi, e come la scelta consapevole determini a sua volta le possibilità di analisi, valorizzazione e coinvolgimento del pubblico.

2. IL PROGETTO REDMIX: UN ARCHIVIO DIGITALE PER LA CO-COSTRUZIONE DELLA MEMORIA

Il progetto REDMIX si propone di studiare persone e gruppi di ascendenza mista nel Mar Rosso, un crocevia secolare di scambi culturali e sociali spesso trascurati o parzialmente analizzati dalla storiografia africanista e asiatista, più concentrata sulle dinamiche economiche e politiche (Miran, 2005; Miran, 2017; Zaccaria, 2018). Il suo obiettivo non è solo accademico, ma sociale: quello di promuovere processi di autorappresentazione e co-produzione della memoria, sviluppati insieme alle persone di ascendenza mista, in una dimensione collettiva e digitale. L'output centrale del progetto sarà un archivio digitale concepito non come un mero repository, ma come una piattaforma di ricerca e di public engagement attiva (Frezza, 2008; Tomasi, 2016). Il progetto si impegna nella preservazione a lungo termine, nell'applicazione rigorosa dei principi FAIR (Wilkinson et al., 2016; Koster & Woutersen-Windhouwer, 2018; Corcho, Oscar et al., 2021; Hermon & Nicolucci, 2021), nell'uso di standard aperti (Aliprandi, 2011) e in un'attenzione peculiare alla sostenibilità infrastrutturale e alla lotta al digital divide (Warschauer, 2002; Ragnedda, 2019), prevedendo un'interfaccia multilingua e funzionalità offline essenziali per i contesti del Mar Rosso.
Inoltre, REDMIX è concepito sin dall'origine come un progetto di co-design. La creazione stessa dell'archivio non sarà un'operazione esclusivamente accademica e top-down, ma un processo partecipativo che coinvolgerà attivamente le comunità di riferimento. Attraverso pratiche di citizen science (European Citizen Science Association, 2015), si procederà a integrare e contestualizzare documenti d'archivio e fotografie storiche con il patrimonio vivo delle testimonianze orali e delle memorie familiari. L'archivio aspira così a diventare uno spazio dinamico di digital public history (Salvatori, 2020; Paci, 2021; Ravveduto & Salvatori, 2023). La sfida tecnico-scientifica è considerevole: gestire materiali multilingue con diversi sistemi grafici, garantire l'accessibilità in contesti a connettività limitata, rappresentare eticamente narrazioni sensibili e decolonizzare le modalità stesse di catalogazione e descrizione.

3. I CASI STUDIO: L'ETEROGENEITÀ DEI MATERIALI COME GUIDA METODOLOGICA

In linea con una prospettiva archivistica che interpreta gli archivi di persona non come insiemi documentari statici, ma come processi dinamici di costruzione della memoria individuale e collettiva, nei quali scrittura, conservazione e rielaborazione si intrecciano nel tempo (Mineo, 2024), per dare inizio a una riflessione metodologica sono stati presi in esame due fondi archivistici di natura eterogenea, le cui specifiche sfide digitali offrono un terreno di prova formativo. Analizzati in fase preparatoria, questi fondi fungeranno da prototipo per le sfide che il progetto incontrerà su scala più ampia.

[1] Sito web del progetto: redmix.eu (cons. 08/04/2026)

3.1 L'Archivio privato di Daniela Aleggiani e la storia del Maggiore Umberto Aleggiani: i racconti dai campi di prigionia

Il primo caso studio fa riferimento alla documentazione contenuta nell'Archivio privato di Daniela Aleggiani, che documenta l'esperienza del Maggiore Umberto Aleggiani, prigioniero degli inglesi in Africa orientale dal 15 maggio 1941 al 1° febbraio 1946. Durante la prigionia, trascorsa in diversi campi di concentramento tra Somalia, Etiopia e Kenya, Aleggiani tenne alcuni diari, redatti in forma dattiloscritta al rientro a Milano con l'esplicito intento di lasciare una testimonianza per la sua famiglia. La narrazione copre l'arco temporale dal 19 marzo 1941 al primo gennaio 1947, includendo sia gli anni della sua prigionia sia quelli dell'attesa del figlio Alessandro (detto Sandro), prigioniero a sua volta nei suoi stessi campi di concentramento e presenza costante nel diario. La testimonianza si articola in undici volumi, di cui i primi dieci si concludono con il suo rimpatrio, mentre l'undicesimo volume funge da appendice dedicata all'epilogo familiare e all'attesa del rimpatrio del figlio Sandro. Oltre al testo dattiloscritto, il diario è un oggetto ibrido e ricchissimo: è corredato da fotografie originali in bianco e nero dei luoghi narrati, ciascuna accompagnata da didascalie autografe, e da disegni colorati di notevole vivacità, raffiguranti scene di vita quotidiana nel campo, abitato dai locali somali, keniani ed etiopi. Questa commistione di supporti e tecniche lo rende un esempio paradigmatico di fonte complessa. La peculiarità e l'interesse di questo archivio risiedono nella sua duplice stratificazione: quella temporale, con la stesura originaria *in situ* e la successiva ribattitura a Milano nel 1946, e quella materiale, data dalla compresenza nello stesso diario di testi dattiloscritti, annotazioni manoscritte, fotografie e disegni a colori.

3.2 L'Archivio privato Famiglia Vento Chiaffredo: le canzoni di un giovane ventenne

Il secondo caso studio riguarda le carte di Chiaffredo Vento, un ragazzo di vent'anni che il 27 dicembre 1912 partì da Genova come soldato di leva per la guerra italo-turca, in Cirenaica, dove trascorse all'incirca nove mesi. Della sua esperienza ci resta una documentazione immediata e intima costituita principalmente da qualche disegno ma, soprattutto, da un corpus di canzoni e poesie manoscritte, che ci parlano di battaglie, d'amore e di guerra. Questi testi, spesso vergati a penna o a matita su fogli sciolti o su quaderni a righe, rappresentano una forma di espressione personale e spontanea, lontana da qualsiasi intento di ufficialità o revisione. La loro natura è quindi radicalmente diversa da quella raccolta nell'Archivio privato di Daniela Aleggiani: sono documenti nati nel fluire dell'esperienza, caratterizzati da una grafia corsiva e dal tratto elementare, che cerca di rispettare le regole scolastiche, lasciandosi andare a qualche errore di ortografia e a qualche inflessione dialettale. Questa materialità "grezza" pone sfide digitali di altra natura, incentrate sulla decifrabilità del segno grafico e sull'interpretazione di testi non standardizzati. Inoltre, anche in questo caso emerge una stratificazione temporale: a una stesura originaria *in situ* seguì una trascrizione, presumibilmente redatta durante il ricovero ospedaliero al rientro dalla Cirenaica, prima della partenza di Vento per il fronte austriaco, dove morì.

4. WORKFLOW METODOLOGICO: APPROCCI COMUNI E DIFFERENZIATI

Di fronte a queste tipologie di fonti, lo storico digitale deve selezionare e orchestrare una serie di tecniche e strumenti, il cui impiego è dettato, in prima istanza, dalla natura materiale del documento.
Un approccio comune e preliminare in entrambi i casi è la digitalizzazione ad alta risoluzione (Allegrezza, 2021). Questo processo non serve solo a creare un duplicato digitale a fini conservativi, ma costituisce una vera e propria strategia di preservazione dall'usura fisica (IFLA, 2015; Istituto centrale per la digitalizzazione del patrimonio culturale – Digital Library, 2022), dal momento che la disponibilità di una copia digitale limita sensibilmente la necessità di manipolare gli originali fragili. È importante sottolineare il verbo "limitare" e non "impedire", poiché lo studio diretto dell'oggetto materiale rimane in molti casi insostituibile. Solo l'originale può restituire pienamente informazioni cruciali come la leggerezza o la trasparenza della carta, la profondità del segno inciso dalla penna, la presenza di filigrane visibili solo in controluce, o le sottili sfumature cromatiche che anche la migliore digitalizzazione può alterare. Tuttavia, è proprio il duplicato digitale ad aprire la strada a quelle procedure di elaborazione e valorizzazione che permettono al documento di "parlare" a un pubblico più ampio, inclusi i non addetti ai lavori, e di essere interrogato attraverso nuove domande di ricerca.

4.1 I diari di Umberto Aleggiani: OCR, analisi del testo e mappatura

Un primo approccio esemplificativo consiste nell'applicare la tecnologia Optical Character Recognition (OCR) al testo dattiloscritto, per ottenere una versione digitale leggibile dalla macchina (Ballerino, 2022).

Su questo estratto testuale, strumenti di analisi computazionale, come script sviluppati in Python, permetterebbero di identificare la frequenza delle parole, individuare entità ricorrenti e rivelare pattern narrativi (AlShammari, 2023). Questa analisi potrebbe servire a mappare la geografia umana e fisica del diario, identificando in modo sistematico i luoghi salienti, i titoli militari o i prestiti linguistici da altre lingue. Tuttavia, l'applicazione dell'OCR a questo materiale incontra limiti strutturali che sono essi stessi dati significativi. Le correzioni e le integrazioni apportate a mano da Aleggiani sfuggono al riconoscimento automatico, così come le didascalie manoscritte delle fotografie, alcune delle quali sono di difficile lettura anche per l'occhio umano a causa del basso contrasto. Inoltre, interruzioni di riga o impaginazioni irregolari possono compromettere il riconoscimento corretto delle parole. Questi limiti non invalidano la tecnica, ma richiedono un workflow ibrido che preveda una fase di post-correzione e integrazione manuale.
Un secondo approccio, complementare, consisterebbe nella creazione di una mappa digitale per visualizzare gli spostamenti di Aleggiani tra i vari campi di prigionia. Strumenti di diversa natura, ovvero open-source come Palladio[2] oppure a pagamento con piani gratuiti per il mondo studentesco e accademico come CARTO[3], permetterebbero di georeferenziare i luoghi menzionati nei diari, creando una rappresentazione spaziale dinamica e temporale della sua esperienza. Questa visualizzazione, di immediata comprensione, è uno strumento potentissimo sia per l'analisi storica, per verificare percorsi e densità degli eventi, che per il public engagement (Dal Zotto & Fusari, 2022), offrendo al visitatore dell'archivio digitale una chiave di lettura intuitiva e coinvolgente della narrazione testuale.

4.2 Le carte di Chiaffredo Vento: HTR, codifica semantica e visualizzazione

Per questi materiali, l'approccio deve necessariamente differenziarsi. L'OCR, ottimizzato per testi a stampa o dattiloscritti, è qui inefficace. La strada alternativa diventa l'Handwritten Text Recognition (HTR) (Cascianelli et al., 2021), implementata attraverso piattaforme come Transkribus (Muehlberger et al., 2019) o eScriptorium (Mazzagufo et al., 2025). Questi strumenti richiedono un investimento iniziale consistente: per addestrare un modello di riconoscimento accurato sulla grafia specifica di Chiaffredo Vento, per quanto piuttosto chiara e riconoscibile, è necessario trascrivere manualmente un campione significativo di pagine. La difficoltà non è banale: i testi sono ricchi di errori grammaticali e presentano un'impaginazione libera, con "a capo" improvvisi e testo disposto in più colonne o spezzoni. Addestrare un modello a riconoscere queste caratteristiche significa, in sostanza, insegnare alla macchina a leggere non solo i segni, ma anche le idiosincrasie espressive dell'autore, preservando così gli errori come dati storici e non come rumore da eliminare. Una volta ottenuta la trascrizione via HTR, l'output (che può essere anche un file XML) diventa la base per successive operazioni di valorizzazione. Questo file può essere arricchito con markup semantico. Attraverso la marcatura di entità nominate (nomi di persona, luoghi geografici, ecc.), il testo da sequenza di parole diventa una rete strutturata di dati interrogabili. Il formato XML così strutturato può quindi essere implementato all'interno di un ambiente di pubblicazione digitale dedicato ai testi letterari e documentari. Uno strumento particolarmente adatto a questo scopo è EVT - Edition Visualization Technology[4] (Rosselli Del Turco, 2019; Cacioli et al., 2022), un software specificamente progettato per creare edizioni digitali a fronte di testi codificati in XML-TEI (Spadini & Turska, 2019; Fenu & Tancredi, 2022). Caricando il file XML su EVT, si realizza un'edizione digitale che affianca all'immagine facsimile della pagina la sua trascrizione, in una visualizzazione chiara e navigabile. La piattaforma permette non solo la semplice visualizzazione a fronte, ma anche la gestione di apparati critici, note, e l'esplorazione interattiva del testo attraverso gli elementi semanticamente marcati. La codifica trasforma il documento in un oggetto di studio stratificato, accessibile sia per analisi linguistiche e tematiche approfondite, sia per una fruizione pubblica guidata, dove un lettore non specialista può esplorare, ad esempio, tutte le poesie che citano un determinato luogo tramite filtri e funzionalità di ricerca.

4.3 VERSO REDMIX: UN METODO IBRIDO PER UN ARCHIVIO PARTECIPATIVO

Le esperienze metodologiche condotte sull'Archivio privato di Daniela Aleggiani e sull'Archivio privato della Famiglia Vento Chiaffredo non costituiscono esercizi isolati, ma rappresentano la messa a punto di un approccio operativo che determina direttamente lo sviluppo del progetto REDMIX. Dall'esperienza con

[2] https://hdlab.stanford.edu/palladio/ (cons. 08/04/2026)
[3] https://carto.com/ (cons. 08/04/2026)
[4] https://www.dish.unito.it/it/progetti/filologia-digitale/evt-edition-visualization-technology (cons. 08/04/2026)

strumenti specifici come Palladio, EVT Viewer o Transkribus, non si trae la prescrizione del loro uso obbligato, ma si estraggono categorie funzionali essenziali: la necessità di strumenti per la visualizzazione spazio-temporale, per l'edizione digitale critica ed esplorabile, e per il riconoscimento del testo manoscritto. Sono queste categorie, e non i singoli software, a dover essere implementate nell'infrastruttura di REDMIX. Le sfide affrontate, dalle limitazioni dell'OCR sui testi ibridi alla necessità di addestrare modelli di HTR per grafie specifiche, delineano un principio guida fondamentale: non esiste uno strumento digitale universale. La scelta, l'integrazione e talvolta l'adattamento degli strumenti devono essere plasmati dalle caratteristiche intrinseche delle fonti. Per REDMIX, questo principio si traduce nella progettazione di workflow distinti e paralleli. Il futuro archivio dovrà gestire una molteplicità di tipologie documentarie ancora maggiore di quella dei casi studio: testi d'archivio multilingue a stampa e manoscritti (in caratteri arabi, latini e geez), fotografie storiche e contemporanee, registrazioni audio di storie orali, materiali epistolari. Ad ognuna di queste categorie dovrà corrispondere una pipeline di trattamento specifica, ispirata alle lezioni apprese. Per i documenti amministrativi coloniali a stampa, si potrà implementare un workflow simile a quello dell'Archivio privato di Daniela Aleggiani, basato su OCR e successiva analisi computazionale del testo. Per le lettere o i diari manoscritti in grafie non latine, sarà necessario ricorrere a strumenti di HTR e, probabilmente, intraprendere campagne di addestramento di modelli specifici, come nel caso Vento, per gestire le varianti grafiche locali. Le fotografie richiederanno un approccio centrato su standard come l'International Image Interoperability Framework (IIIF) (Salarelli, 2017) per una visualizzazione ad alta definizione e interoperabile, mentre per le registrazioni orali si dovranno stabilire protocolli per la trascrizione, la traduzione e l'annotazione semantica. Tuttavia, la vera innovazione metodologica di REDMIX risiede nel fatto che tali workflow tecnici non sono concepiti come processi chiusi e confinati al gruppo di ricerca. Al contrario, sono progettati per essere aperti e diventare essi stessi strumenti di public engagement e di co-costruzione del sapere. La fase di post-correzione dell'OCR, necessaria per gestire le inevitabili imperfezioni del riconoscimento automatico, può essere trasformata in un'attività di citizen science (European Citizen Science Association, 2015; Ballerino, 2022), coinvolgendo studenti, appassionati di storia o membri delle comunità di riferimento. Allo stesso modo, la trascrizione e l'annotazione semantica di documenti manoscritti o la geolocalizzazione di fotografie storiche possono diventare esercizi partecipativi in contesti educativi o in laboratori territoriali[5]. In questo modo, il processo metodologico di trasformazione della fonte in dato passa da essere tradizionalmente "invisibile" e specialistico ad un'occasione di coinvolgimento, formazione e arricchimento collettivo dell'archivio. L'applicazione rigorosa dei principi FAIR a tutti i dati prodotti, siano essi metadati, testi trascritti o annotazioni, non è quindi solo una buona pratica di scienza aperta, ma la condizione tecnica che rende possibile questo riuso creativo e partecipato. Un dato ben strutturato e descritto può essere trovato, compreso e rielaborato non solo da un altro ricercatore, ma anche da un insegnante per preparare una lezione o da una persona con ascendenza mista per ricostruire la propria storia familiare.

5. CONCLUSIONI: DALLA PLASTICITÀ DEL METODO ALL'INCLUSIONE DELLA MEMORIA

Il percorso tracciato da questo contributo muove dalla concretezza materiale dei documenti, in questo caso i diari ibridi di Aleggiani e i testi manoscritti di Vento, per arrivare alla progettazione di un archivio digitale complesso e partecipativo come quello del progetto REDMIX. Il filo conduttore è la plasticità metodologica: la capacità critica di selezionare, combinare e adattare strumenti digitali in risposta alle specificità delle fonti, senza mai subirli acriticamente. Questa attenzione al metodo non è un tecnicismo fine a se stesso, ma si rivela la premessa indispensabile per realizzare gli obiettivi più ambiziosi delle Digital Humanities. È proprio in questa ottica di validazione metodologica che va compresa la scelta, in questa fase, di lavorare su fonti esclusivamente in lingua italiana. Tale decisione non rappresenta una limitazione, ma una strategia di fondo. L'italiano ha costituito un terreno di prova controllato permettendo di isolare, analizzare e risolvere le sfide tecniche e procedurali fondamentali, dalla scelta tra OCR e HTR alla progettazione di workflow per la georeferenziazione e la codifica semantica, senza l'ulteriore e complesso livello della mediazione linguistica interculturale. Padroneggiare queste complessità su un corpus omogeneo è stato il presupposto necessario per progettare soluzioni infrastrutturali che, in una fase successiva, possano essere estese e adattate alla gestione del multilinguismo radicale (arabo, amarico, tigrino, ecc.) che caratterizza il bacino del Mar Rosso e che costituisce il cuore della sfida per

[5] Si veda l'iniziativa "Collezioni fotografiche del colonialismo italiano (1861-1960)" (cons. 08/04/2026)

REDMIX. I casi studio italiani hanno dunque fornito la mappa delle questioni da risolvere, offrendo un modello di adattamento piuttosto che un modello preconfezionato. Solo un processo di digitalizzazione e di trasformazione in dati che sia consapevole dei limiti e delle potenzialità di ogni strumento può generare risorse digitali di qualità, affidabili e durevoli. Solo dati costruiti con questo rigore possono essere efficacemente condivisi, interoperare con altri archivi e, soprattutto, essere messi a disposizione di pubblici non specialistici in forme accessibili e coinvolgenti, come mappe interattive o edizioni digitali esplorabili. La lotta al digital divide, esplicita negli obiettivi del progetto REDMIX con un'interfaccia multilingua e funzionalità offline, inizia proprio qui: nella progettazione di una architettura dell'informazione che non escluda a priori chi ha una connettività limitata o competenze digitali basilari. In conclusione, REDMIX si propone come un modello operativo che dimostra come la riflessione sul metodo, saldamente ancorata alla materialità delle fonti, costituisca il ponte essenziale tra l'analisi storica specialistica e la co-costruzione democratica della memoria. La cura per la specificità di un diario o di una poesia diventa il fondamento etico e tecnico per costruire archivi che siano sia robusti depositi del patrimonio culturale che piazze digitali inclusive. In tali spazi, la molteplicità delle voci (le fonti d'archivio, i cittadini, i ricercatori) può finalmente trovare espressione e dialogare, contribuendo a decolonizzare le narrazioni storiche e a scrivere, collettivamente, una storia più ricca e rappresentativa del complesso passato del Mar Rosso e dei suoi abitanti.

FINANZIAMENTI

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.
This work is supported by ERC Grant REDMIX Agreement 101124725.

RINGRAZIAMENTI

Si ringraziano le famiglie Aleggiani e Vento per aver permesso la consultazione dei loro archivi privati per la conduzione di queste ricerche storiche.
Si ringrazia Valentina Fusari per i preziosi suggerimenti forniti durante la stesura di questo lavoro.

BIBLIOGRAFIA

Aliprandi, S. (2011). *Apriti standard!*. Milano: Ledizioni. https://doi.org/10.4000/books.ledizioni.171

Allegrezza, S. (A c. Di). (2021). La digitalizzazione del patrimonio culturale. Linee guida, standard, esperienze. Collana di studi archivistici. Civita Editoriale.

AlShammari, A. F. (2023). Implementation of text similarity using word frequency and cosine similarity in Python. *International Journal of Computer Applications*, *185*(36), 54-59

Ballerino, J. (2022). Artificial Intelligence and Pattern Recognition Technologies for Cultural Heritage: Involvement of Optical Character Recognition Software for Citizen Science in the processes for Crowdsourcing of Ancient Italian Texts. (Dissertation). Retrieved from https://urn.kb.se/resolve?urn=urn:nbn:se:lnu:diva-114820

Cacioli, E., Di Pietro, C., Maenza, S., Rosselli Del Turco, R., & Zenzaro, S. (2022). There and back again: What to expect in the next EVT version. In F. Ciracì, G. Miglietta, & C. Gatto (A c. Di), *AIUCD 2022—Culture digitali. Intersezioni: Filosofia, arti, media. Proceedings della 11a conferenza nazionale, Lecce, 2022* (pp. 212–217). Quaderni di Umanistica Digitale.

Cascianelli, S., Cornia, M., Baraldi, L., Piazzi, M. L., Schiuma, R., & Cucchiara, R. (2021). Learning to Read L'Infinito: Handwritten Text Recognition with Synthetic Training Data. In N. Tsapatsoulis, A. Panayides, T. Theocharides, A. Lanitis, C. Pattichis, & M. Vento (A c. Di), *Computer Analysis of Images and Patterns. CAIP 2021* (Vol. 13053, pp. 340–350). Springer, Cham. https://doi.org/10.1007/978-3-030-89131-%25202_31

Corcho, Oscar, Eriksson, M., Kurowski, K., & Ojsteršek, M. (2021). *EOSC Interoperability Framework: Report from the EOSC Executive Board Working Groups FAIR and Architecture*. Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko. https://www.afs.enea.it/project/madia/Documenti/web/docs/EOSC-Interoperability-Framework-KI0221055ENN.pdf

Dal Zotto, E., & Fusari, V. (2022). Diamo rifugio ai talenti. Nuove forme di inclusione e internazionalizzazione attraverso l'Università di Pavia. *Antropologia Pubblica*, *8*(1), 225-244.

European Citizen Science Association. (2015). *Ten principles of citizen science*. https://www.ecsa.ngo/10-principles/

Fenu, C., & Tancredi, G. (2022). XML-TEI: Un modello per la filologia d'autore. In F. Ciracì, G. Miglietta, & C. Gatto (A c. Di), *AIUCD 2022—Culture digitali. Intersezioni: Filosofia, arti, media. Proceedings della 11a conferenza nazionale, Lecce, 2022* (pp. 218–222). Quaderni di Umanistica Digitale. https://doi.org/10.6092/unibo/amsacta/6848

Frezza, G. (A c. Di). (2008). L'arca futura. Archivi mediali digitali, audiovisivi, web. Meltemi Editore.

Hermon, S., & Nicolucci, F. (2021). FAIR Data and Cultural Heritage Special Issue Editorial Note. *International Journal on Digital Libraries*, *22*(3), 251–255. https://doi.org/10.1007/s00799-021-00309-8

IFLA. (2015). *Linee guida per pianificare la digitalizzazione di collezioni di libri rari e manoscritti* (Gruppo di lavoro della Biblioteca digitale BEIC, Trad.). IFLA.

Istituto centrale per la digitalizzazione del patrimonio culturale – Digital Library (A c. Di). (2022). *Piano nazionale di digitalizzazione del patrimonio culturale* (v. 1.0). https://docs.italia.it/italia/icdp/icdp-pnd-digitalizzazione-docs/it/v1.0-giugno-2022/index.html

Koster, L., & Woutersen-Windhouwer, S. (2018). FAIR Principles for Library, Archive and Museum Collections: A Proposal for Standards for Reusable Collections. *The Code4Lib Journal*, (40). https://journal.code4lib.org/articles/13427

Mazzagufo, L., Sichera, P., Bruno, D., Cristofaro, S., Del Grosso, A. M., & Spampinato, D. (2025, October). An Infrastructural Solution for Digital Publication starting from Automatic Layout and Text Recognition: Insights from Italian Literary Journals. In *2025 IEEE 8th Congress on Information Science and Technology (CiSt)* (pp. 494-499). IEEE.

Mineo, L. (2024). *Vivere, archiviare, scrivere. Brevi riflessioni a margine di una recente pubblicazione*. Archivi, 19(2), 119–125.

Miran, J. (2017). The Red Sea. In D. Armitage, A. Bashford, & S. Sivasundaram (A c. Di), *Oceanic Histories* (1ª ed., pp. 156–181). Cambridge University Press. https://www.cambridge.org/core/product/identifier/9781108399722%23CN-bp-6/type/book_part

Miran, J. (2005). A Historical Overview of Islam in Eritrea. *Die Welt Des Islams*, *45*(2), 177–215. http://www.jstor.org/stable/1571280

Muehlberger, G., Seaward, L., Terras, M., Ares Oliveira, S., Bosch, V., Bryan, M., Colutto, S., Déjean, H., Diem, M., Fiel, S., & Gatos, B. (2019). Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study. *Journal of Documentation*, *75*(5), 954–976. https://doi.org/10.1108/JD-07-2018-0114

Ragnedda, M. (2019). 2. Conceptualising the digital divide. In B. Mutsvairo & M. Ragnedda (Ed.), *Mapping the Digital Divide in Africa: A Mediated Analysis* (pp. 27-44). Amsterdam: Amsterdam University Press. https://doi.org/10.1515/9789048538225-004

Ravveduto, M., & Salvatori, E. (2023). Storia digitale e digital public history: Le novità di un antico mestiere. In *Digital Humanities: Metodi, strumenti, saperi* (pp. 229–254). Carocci.

Rosselli Del Turco, R. (2019). Designing an advanced software tool for Digital Scholarly Editions: The Inception and Development of EVT (Edition Visualization Technology). *Textual Cultures*, *12*(2), Articolo 2. https://doi.org/10.14434/textual.v12i2.27690

Salarelli, A. (2017). International image interoperability framework (IIIF): A panoramic view. *JLIS: Italian Journal of Library, Archives and Information Science= Rivista italiana di biblioteconomia, archivistica e scienza dell'informazione: 8, 1, 2017*, 50-66.

Salvatori, E. (2020). Digital Public History Inside and Outside the Box. *Magazén*, *2*. https://doi.org/10.30687/mag/2724-3923/2020/02/003

Spadini, E., & Turska, M. (2019). XML-TEI Stand-off Markup: One Step Beyond. *Digital Philology: A Journal of Medieval Cultures*, *8*, 225–239. https://doi.org/10.1353/dph.2019.0025

Tomasi, F. (2016). Edizioni o archivi digitali? Knowledge sites e apporti disciplinari. In P. Italia & C. Bonsi (A c. Di), Edizioni critiche digitali (pp. 129–136). Sapienza Università Editrice. https://www.editricesapienza.it/sites/default/files/5369_Italia_Bonsi_EdizioniCriticheDigitali.pdf

Warschauer, M. (2002). Reconceptualizing the Digital Divide. *First Monday*, *7*(7). https://doi.org/10.5210/fm.v7i7.967

Wilkinson, M., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(160018). https://doi.org/10.1038/sdata.2016.18

Zaccaria, M. (2018). Migration from the Horn of Africa: Rethinking space and time. In *Migration and the Contemporary Mediterranean Shifting Cultures in Twenty-First-Century Italy and Beyond* (pp. 263–280). Oxford: Peter Lang.

[2/A] Large Language Models: applicazioni umanistiche

Scholarly Opinion Mining using LLMs and Knowledge Graphs: the case of the Van den vos Reynaerde

ANDREA SCHIMMENTI[1], JORIS VAN ZUNDERT[2], FABIO VITALI[3], MARIEKE VAN ERP[4]

[1]LILEC, UNIVERSITÀ DI BOLOGNA, ITALY, ANDREA.SCHIMMENTI2@UNIBO.IT
[2]COMPUTATIONAL LITERARY RESEARCH - HUYGENS INSTITUTE, KNAW HUMANITIES CLUSTER, NETHERLANDS, JORIS.VAN.ZUNDERT@HUYGENS.KNAW.NL
[3] DISI, UNIVERSITÀ DI BOLOGNA, ITALY, FABIO.VITALI@UNIBO.IT
[4]DHLAB - KNAW HUMANITIES CLUSTER, NETHERLANDS, MARIEKE.VAN.ERP@DH.HUC.KNAW.NL

ABSTRACT (ENGLISH)

Extracting knowledge graphs from full-length scholarly articles represents an open challenge in Digital Humanities, particularly for enabling computational analysis of interpretative debates across large corpora. We present a Retrieval-Augmented Generation pipeline that extracts nanopublications implementing the Digital Hermeneutics model (factual knowledge, scholarly assertions, hermeneutical context, and computational provenance) guided by domain expert questions. Applied to five articles on the *Van den vos Reynaerde* authorship debate using CIDOC CRM, HiCO and PROV-O, the structured representation enables comparative analysis of interpretative patterns, argumentative strategies, and scholarly consensus formation across publications, supporting distant reading approaches to philological debates.
Keywords: Retrieval-Augmented Generation; Large Language Models; Scholarly Argumentation; Semantic Web; CIDOC CRM;

ABSTRACT (ITALIANO)

Estrazione di argomentazioni filologiche tramite LLM e grafi della conoscenza: il caso del Van den vos Reynaerde. L'estrazione di grafi di conoscenza da articoli scientifici completi rappresenta una sfida aperta nell'Informatica Umanistica, specialmente per abilitare analisi computazionali di dibattiti interpretativi su grandi corpora. Presentiamo una pipeline basata su Retrieval-Augmented Generation che, guidata da domande di un esperto, estrae nanopubblicazioni implementando il modello Digital Hermeneutics su quattro livelli: conoscenze fattuali, asserzioni scientifiche, contesto ermeneutico e provenienza computazionale. Applicata a cinque articoli sul dibattito autoriale del *Van den vos Reynaerde* usando CIDOC CRM, HiCO e PROV-O, i grafi di conoscenza permettono analisi comparative di pattern interpretativi, strategie argomentative e formazione del consenso scientifico tra pubblicazioni, supportando approcci di distant reading ai dibattiti filologici.
Parole chiave: Retrieval-Augmented Generation; Large Language Models; Dibattito Scientifico; Semantic Web; CIDOC CRM;

1. INTRODUCTION

The extraction of structured knowledge from scholarly discourse represents a persistent challenge in digital humanities research, yet one that promises to transform how interpretative debates are analyzed across large corpora. While close reading remains foundational to humanities scholarship (Bode, 2020), the accumulation of decades of specialized publications creates opportunities for computational approaches to scholarly opinion mining—identifying patterns of consensus, divergence, and argumentative evolution that emerge only at scale. Large Language Models (LLMs) have demonstrated capacity for processing concise textual sources such as encyclopedia articles or archival finding aids (Hyvönen et al., 2017; Giagnolini et al., 2025), yet their application to full-length academic publications introduces distinct methodological complexities. Academic articles extend arguments across tens of thousands of words, distribute discourse across multiple structural sections, employ specialized disciplinary terminology, and embed references to broader scholarly networks that may remain implicit. Recent advances in text-to-knowledge graph (text-to-KG) extraction have demonstrated potential across diverse Cultural Heritage (CH) domains (Hyvönen et al., 2017; Lisena et al., 2022; Morales Tirado et al., 2024), yet these implementations have focused primarily on encyclopedic summaries, archival metadata, or semi-structured biographical narratives rather than full-length scholarly articles. The conversion of

interpretative discourse into machine-readable formats remains underexplored, limiting the application of distant reading methodologies to the analysis of scholarly argumentation itself. This work presents a Retrieval-Augmented Generation (RAG) pipeline that extracts scholarly interpretations from full-length articles on the *Van den vos Reynaerde* authorship debate as structured nanopublications following the Digital Hermeneutics model (Daquino et al., 2020), using as ontologies CIDOC CRM,[1] HiCO,[2] FaBiO[3] and PROV-O.[4] Traditional text-to-KG approaches have achieved precision rates from 75% to 99% depending on domain and source type (Hyvönen et al., 2017; Koho et al., 2020), but performance degrades substantially with scholarly articles. Their length requires mechanisms to locate relevant passages for targeted content retrieval as they can easily surpass the common context windows of LLMs (Fan et al., 2024). The code for the experiment is available on a GitHub repository.[5]

2. SETTING

Van den vos Reynaerde is a Middle Dutch beast epic dated to the mid-thirteenth century, composed by an unknown author, whose speculated name is *Willem*, given a verse in the incipit ("Willem die Madoc maecte") and an acrostic signature which repeats the name Willem. (Bouwman & Besamusca, 2009). Scholarship has generated debates regarding authorship, temporal localization, geographic origin (focusing on the Flemish region), and the relationship between the Middle Dutch text and its French source material, the *Roman de Renart* (Van Daele, 2005). The most active debate concerns the identification of Willem with the historic Willem van Boudelo, a thirteenth-century Cistercian monk (Van Daele, 2005). The experiment started from a set of competency questions (CQs) formulated by a computer scientist and philologist, expert in the *Van den vos Reynaerde,* designed to capture the different facets of the authorship debate: (1) authorship attribution; (2) temporal localization (3) spatial localization; (4) thematic and referential content; (5) cultural and institutional affiliations; (6) expertise and competencies; (7) methodological approaches; (8) epistemic stance. The domain expert suggested five articles, of which two — Wackers (2016) and Wackers et al. (2000) — do not advance authorship attributions to the *Reynaerde* corpus, and were included to test whether the pipeline correctly produces empty or minimal assertion layers rather than generating false positives when the source text does not support the target relation type. The selected articles for this experiment were:

- *Of Reynaert the Fox: Text and facing translation of Van den vos Reynaerde*, a comprehensive critical edition, addressing the authorship question (Bouwman & Besamusca, 2009);
- *Historiciteit en chronologie in 'Van den vos Reynaerde'*, which examines evidence to support the identification of Willem van Baudelo as the author (Peeters, 1973)
- *De robotfoto van de Reynaertdichter*, which analyzes the Willem van Baudelo hypothesis through analysis of Cistercian connections, courtly context, and source material (Van Daele, 2005);
- *Wat staat er eigenlijk? Over het editeren van Van den vos Reynaerde*, on the epic's editorial history (Wackers, 2016);
- *Medieval French and Dutch Renardian Epics: Between Literature and Society*, on the intertextual practices and social contexts across the French and Dutch traditions (Wackers et al., 2000);

The **Digital Hermeneutics model** (Daquino et al., 2020) was selected as the representational framework because humanities scholarly claims operate at distinct epistemic levels that a flat triple structure cannot differentiate. For instance, a statement saying Willem likely composed the *Reynaerde* in a Cistercian convent carries a different epistemic status than the statement that the *Reynaerde* exists as a Middle Dutch text, and both differ from the metadata describing who made the claim and on what evidential basis. The model addresses this through a layered **nanopublication** architecture (Groth et al., 2010) in which each layer captures a distinct type of knowledge.
Layer 0 ("Factual Data") contains background knowledge that participants in scholarly discourse treat as established — in this case, the existence of the text, its language, and the manuscript tradition.
Layer 1 ("Scholarly Assertions") captures interpretative claims about the *Reynaerde*'s creation and proposed information about Willem, using **CIDOC CRM** (see Fig. 1, Fig. 2). Competence and expertise in

[1] https://cidoc-crm.org/
[2] https://marilenadaquino.github.io/hico/
[3] https://sparontologies.github.io/fabio/current/fabio.html
[4] The image's source is Daquino et al., (2020). It is shown here only for practical purposes
[5] https://github.com/aschimmenti/digital_hermeneutics_reynaerde

different domains are modelled as crm:E55_Type (see Figure 2), following the CIDOC official forum (see Fig. 2).[6]

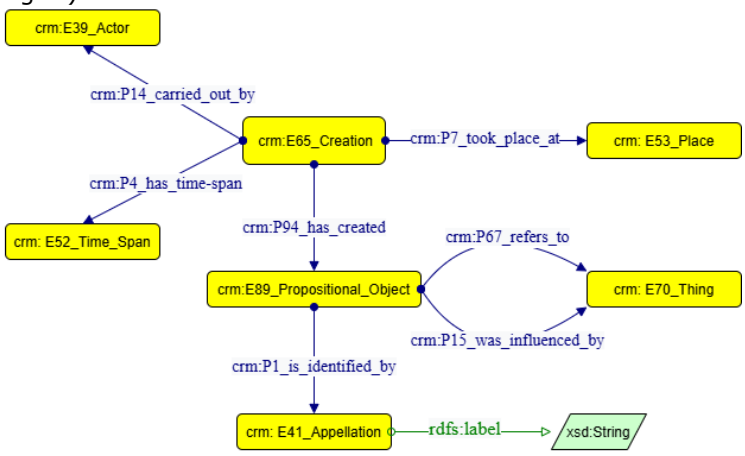


Figure 1. Layer 1 (Scholarly Assertions) representation of a creation event.[7]

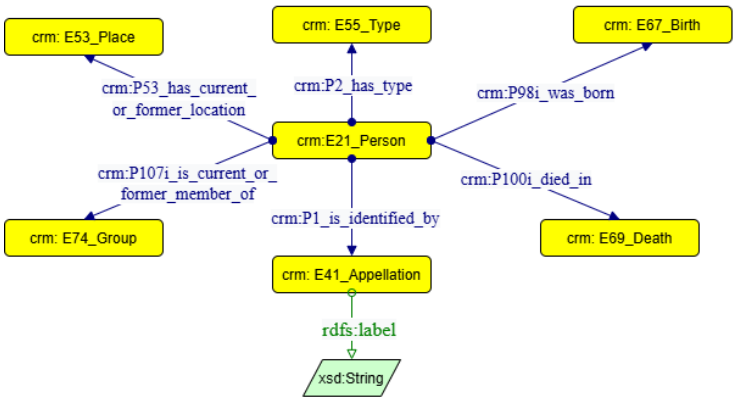


Figure 2. Layer 1 (Scholarly Assertions) representation of biographical information associated with an author.[8]

Layer 2 ("Hermeneutical Context") implements the HiCO model (Daquino & Tomasi, 2015) to represent the interpretation act itself: each scholarly article generates an instance of hico:InterpretationAct, including hico:hasInterpretationType, hico:hasInterpretationCriterion, and hico:hasCertaintyLevel. It captures, for instance, whether Van Daele (2005) grounds his attribution in philological or institutional evidence, and with what expressed degree of certainty (see Fig. 3).[9]

[6] CIDOC CRM Discussion about the language proficiency of a crm:E39_Actor, https://cidoc-crm.org/Issue/ID-429-p72-has-language

[7] The central node instantiates crm:E65_Creation, linked via crm:P14_carried_out_by to the proposed author entity (Willem / Willem van Boudelo depending on the source article) and via crm:P15_was_influenced_by to works and cultural entities identified as influences. Readers should note the distinction between the work-level node (the Reynaerde as crm:E33_Linguistic_Object) and the influence relations, which vary across articles

[8] Person-level attributes — name, occupation, institutional affiliation, geographic location, birth and death — are attached to a crm:E21_Person node. Occupational and expertise categories are typed as crm:E55_Type instances. The affiliation relation (crm:P107i_is_current_or_former_member_of linking to a crm:E74_Group instantiating, e.g., the Cistercian order) would represent information such as the historically contested claim that distinguishes Van Daele (2005) and Peeters (1973) from Bouwman & Besamusca (2009).

[9] The image's source is Daquino et al., (2020). It is shown here only for practical purposes

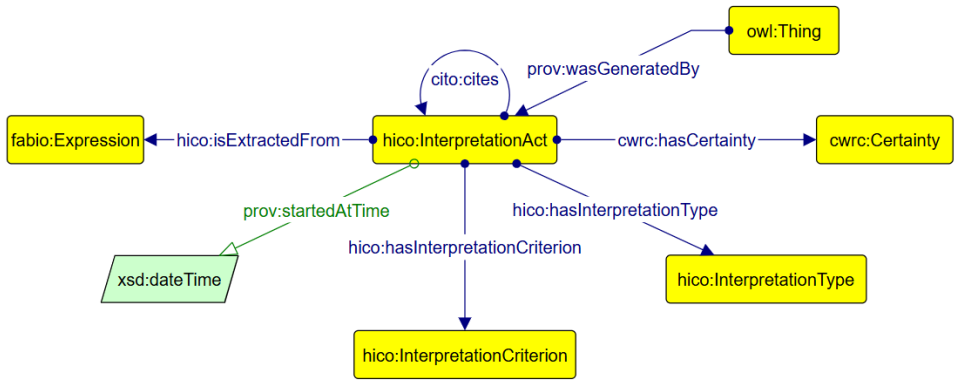
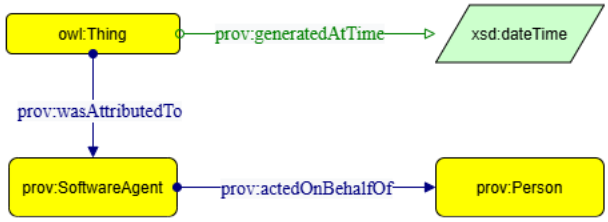


Figure 3. Layer 2 (Hermeneutical Context) representation using the HiCO ontology (Daquino & Tomasi, 2015)

Layer 3 ("Provenance") documents the computational origin of each assertion using PROV-O, recording software agents (prov:wasAttributedTo), timestamps (prov:generatedAtTime), and procedural details, enabling traceability from any extracted triple back to the pipeline stage that produced it (see Fig. 4).



Layer 3 (Provenance) representation using PROV-O.

3. PIPELINE

Retrieval-Augmented Generation (RAG) was implemented within the pipeline to address the challenge that scholarly articles extend arguments across tens of thousands of words (see Fig. 5).

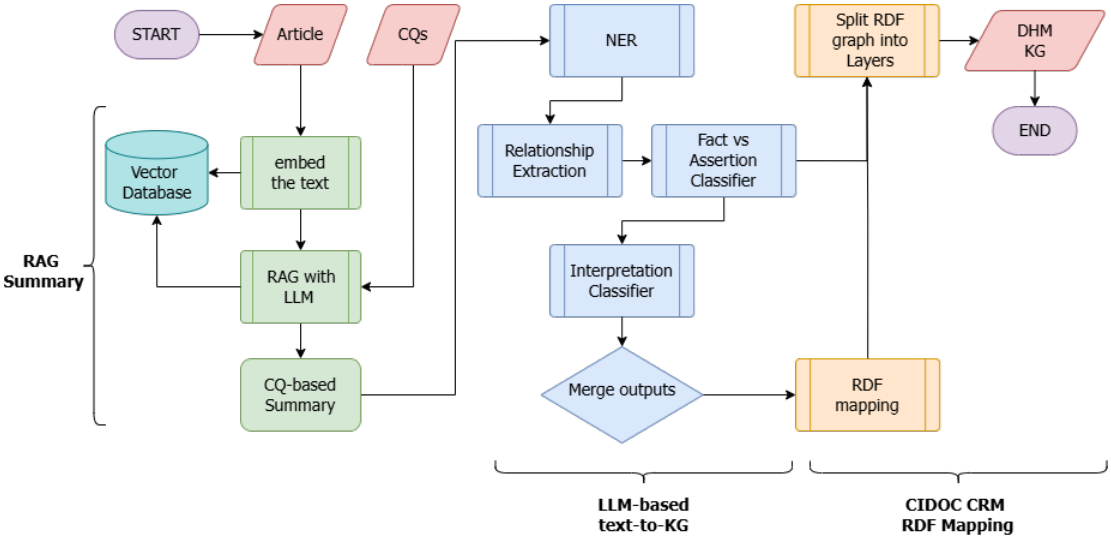


Figure 5. Pipeline Flowchart

RAG combines retrieval mechanisms with generative language models to ground generated text in retrieved evidence (Lewis et al., 2020), constraining LLM outputs to information grounded in retrievable passages from the source document. In this experiment, RAG is used to steer the summarization of the article by selecting only relevant passages for the question at hand. The pipeline implements the following stages:

1. **Document Preprocessing and Question Answering.** The system embeds the articles using a FAISS index (Douze et al., 2025) with contextualized embeddings generated by Voyage AI (model: voyage-context-3).[10] Sequentially, the Competency Questions (CQs) identified by the domain expert are used to retrieve selective summaries of the article. For each question, the system: (1) generates a query embedding; (2) retrieves the *top-k* most relevant chunks from the FAISS index; (3) re-ranks retrieved chunks using Voyage AI's reranking model (rerank-2); (4) constructs a prompt that includes document metadata, retrieved passages, previously answered questions, and few-shot examples; (5) generates an answer to the question using GPT-4o-mini.[11]
2. **Entity Extraction.** Entity extraction identifies entities mentioned in the answers using GPT-4o-mini. The entities belong to fourteen types: person, reference, role, place, work, date, historical_context, organization, language, methodology, theory, genre, concept, event, group, activity, characteristic, and material.
3. **Relationship Extraction and Interpretation Metadata.** As in the previous steps, the relations mirror the CIDOC CRM properties selected above (e.g. created_by, influenced_by, educated_at, has_occupation…). This stage also assigns a claim_type attribute to each relation, so that each relation can be connected back to the specific layer (established_fact => Layer 0, authorial_argument => Layer 1). The final sub-step extracts relations about the interpretation (Layer 2).
4. **RDF Generation.** The final stage transforms the extracted KG into RDF triples following the ontological model, generating nanopublications in TriG format with separate named graphs for facts (Layer 0), assertions (Layer 1), interpretation (Layer 2) and provenance (Layer 3).

4. RESULTS

The text-to-KG pipeline processed five scholarly articles through the RAG-based extraction system. The corpus includes three articles proposing authorship attributions and two articles on editorial and intertextuality, which served as negative examples to evaluate the pipeline's robustness in distinguishing relevant from out-of-scope content (see Table 1 for the number of triples per document).

Source	Layer 0	Layer 1	Layer 2
Van Daele (2005)	44	17	9
Bouwman & Besamusca (2009)	31	18	17
Peeters (1973)	53	9	6
Wackers (2016)	28	5	7
Wackers et al. (2000)	32	0	8
Total	188	49	47

Table 1. Extraction statistics for nanopublications

The three main articles (Van Daele, Bouwman & Besamusca, Peeters) discussed substantial arguments. Wackers et al. (2000) correctly resulted in an empty assertion layer, while Wackers (2016), focusing on editorial history and manuscript tradition, produced minimal assertions that discussed authorship (3 out of 5 triples), while two represented false positives regarding manuscript references.

Competency Question	Van Daele	Besamusca	Peeters	Wackers (2016)	Wackers (2000)
Author proposed?	Yes	Partial	Yes	No	No
Creation date?	Yes	Partial	Yes	No	No

[10] https://www.voyageai.com/
[11] The model was chosen as one of the cheapest mid-size models for inference while maintaining overall competitive performance and the ability to return answers following strict JSON schemas (estimated size is around 20B parameters) https://openai.com/it-IT/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

Creation location?	Yes	No	Yes	No	No
Linguistic competencies?	No	Yes	No	No	No
Social context?	Yes	Yes	Yes	No	No
Text references?	Yes	Partial	Yes	No	No
Methodologies?	Yes	Yes	Yes	No	No

Table 2. Competency question coverage

Table 2 presents the competency question coverage analysis across the five generated nanopublications. The three authorship-focused articles demonstrate comprehensive coverage of the CQs, with partial coverage designations reflecting structural differences in representation rather than extraction failures. Table 3 summarizes the authorship attributions extracted from the three main articles, demonstrating the pipeline's ability to capture different levels of specificity in scholarly claims. Comparing the different KGs, we can draw a picture of the different opinion on the epic: Van Daele (2005) and Peeters (1973) both proposed Willem van Boudelo as author linking to the historical individual, while Besamusca & Bouwman (2009) refuse to support an identification, leaving the issue open (see Table 3).

Article	Author	Historical	Dating	Social Context
Van Daele (2005)	Willem van Boudelo	Yes	circa 1250	Cistercian conversi
Besamusca (2009)	Willem	No	13th century	Flemish aristocracy
Peeters (1973)	Willem van Baudelo	Yes	after 1258	Cistercian cleric

Table 3. Authorship attribution patterns

The domain expert identified several points of improvement. The primary error category involved false positive manuscript references in Wackers (2016), where Comburg and Dyck (manuscript designations) were incorrectly represented as disjoint entities, rather than as physical carriers of the text, a category confusion between work-level and transmission-level entities requiring FRBR-level differentiation. Additional challenges included temporal expression complexity (phrases such as “circa 1250”), argumentative structure preservation, rather than the result of the interpretation itself.

5. CONCLUSIONS

This work examined the challenges of extracting and representing complex scholarly reasoning from long-form philological texts within a text-to-KG framework. The proposed pipeline constitutes a proof of concept demonstrating that LLM-based systems can convert extensive interpretative discourse into structured, machine-readable nanopublications in the CH domain. The central contribution is the integration of a RAG module with CQs to improve grounding, explainability, and relevant information summarization. The RAG component proved valuable both to steer the summarization from the LLM and as an explainability aid, as errors in the KG could be traced back to retrieved passages and subsequent pipeline steps, allowing also direct traceability to the original article paragraphs for source verification. The Digital Hermeneutics successfully differentiates the different layers of the scholarly interpretations about *Van den vos Reynaerde*. The CIDOC CRM implementation demonstrated flexibility across multiple dimensions. The influence relations (crm:P15_was_influenced_by) represented historical influences, other works, and cultural entities; the activity graphs (crm:E7_Activity) modeled multiple facets of the creator's life, from occupation to institutional affiliations. Several methodological limitations emerged that require attention in future work. Future work should concentrate on expanding the covered schemas (e.g. FRBR-level differentiation) and on including explicit steps to detect argumentative dependencies and reassemble premises and conclusions spanning non-contiguous passages. A human-in-the-loop web application supporting supervised resolution across pipeline stages would allow domain experts to inspect retrieved evidence, validate or correct entity and relation candidates, and iteratively refine nanopublications before publication.

ACKNOWLEDGEMENTS

This research is supported by INFINITY: a EU Horizon Europe project under Grant Agreement No 101233051. Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or EC-REA. Neither the European Union nor the granting authority can be held responsible for them.

REFERENCES

Bode, K. (2020). Why You Can't Model Away Bias. Modern Language Quarterly, 81(1), 95–124. https://doi.org/10.1215/00267929-7933102

Bouwman, A., & Besamusca, B. (Eds.). (2009). Of Reynaert the Fox: Text and Facing Translation of the Middle Dutch Beast Epic Van den vos Reynaerde. Amsterdam University Press. https://doi.org/10.5117/9789089640246

Cigliano, A., & Fallucchi, F. (2025). The Convergence of Open Data, Linked Data, Ontologies, and Large Language Models: Enabling Next-Generation Knowledge Systems. In M. Sfakakis, E. Garoufallou, M. Damigos, A. Salaba, & C. Papatheodorou (Eds.), Metadata and Semantic Research (Vol. 2331, pp. 197–213). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-81974-2_17

Daquino, M., Pasqual, V., & Tomasi, F. (2020). Knowledge Representation of digital Hermeneutics of archival and literary Sources. JLIS, (3). https://doi.org/10.4403/jlis.it-12642

Daquino, M., & Tomasi, F. (2015). Historical Context Ontology (HiCO): A Conceptual Model for Describing Context Information of Cultural Heritage Objects. In E. Garoufallou, R. J. Hartley, & P. Gaitanou (Eds.), Metadata and Semantics Research (Vol. 544, pp. 424–436). Springer International Publishing. https://doi.org/10.1007/978-3-319-24129-6_37

Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.-E., Lomeli, M., Hosseini, L., & Jégou, H. (2025). THE FAISS LIBRARY. IEEE Transactions on Big Data, 1–17. https://doi.org/10.1109/TBDATA.2025.3618474

Fan, W., Ding, Y., Ning, L., Wang, S., Li, H., Yin, D., Chua, T.-S., & Li, Q. (2024). A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models. Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 6491–6501. https://doi.org/10.1145/3637528.3671470

Giagnolini, L., Schimmenti, A., Bonora, P., & Tomasi, F. (2025). Expliciting Contexts: Semantic Knowledge Extraction from Traditional Archival Descriptions. Umanistica Digitale, 9(20), 115–144. https://doi.org/10.6092/ISSN.2532-8816/21229

Hyvönen, E., Leskinen, P., Heino, E., Tuominen, J., & Sirola, L. (2017). Reassembling and Enriching the Life Stories in Printed Biographical Registers: Norssi High School Alumni on the Semantic Web. In J. Gracia, F. Bond, J. P. McCrae, P. Buitelaar, C. Chiarcos, & S. Hellmann (Eds.), Language, Data, and Knowledge (Vol. 10318, pp. 113–119). Springer International Publishing. https://doi.org/10.1007/978-3-319-59888-8_9

Koho, M., Leskinen, P., & Hyvönen, E. (2020). Integrating Historical Person Registers as Linked Open Data in the WarSampo Knowledge Graph. In E. Blomqvist, P. Groth, V. De Boer, T. Pellegrini, M. Alam, T. Käfer, P. Kieseberg, S. Kirrane, A. Meroño-Peñuela, & H. J. Pandit (Eds.), Semantic Systems. In the Era of Knowledge Graphs (Vol. 12378, pp. 118–126). Springer International Publishing. https://doi.org/10.1007/978-3-030-59833-4_8

Lisena, P., Schwabe, D., van Erp, M., Troncy, R., Tullett, W., Leemans, I., Marx, L., & Ehrich, S. C. (2022). Capturing the Semantics of Smell: The Odeuropa Data Model for Olfactory Heritage Information. In P. Groth, M.-E. Vidal, F. Suchanek, P. Szekley, P. Kapanipathi, C. Pesquita, H. Skaf-Molli, & M. Tamper (Eds.), The Semantic Web (pp. 387–405). Springer International Publishing.

Martinez-Rodriguez, J. L., Hogan, A., & Lopez-Arevalo, I. (2020). Information extraction meets the Semantic Web: A survey. Semantic Web, 11(2), 255–335. https://doi.org/10.3233/SW-180333

Morales Tirado, A., Carvalho, J., Ratta, M., Uwasomba, C., Mulholland, P., Barlow, H., Herbert, T., & Daga, E. (2024). Musical Meetups Knowledge Graph (MMKG): A Collection of Evidence for Historical Social Network Analysis. In A. Meroño Peñuela, A. Dimou, R. Troncy, O. Hartig, M. Acosta, M. Alam, H. Paulheim, & P. Lisena (Eds.), The Semantic Web (Vol. 14665, pp. 110–127). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-60635-9_7

Peeters, L. (1973). Historiciteit en chronologie in Van den vos Reynaerde. Spektator. Tijdschrift Voor Neerlandistiek, 3, 157–175.

Schimmenti, A., Pasqual, V., Tomasi, F., Vitali, F., & Marieke Van Erp. (2024). Structuring Authenticity Assessments on Historical Documents using LLMs. https://doi.org/10.6092/unibo/amsacta/7927

Van Daele, R. (2005). De robotfoto van de Reynaertdichter. Bricoleren met de overgeleverde wrakstukken: “cisterciënzers”, “grafelijk hof” en “Reynaertmaterie.” Tiecelijn : Nieuwsbrief Voor Reynaerdofielen. - Sint-Niklaas, 1988, Currens, 18 :3 (2005), pp. 179–205. https://repository.uantwerpen.be/docstore/d:irua:20388

Wackers, P., Block, E. C., Dufournet, J., Goossens, J., Jill Mann, Pastré, J.-M., Schouwink, W., Stephenson, R., Subrenat, J., Suomela-Härmä, E., Daele, R. van, & Kenneth Varty. (2000). Medieval French and Dutch Renardian Epics: Between Literature and Society. In Reynard the Fox: Cultural Metamorphoses and Social Engagement in the Beast Epic from the Middle Ages to the Present (1st ed., pp. 55–72). Berghahn Books. http://www.jstor.org/stable/j.ctt1c0gkr9.9

Wackers, P. W. M. (2016). Wat staat er eigenlijk? Over het editeren van Van den vos Reynaerde. Tiecelijn, 29, 98–118.

Un LLM per le Scienze Umane e Sociali? ReSearch_SSH come esperimento europeo

ADAM FACI[1], ALESSIO MIASCHI[2], ANNE COMBE[3], PASCAL CUXAC[4], FRANCESCA FRONTINI[2], NICOLAS LARROUSSE[1], STÉPHANE POUYLLAU[1]

1 HUMA-NUM, CNRS, FRANCE - {NAME.SURNAME}@HUMA-NUM.FR
2 CNR-ILC, ITALY - {NAME.SURNAME}@ILC.CNR.IT
3 INRIA, FRANCE - ANNE.COMBE@INRIA.FR
4 INIST, CNRS, FRANCE - PASCAL.CUXAC@INIST.FR

ABSTRACT (ITALIANO)

L'adozione dei Large Language Models (LLM) nel contesto della ricerca scientifica, in particolare per attività di ricerca e scoperta bibliografica, solleva questioni rilevanti per le Scienze Umane e Sociali (SSH), soprattutto in relazione alla diversità delle pratiche disciplinari e alla valutazione dei risultati. Questo contributo presenta ReSearch_SSH, un caso d'uso sviluppato nel contesto del progetto europeo LLMs4EU e dell'infrastruttura ALT-EDIC, come esperimento di adattamento e valutazione di modelli linguistici orientato alle pratiche di ricerca nelle SSH e, più specificamente, nelle Digital Humanities. Il lavoro descrive l'architettura metodologica del caso d'uso, basata su strategie di *domain alignment*, *task-oriented fine-tuning* e su un approccio GraphRAG integrato con infrastrutture di scienza aperta come ISIDORE e ISTEX. Particolare attenzione è dedicata alle risorse dati multilingui, ai requisiti di fattualità e tracciabilità delle fonti e a un modello di valutazione qualitativa fondato sul coinvolgimento di un panel di esperti delle Digital Humanities. Il contributo si colloca nel dibattito delle Digital Humanities sugli LLM come componenti infrastrutturali per la ricerca, proponendo un approccio orientato alle pratiche disciplinari e alla partecipazione delle comunità scientifiche per guidare in modo critico lo sviluppo e la valutazione di sistemi di ricerca e scoperta basati su LLM in ambito SSH.
Parole chiave: Large Language Models; fine-tuning; valutazione qualitativa.

ABSTRACT (ENGLISH)

An LLM for the Social Sciences and Humanities? ReSearch_SSH as a European Experiment. The adoption of Large Language Models (LLMs) in the context of scientific research, particularly for bibliographic research and discovery, raises significant challenges for the Social Sciences and Humanities (SSH), especially with regard to the diversity of disciplinary practices and the evaluation of results. This paper presents ReSearch_SSH, a use case developed within the European project LLMs4EU and the ALT-EDIC infrastructure, as an experiment in adapting and evaluating language models grounded in SSH research practices, with a specific focus on Digital Humanities. The contribution describes the methodological architecture of the use case, based on *domain alignment*, *task-oriented fine-tuning*, and a GraphRAG approach integrated with open science infrastructures such as ISIDORE and ISTEX. Particular attention is devoted to multilingual data resources, requirements of factuality and source traceability, and a qualitative evaluation framework based on the involvement of a panel of Digital Humanities experts. The paper situates itself within the Digital Humanities debate on LLMs as infrastructural components for research, advocating a practice-oriented and community-driven approach to critically guide the development and evaluation of LLM-based research and discovery systems in the SSH domain.

Keywords: Large Language Models; fine-tuning; qualitative evaluation.

1. INTRODUZIONE

Negli ultimi anni, i Large Language Models (LLM) hanno trasformato profondamente il panorama delle tecnologie linguistiche e delle pratiche di ricerca nelle Digital Humanities (si vedano, tra i molti contributi possibili, Ciotti, 2023, e per un esempio di utilizzo degli LLM nelle Scienze Sociali e Umanistiche, Arachchige et al., 2025). Tuttavia, lo sviluppo di LLM e di piattaforme di *scientific discovery* a supporto della ricerca scientifica è oggi fortemente concentrato su lingue e contesti culturali dominanti e tende a privilegiare forme di produzione e validazione della conoscenza tipiche delle scienze sperimentali. Strumenti ampiamente diffusi per l'esplorazione della letteratura scientifica — come motori semantici,

sistemi di *question answering* o piattaforme di sintesi automatica basate su grandi corpora di articoli — si fondano prevalentemente su pubblicazioni in lingua inglese e su modelli epistemici centrati su articoli journal-based, metriche citazionali e risultati sperimentali. Questo orientamento, evidente in molte piattaforme di scoperta scientifica oggi disponibili, comporta il rischio di marginalizzare sia la diversità linguistica europea sia le specificità epistemiche e metodologiche della ricerca nelle Scienze Umane e Sociali, dove assumono rilievo altre tipologie di fonti, pratiche interpretative e forme di argomentazione. Come osservato da Fenlon (2017), infatti, la ricerca umanistica (ed in particolare le DH) ha visto di recente l'emergere di nuove forme di produzione scientifica, quali blog di ricerca, prodotti multimediali e grandi corpora digitali, che mettono in discussione le convenzioni tradizionali della pubblicazione accademica. In questo scenario si colloca il caso d'uso ReSearch_SSH che nasce dalla collaborazione tra infrastrutture SSH francesi e italiane nel contesto del progetto LLMs4EU (Large Language Models for the European Union), un progetto finanziato dalla Commissione Europea e coordinato da ALT-EDIC (Alliance for Language Technologies – European Digital Infrastructure Consortium[1]).
Alla luce di questo quadro, il contributo si articola come segue. Dopo aver inquadrato il tema dei Large Language Models a partire dalle pratiche di ricerca proprie delle Scienze Umane e Sociali, il lavoro presenta il progetto europeo LLMs4EU nel contesto della nuova infrastruttura ALT-EDIC, soffermandosi in particolare sul contributo e sul ruolo degli attori italiani. Viene quindi illustrato ReSearch_SSH come caso di studio di adattamento e fine-tuning di modelli linguistici orientato alle SSH, con attenzione alle risorse dati, all'architettura metodologica e alle strategie di valutazione adottate. Infine, il contributo discute il coinvolgimento delle comunità di ricerca — in particolare delle Digital Humanities — attraverso la costituzione di un panel di esperti, delineando possibili modalità di partecipazione della comunità AIUCD allo sviluppo e alla valutazione di LLM a supporto della ricerca.

2. PROGETTARE LLM A PARTIRE DALLE PRATICHE DELLE HUMANITIES

La ricerca nelle Scienze Umane e Sociali (SSH) è caratterizzata da pratiche interpretative, contestuali e riflessive, in cui dati, testi e metadati non costituiscono entità neutre, ma sono costruiti e reinterpretati nel tempo in relazione a quadri teorici, domande di ricerca e contesti storici e culturali. La letteratura sulle infrastrutture di ricerca per le *humanities* ha mostrato come tali pratiche richiedano ambienti digitali progettati a partire dalle esigenze delle comunità scientifiche: infrastrutture disciplinari come CLARIN ERIC, DARIAH ERIC, OPERAS o lo SSHOC Marketplace sono state concepite per sostenere l'uso condiviso di risorse digitali mantenendo un forte legame con la diversità disciplinare e linguistica (Branco et al., 2023; Dumouchel et al., 2020; König et al., 2023). Studi più recenti sulle infrastrutture di ricerca e sui servizi di archiviazione dati sottolineano inoltre come i dati umanistici e scientifici siano il risultato di pratiche comunitarie dinamiche, in cui archivi e service provider svolgono un ruolo attivo non solo di supporto, ma anche di mediazione e co-costruzione delle comunità di ricerca (Morselli et al., 2025).
Un ulteriore elemento distintivo della ricerca nelle humanities è la pluralità degli output scientifici e la centralità del multilinguismo come dimensione epistemica e metodologica della ricerca, ampiamente discussa negli studi di Digital Humanities multilingui (Balula & Leão, 2021; Viola & Spence, 2024). Accanto agli articoli, la produzione di conoscenza comprende dataset, corpora, workflows, edizioni digitali e forme di scrittura scientifica come i blog di ricerca, la cui rilevanza epistemica e la necessità di pratiche adeguate di valorizzazione e citazione sono ampiamente documentate (Barbot et al., 2024; Mayeur, 2017). In questo quadro, l'integrazione di funzionalità supportate da LLM per la ricerca bibliografica e documentale nelle SSH non può limitarsi a un uso generalista: i modelli linguistici devono essere progettati e valutati in modo da rispecchiare le pratiche di ricerca proprie delle *humanities*, favorendo approcci interdisciplinari e transdisciplinari e facilitando i contatti tra le comunità scientifiche. Questi principi trovano una prima concretizzazione nello use case ReSearch_SSH, elaborato nel contesto del progetto LLMs4EU guidato dall'infrastruttura ALT-EDIC. Nelle sezioni seguenti presenteremo questo progetto che mira a coinvolgere le comunità SSH ed in particolare DH nel fine-tuning e valutazione di modelli linguistici.

[1] https://www.alt-edic.eu/ (cons. 7/2/2026)

3. ALT-EDIC E LLMS4EU

ALT-EDIC è un European Digital Infrastructure Consortium[2], ovvero una forma giuridica europea pensata per consentire a più Stati membri di sviluppare e gestire congiuntamente infrastrutture digitali strategiche di interesse comune. In questo quadro, ALT-EDIC mira a sostenere l'eccellenza scientifica europea nel settore delle tecnologie del linguaggio e a promuovere la diversità linguistica europea, seguendo un modello di cooperazione che coinvolge istituzioni pubbliche, industria, società civile e ricerca (da cui la stretta collaborazione con CLARIN ERIC, l'infrastruttura di ricerca per le risorse linguistiche). L'Italia partecipa ad ALT-EDIC; alla guida del consorzio nazionale ci sono il CNR-ILC (istituto che ospita anche CLARIN-IT), la Fondazione Bruno Kessler (FBK) e il CINECA, che fornisce anche l'infrastruttura di calcolo.
Costituitosi nel 2024, ALT-EDIC sta consolidando le sue attività anche attraverso la partecipazione a diversi progetti europei tra loro complementari: tra questi, una menzione importante va a OpenEuroLLM, orientato allo sviluppo di modelli linguistici aperti e multilingui, che rappresentano un tassello chiave della strategia europea per un'Intelligenza Artificiale aperta, trasparente e allineata ai valori dell'Unione.
Il presente contributo origina invece dalle attività di un secondo progetto, LLMs4EU, che mira a potenziare gli aspetti più applicativi degli LLM nei vari settori strategici, con un approccio fortemente orientato ai casi d'uso. Il progetto si articola in cinque domini applicativi (turismo, servizi pubblici, telecomunicazioni, energia e scienza) e si sviluppa attorno al fine-tuning e all'adattamento di modelli aperti a bisogni specifici.

4. LO USE CASE RESEARCH_SSH

ReSearch_SSH è uno dei casi d'uso del dominio *scienza* ed è dedicato al supporto alle pratiche di ricerca nelle Scienze Umane e Sociali (SSH). Mira allo sviluppo e alla sperimentazione di modelli linguistici generativi adattati a compiti avanzati di scoperta, interrogazione e sintesi della letteratura scientifica, con un'attenzione specifica ai requisiti di fattualità, tracciabilità delle fonti ed esplicabilità dei risultati.
Il caso d'uso si fonda su una stretta integrazione tra partner infrastrutturali chiave e risorse europee, in particolare CNRS (attraverso le infrastrutture Huma-Num e INIST/ISTEX) e CNR-ILC (CLARIN-IT). L'obiettivo è sostenere flussi di lavoro realistici della ricerca SSH – quali la costruzione, l'estensione e l'aggiornamento dello stato dell'arte, nonché l'interrogazione ragionata di insiemi documentali – attraverso modelli multilingui (il caso d'uso si concentra per ora su francese e italiano, oltre che inglese) integrati in una piattaforma di ricerca ampiamente utilizzata come ISIDORE.[3]
Dati e focalizzazione disciplinare. Il fine-tuning partirà da modelli aperti, quali Salamandra (Gonzalez-Agirre et al., 2025) e il già citato OpenEuroLLM, e avverrà inizialmente sul corpus ISTEX (de Salabert & Barreaux, 2020) nella sua componente SSH (3 milioni di documenti, in diverse lingue, tra cui inglese, francese e italiano, per un totale di 2 miliardi di token circa), che rappresenta una base ampia e multidisciplinare della produzione scientifica in questo ambito. Questo corpus, costituito da testi completi in formato XML/TEI e da metadati ricchi e standardizzati, sarà utilizzato principalmente per le attività di *domain alignment*, al fine di adattare i modelli al discorso scientifico e alla terminologia propria delle SSH nel loro complesso.
Su questa base generalista, si introduce una focalizzazione progressiva sulle Digital Humanities, attraverso l'integrazione di risorse mirate provenienti sia dal contesto francese sia da quello italiano. In particolare, verranno utilizzati contenuti scientifici e blog accademici della piattaforma Hypotheses e dati e metadati di ricerca provenienti dal repository dati Nakala.[4] In seguito, potranno essere inseriti anche altri dati, tra cui atti di conferenze e pubblicazioni rappresentative delle comunità DH italiane, tra cui i *proceedings* di AIUCD, la rivista Umanistica Digitale, gli atti della conferenza CLiC e l'*Italian Journal of Computational Linguistics*. Queste risorse, rilasciate con licenze aperte, consentiranno di rafforzare l'allineamento multilingue e interculturale tra pratiche di ricerca SSH francesi, italiane e anglofone, e di ancorare il modello a comunità disciplinari ben definite.
Approccio metodologico e fine-tuning. L'approccio proposto privilegia un'architettura di Knowledge Graph–based Retrieval-Augmented Generation (GraphRAG, si veda tra gli altri Edge et al., 2025), concepita per garantire che le risposte generate siano sempre esplicitamente fondate su documenti di

[2] https://digital-strategy.ec.europa.eu/en/policies/edic (cons. 7/2/2026)
[3] https://isidore.science/ (cons. 7/2/2026)
[4] https://hypotheses.org/ (cons. 7/2/2026) ; https://www.nakala.fr/ (cons. 7/2/2026)

riferimento. In generale, i metodi di Retrieval-Augmented Generation (RAG) combinano modelli generativi con un modulo di retrieval, che seleziona dinamicamente porzioni rilevanti di conoscenza da una base documentale esterna, mitigando fenomeni di allucinazione e migliorando l'aderenza fattuale delle risposte. GraphRAG estende questo paradigma introducendo una rappresentazione strutturata della conoscenza sotto forma di grafo, che consente di modellare esplicitamente relazioni tra entità e documenti e di supportare processi di recupero e inferenza più contestualizzati e multi-hop.
Il processo di adattamento del modello si articola in più fasi. La prima fase consiste nell'allineamento di dominio (*domain alignment*), realizzato tramite pre-training su corpora SSH in francese e italiano, con l'obiettivo di migliorare la competenza linguistica e terminologica del modello senza introdurre comportamenti orientati a compiti specifici. La seconda fase è dedicata al task-*oriented fine-tuning*, in cui il modello viene adattato a compiti di ricerca attraverso *instruction tuning* integrato direttamente in una pipeline che implementa anche GraphRAG. I compiti target includono: comprensione di query di ricerca espresse in linguaggio naturale; recupero di documenti scientifici rilevanti; organizzazione dei risultati secondo criteri propri delle rassegne di letteratura; generazione di sintesi brevi e contestualizzate; risposta a domande su insiemi documentali strutturati; confronto e collegamento tra pubblicazioni sulla base di temi, autori e riferimenti. L'uso di grafi di conoscenza – si potranno usare quelli di Wikidata e OpenAIRE, oltre ai metadati di ISIDORE e Nakala – consente di arricchire il *retrieval*, migliorare il linking tra documenti e rafforzare l'esplicabilità dei risultati. Meccanismi interni di ragionamento (ad esempio rappresentazioni intermedie strutturate) possono essere impiegati a supporto dell'interpretazione delle query e della coerenza delle risposte, senza essere esposti direttamente all'utente finale.
Deployment e valutazione. Il modello così specializzato sarà messo in opera su una versione della piattaforma ISIDORE, concepita come una sorta di "ISIDORE AI", in cui sperimentare in modo controllato funzionalità avanzate di assistenza alla ricerca. Il sistema sarà integrato con il database e i grafi di conoscenza di ISIDORE e consentirà agli utenti di formulare interrogazioni complesse e ottenere risultati organizzati, sintetizzati e corredati da riferimenti espliciti alle fonti. L'interrogazione multilingue potrà essere supportata anche dall'uso di un modello traduttore.
La valutazione del sistema non si limiterà a metriche automatiche, ma includerà una valutazione "a panel" con il coinvolgimento di comunità di ricerca DH francesi e italiane. I panel di esperti saranno chiamati a valutare la qualità, l'utilità e l'affidabilità dei risultati prodotti, contribuendo anche a processi di ottimizzazione delle preferenze, orientati a migliorare l'allineamento del modello alle pratiche e ai criteri di rilevanza propri delle Digital Humanities.

5. QUESTIONI LEGALI E ADERENZA ALLE NORMATIVE EUROPEE

ReSearch_SSH è progettato in piena aderenza al quadro normativo europeo in materia di Intelligenza Artificiale, protezione dei dati e diritto d'autore. In primo luogo, le attività di adattamento e messa in opera dei modelli tengono conto dei vincoli introdotti dall'AI Act (Regulation (EU) 2024/1689, 2024), con particolare attenzione ai requisiti di trasparenza, tracciabilità delle fonti, gestione dei rischi e supporto alla valutazione umana, aspetti particolarmente rilevanti nel contesto di sistemi di supporto alla ricerca scientifica. L'adozione di un'architettura GraphRAG e l'utilizzo in una sandbox sperimentale contribuiscono a mitigare i rischi associati all'uso di modelli generativi in contesti sensibili come quello accademico.
Dal punto di vista della protezione dei dati personali, il progetto opera nel rispetto del GDPR, in particolare per quanto riguarda i dati raccolti nell'ambito delle attività di valutazione basate su panel di esperti. Tali dati, che possono includere opinioni, preferenze e valutazioni qualitative riconducibili a persone identificabili, sono trattati come dati sensibili, secondo principi di minimizzazione, limitazione delle finalità e conservazione controllata, e utilizzati esclusivamente per finalità di validazione scientifica e miglioramento del sistema.
Un'attenzione specifica è inoltre dedicata alle questioni di copyright e licensing. Tutti i dataset impiegati sono analizzati rispetto alle condizioni di utilizzo previste dalle rispettive licenze, incluse licenze non commerciali e, in alcuni casi, accordi specifici di riservatezza (*Non-Disclosure Agreements*). In linea con il dibattito recente, il progetto assume che la disponibilità di contenuti con licenza aperta (ad esempio CC-BY) non implichi automaticamente la loro riutilizzabilità per il fine-tuning di modelli linguistici. Come discusso da Spichtinger (2026), esistono infatti tensioni strutturali tra l'ethos della scienza aperta — fondato su condivisione, riuso e apertura dei dati — e l'"appetito" dei sistemi di Intelligenza Artificiale per grandi quantità di dati, che solleva interrogativi giuridici ed etici sul riuso dei contenuti scientifici al di là

delle finalità originarie per cui sono stati resi disponibili[5]. Ne consegue la necessità di una valutazione caso per caso delle condizioni di riutilizzo, incluse le eccezioni per il text and data mining previste dalla normativa europea.
In questo contesto, LLMs4EU adotta un approccio strutturato alla conformità legale ed etica: per ciascun caso d'uso sono previsti un Data Plan e un Legal and Ethics Compliance Plan, e il progetto include un intero Work Package dedicato alle questioni legali ed etiche. Questo assetto garantisce che la selezione dei dati, le strategie di fine-tuning e le modalità di deployment siano costantemente valutate alla luce delle normative europee vigenti e dei valori di responsabilità, affidabilità e sostenibilità che guidano lo sviluppo delle tecnologie linguistiche in ambito europeo.

6. VERSO UN PANEL DI ESPERTI

Un elemento centrale di ReSearch_SSH è la costituzione di un panel di esperti, composto da studiosi e professionisti delle Scienze Umane e Sociali, con un focus specifico sulle Digital Humanities. Questa scelta non è casuale: le DH rappresentano infatti un ambito delle SSH in cui l'uso di infrastrutture digitali, la riflessione metodologica e la sperimentazione di strumenti computazionali sono già parte integrante delle pratiche di ricerca. Il panel è quindi concepito non solo come strumento di valutazione qualitativa nelle fasi iniziali del progetto, ma come spazio di co-progettazione, in cui le comunità disciplinari partecipano attivamente alla definizione di scenari d'uso, requisiti metodologici e criteri di rilevanza. In questo senso future fasi di progettazione potranno utilizzare il feedback del panel per un ulteriore adattamento dei modelli.
Le attività di valutazione non riguarderanno esclusivamente gli output dei modelli, ma l'approccio del caso d'uso nel suo complesso, includendo aspetti quali la pertinenza, l'affidabilità e l'utilità dei risultati nei contesti di ricerca SSH, nonché l'individuazione di potenziali rischi epistemici, come allucinazioni, bias disciplinari o effetti di semplificazione. Il panel supporterà inoltre la sperimentazione di approcci di valutazione human-in-the-loop e di task che vadano oltre *benchmark* generalisti, privilegiando scenari complessi e situati, tipici della ricerca umanistica. Grazie all'integrazione con infrastrutture disciplinari per la scienza aperta e alla natura pubblica e federata degli strumenti sviluppati, ReSearch_SSH ambisce a creare le condizioni per un coinvolgimento duraturo delle comunità DH e SSH. In questa prospettiva, il panel non è concepito come un meccanismo limitato alla durata di LLMs4EU, ma come un modello di partecipazione e valutazione che possa essere mantenuto e fatto evolvere nel tempo all'interno delle infrastrutture di ricerca nazionali ed europee, anche oltre il ciclo di vita del progetto.

7. CONCLUSIONI

Nato nel contesto di LLMs4EU e dell'infrastruttura ALT-EDIC, lo use case ReSearch_SSH si configura come un laboratorio di sperimentazione per un uso responsabile e situato dei modelli linguistici generativi a supporto della ricerca nelle Scienze Umane e Sociali. Attraverso l'integrazione di strategie di adattamento dei modelli, architetture GraphRAG, infrastrutture di scienza aperta e processi di valutazione qualitativa basati su comunità di esperti, il caso d'uso propone un approccio che mette al centro l'impatto epistemico degli LLM sulle pratiche di scoperta, interpretazione e costruzione della conoscenza.
In una prospettiva di medio e lungo periodo, ReSearch_SSH apre a possibili collaborazioni strutturate con attori istituzionali e comunitari attivi nel panorama delle DH e delle SSH. In ambito italiano, un dialogo con il Dipartimento di Scienze Umane e Sociali, Patrimonio Culturale del CNR (CNR-DSU), e in particolare con il DSU Digital Humanities Center, potrebbe rafforzare l'integrazione tra sviluppo infrastrutturale, ricerca metodologica e comunità disciplinari. Analogamente, sul versante francese, il coinvolgimento del CNRS Humanities & Social Sciences (CNRS-INSHS) appare naturale per consolidare il legame tra strumenti di supporto alla ricerca, infrastrutture di scienza aperta e pratiche disciplinari condivise.
Un ruolo centrale è infine attribuito alle comunità scientifiche, in particolare ad AIUCD e alla corrispondente associazione francofona Humanistica, come spazi privilegiati per la definizione di scenari d'uso, criteri di valutazione e per una riflessione critica e continuativa sull'impatto dei modelli linguistici generativi nelle humanities. In questo senso, ReSearch_SSH ambisce a costituire non solo un risultato

[5] Dal punto di vista delle *Digital Humanities*, tale tensione non è un effetto collaterale, ma un nodo epistemologico e politico: la promessa di apertura, riuso e circolazione della conoscenza scientifica si confronta con pratiche di addestramento che rischiano di trasformare dati aperti in risorse opache, riaggregate e sottratte al controllo delle comunità che le hanno prodotte.

progettuale, ma un punto di partenza per un dialogo stabile tra infrastrutture, tecnologie e comunità di ricerca.

RINGRAZIAMENTI

Questo lavoro è stato sostenuto dal progetto LLMs4EU 'Large Language Models for the European Union', finanziato dall'Unione Europea tramite il Programma Europa Digitale (DIGITAL-2024-AI-B-06-LANGUAGE - GA 101198470) nell'ambito dell'accordo di sovvenzione 101198470.
Per la redazione dei Capitoli 4 e 5 di questo contributo è stata utilizzata IA generativa come supporto per la traduzione e la rielaborazione di passaggi originariamente redatti in lingua inglese.

BIBLIOGRAFIA

Arachchige, I. N., Frontini, F., Mitkov, R., & Rayson, P. (Eds). (2025). *Proceedings of the First on Natural Language Processing and Language Models for Digital Humanities*. INCOMA Ltd., Shoumen, Bulgaria. https://aclanthology.org/2025.lm4dh-1

Balula, A., & Leão, D. (2021). Multilingualism within Scholarly Communication in SSH. A literature review. *JLIS.It*, *12*(2), 88–98. https://doi.org/10.4403/jlis.it-12672

Barbot, L., Dolinar, M., Gray, E. J., Grisot, C., Illmayer, K., Kurzmeier, M., & McGillivray, B. (2024). Contextualizing Research Tools & Services Through Workflows in the SSH Open Marketplace. *Journal of Open Humanities Data*, *10*(1). https://doi.org/10.5334/johd.192

Branco, A., Eskevich, M., Frontini, F., Hajič, J., Hinrichs, E., Jong, F. de, Kamocki, P., König, A., Lindén, K., Navarretta, C., Piasecki, M., Piperidis, S., Pitkänen, O., Simov, K., Skadiņa, I., Trippel, T., Witt, A., & Zinn, C. (2023). The CLARIN infrastructure as an interoperable language technology platform for SSH and beyond. *Language Resources and Evaluation*. https://doi.org/10.1007/s10579-023-09658-z

Ciotti, F. (2023). Minerva e il pappagallo: IA generativa e modelli linguistici nel laboratorio dell'umanista digitale. *Testo e Senso*, (26), 289–315. https://doi.org/10.58015/2036-2293/671

de Salabert, C., & Barreaux, S. (2020). Vers un corpus optimal pour la fouille de textes: Stratégie de constitution de corpus spécialisés à partir d'ISTEX. In C. Benzitoun, C. Braud, L. Huber, D. Langlois, S. Ouni, S. Pogodalla, & S. Schneider (Eds), *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 4: Démonstrations et résumés d'articles internationaux* (pp. 66–69). ATALA. https://hal.science/hal-02768520

Dumouchel, S., Blotière, E., Barbot, L., Breitfuss, G., Chen, Y., Donato, F. D., Forbes, P., Petitfils, C., & Pohle, S. (2020). TRIPLE project: Building a discovery platform to enhance collaboration. *ITM Web of Conferences*, *33*, 03005. https://doi.org/10.1051/itmconf/20203303005

Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., Metropolitansky, D., Ness, R. O., & Larson, J. (2025). *From Local to Global: A Graph RAG Approach to Query-Focused Summarization* (arXiv:2404.16130). arXiv. https://doi.org/10.48550/arXiv.2404.16130

Fenlon, K. S. (2017). *Thematic research collections: Libraries and the evolution of alternative scholarly publishing in the humanities* [Text, University of Illinois at Urbana-Champaign]. https://hdl.handle.net/2142/99380

Gonzalez-Agirre, Aitor, et al. "Salamandra technical report." arXiv preprint arXiv:2502.08489 (2025).

König, A., Barbot, L., Grisot, C., Kurzmeier, M., & Gray, E. J. (2023). The SSH Open Marketplace and CLARIN. *Proceedings of the CLARIN Annual Conference*. https://doi.org/10.3384/ecp210006

Mayeur, I. (2017). Imparting Knowledge in Humanities. About Some Practices of Scientific Blogging on Hypothèses. In *Expanding Perspectives on Open Science: Communities, Cultures and Diversity in Concepts and Practices* (pp. 75–84). IOS Press. https://doi.org/10.3233/978-1-61499-769-6-75

Morselli, F., Touber, J., & Scharnhorst, A. (2025). *Fostering Data Communities—Perspective from a Data Archive Service Provider* (arXiv:2502.02321). arXiv. https://doi.org/10.48550/arXiv.2502.02321

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). (2024). *Official Journal of the European Union*, L, 2024/1689.

Spichtinger, D. (2026). *Perspective Chapter: Fit for Purpose? Creative Commons Licensing for Research Data in the Age of Artificial Intelligence*. IntechOpen. https://doi.org/10.5772/intechopen.1013402

Viola, L., & Spence, P. (Eds). (2024). *Multilingual Digital Humanities*. Routledge.

L’AI a servizio della Musicologia: costruzione di thesauri musicologici con il supporto dei Large Language Models[1]

MANUELA GRILLO[1], PAOLO BONORA[2]
[1] SAPIENZA UNIVERSITÀ DI ROMA, MANUELA.GRILLO@GMAIL.COM
[2] ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA, PAOLO.BONORA@UNIBO.IT

ABSTRACT (ITALIANO)

La necessità di utilizzare linguaggi controllati e strumenti di indicizzazione semantica per il recupero dell’informazione specialistica, sia nei database della ricerca di settore che nei cataloghi degli istituti di conservazione delle risorse di specifici ambiti disciplinari, rende necessario riflettere sulle modalità e sulla sostenibilità di costruzione di linguaggi controllati specialistici.
Il contributo si occupa in particolare dell’ambito musicologico, proponendo una riflessione su una proof of concept di costruzione di thesaurus attraverso l'uso dell'AI, alla luce dell’esperienza di redazione di un thesaurus per le arti performative nell’ambito del Progetto ERC PerformArt e del convegno internazionale Controlled Languages for Musical Heritage. State of the Art, Tools, Perspectives (Roma, 3-4 giugno 2024). Si è condotto un tentativo di categorizzazione di termini di thesaurus di ambito musicologico attraverso l’utilizzo di sistemi Large Language Model (LLMs). Si indaga la loro efficacia nel classificare i termini proposti come casi di studio in base ad una serie di macrocategorie di riferimento nel rispetto degli standard internazionali sulla costruzione di thesauri.
Parole chiave: Information retrieval; LLMs; Musicologia; Sematic web; Thesauri

ABSTRACT (ENGLISH)

AI and Musicology: building thesauri with the support of Large Language Models.
The need for controlled languages and semantic indexing tools for retrieving specialized information makes it necessary to reflect on the methods and sustainability of constructing specialized controlled languages. This is relevant both in sector-specific research databases and in the catalogs of institutions that preserve resources in specific disciplinary fields.
This contribution focuses in particular on the field of musicology, offering a reflection on attempts to categorize thesaurus terms through the use of AI, based on the experience of compiling a thesaurus for the performing arts as part of the ERC PerformArt Project and the international conference Controlled Languages for Musical Heritage. State of the Art, Tools, Perspectives (Rome, June 3-4, 2024). An attempt to categorize musicological thesaurus terms was made using Large Language Model (LLM) systems. The contribution investigates the effectiveness of LLMs in classifying the terms proposed as case studies according to a series of reference macro-categories, in compliance with international standards on thesaurus construction.
Keywords: Information retrieval; LLMs; Musicology; Sematic web; Thesauri

1. INTRODUZIONE

Il contributo ha per oggetto la costruzione, attraverso sistemi di intelligenza artificiale, di strumenti di indicizzazione semantica, ovvero di indicizzazione del contenuto, delle risorse informative per il recupero dell’informazione.
In un contesto in cui e la quantità di risorse informative disponibili cresce a ritmo esponenziale e in cui il tradizionale compito di garantire l’accesso, che le biblioteche hanno tradizionalmente in condizioni di scarsità e difficoltà di reperimento dell’informazione, viene ribaltato nella necessità di costante selezione ragionata che si contrappone all’information overload (Meschini, 2024: 160), è cruciale la possibilità di trovare - e trovare rapidamente - ciò che è rilevante per soddisfare le esigenze informative. In particolare, indicizzare la semantica delle risorse, descriverne il contenuto tramite parole chiave o

[1] Manuela Grillo ha elaborato i paragrafi 1., 2., 4.; Paolo Bonora ha elaborato il paragrafo 3.; il paragrafo 5. è stato scritto a quattro mani.

classificazioni che ne facilitino l'accesso, è un'attività che crea le condizioni indispensabili per trasformare l'informazione potenziale in conoscenza realmente utilizzabile. L'indicizzazione è infatti l'operazione complessa di valutazione del messaggio, cioè del contenuto concettuale di un testo (Guarasci & Guerrini, 2022: 13).
Nei domini disciplinari specialistici questo è tanto più necessario, sebbene gli strumenti di organizzazione della conoscenza in uso negli istituti culturali non sempre riescano a garantire un'adeguata copertura per gli specifici concetti da rappresentare nell'indicizzazione semantica.
In particolare per l'ambito musicologico – soprattutto se pensiamo a biblioteche e centri di documentazione non di settore, come le grandi biblioteche di conservazione, che custodiscono materiali eterogenei e spesso unici (partiture manoscritte o a stampa, registrazioni sonore, bozzetti di scenografie d'opera, libretti, epistolari, programmi di sala, archivi di compositori e istituzioni musicali) – l'attività di descrizione della semantica non sempre è sostenuta da knowledge organization systems (KOS) sufficientemente rappresentativi: la conseguenza è che alcune di queste tipologie di risorse, senza un'adeguata indicizzazione semantica, rischiano di essere difficilmente trovabili, nonostante il loro valore scientifico e culturale. Si tratta di un passaggio essenziale per garantire ricercabilità, accessibilità e contestualizzazione.
In questo contesto si inserisce il possibile contributo dell'intelligenza artificiale, i cui principali possibili campi di applicazione in biblioteca, segnalati già da tempo (Spark Jones, 1999; Griffey, 2019; Brzustowicz, 2023), sono assistere da un lato l'utente nella ricerca e dall'altro il bibliotecario nella catalogazione e nell'indicizzazione per soggetto (Cheti, 2024: 347). Su quest'ultimo campo di azione, scrive Gino Roncaglia che le intelligenze artificiali generative potranno lavorare alla produzione automatica di metadati in contesti come archivi, biblioteche e musei (Roncaglia, 2023: 119).
In particolar modo i Large Language Models o LLMs (Frické, 2024: 101-128) - ma anche i Large Multimodal Models o LMMs che trattano non solo oggetti testuali ma anche immagini, audio e video (Frické, 2024: 130-135) – possono aprire nuove prospettive: la Retrieval-Augmented Generation o RAG è una funzionalità interessante soprattutto per il recupero dell'informazione da archivi e biblioteche (La Gorga, Testa & Verna, 2025).
Si indaga in questo contributo l'efficacia degli LLMs nel classificare i termini musicologici o relativi al campo delle arti performative, proposti come casi di studio.

2. IL CONTESTO DI IDEAZIONE

Allestire un thesaurus è una operazione intellettuale che si basa su standard internazionali e che prevede una serie di controlli su fonti affidabili per attestare i confini definitori di ogni descrittore, quindi testare su questa attività sistemi di intelligenza artificiale generativa è stata considerata da chi scrive una prova interessante, soprattutto nella prospettiva dei grandi numeri connessi a insiemi documentari digitali sempre più ampi, difficilmente lavorabili con forze di soli agenti umani, ovvero professionisti dell'informazione.
Gli LLMs riescono a "capire" il contesto di riferimento della domanda che viene posta: si tratta di una comprensione data dalla vicinanza "semantica" tra i termini presenti nella domanda e quelli prodotti in risposta. I modelli generativi sono quindi in grado di individuare correlazioni terminologiche, individuare nessi e derivazioni basandosi sulla correlazione statistica presente nei testi impiegati per il loro addestramento. Sono quindi in grado di proporre raggruppamenti dei termini che descrivono le caratteristiche di una risorsa in base a criteri di prossimità semantica, esplicitare le relazioni semantiche tra i termini individuati e proporne una classificazione, individuare corrispondenze con vocabolari esistenti al fine di ampliare le chiavi di accesso, generare definizioni e note d'uso a partire dai vocabolari dati.
A partire dall'esperienza di indicizzazione semantica dei documenti archivistici contenuti nel repository digitale del progetto di ricerca PerformArt – progetto di ricerca ERC dedicato alle arti dello spettacolo e diretto da Anne-Madeleine Goulet del Centre d'études supérieures de la Renaissance di Tours – per il quale è stato allestito un thesaurus specializzato in campo musicologico, teatrale e coreutico (Grillo, 2023; Grillo & Berti, 2024), in collaborazione con il Nuovo Soggettario della Biblioteca nazionale centrale di Firenze, si è quindi voluta tentare una proof of concept di costruzione delle gerarchie dei termini thesaurali per il tramite di vari sistemi LLMs (Microsoft Copilot, Google Gemini e Meta LLama).
L'interesse nei confronti dei thesauri è legato alla loro possibilità di lavorare come ponti tra le risorse di istituzioni come biblioteche, archivi e musei, che hanno descrizioni formali tradizionalmente diverse, ma che possono avere la semantica come trait-d'union. Un esempio è quello offerto a partire dai termini del

Nuovo Soggettario che, per incrementare la fruibilità e l'integrazione delle risorse degli istituti culturali, sono collegati a voci corrispondenti in altri vocabolari o altri strumenti impiegati per l'indicizzazione, la descrizione o denotazione delle risorse in archivi o musei, come le Gallerie degli Uffizi o l'Archivio Ricordi. Se la semantica è di valido aiuto per l'information retrieval nelle biblioteche, lo è in egual misura nei musei e negli archivi, in special modo negli archivi digitali, laddove la varietà di fonti e la granularità di corpora documentari 'creati' dall'assemblaggio delle singole risorse rendono l'indicizzazione semantica e l'uso di un thesaurus utili per colmare la mancanza degli elementi di contesto (Valacchi, 2023). L'uso di tecnologie legate al Semantic Web ha peraltro contribuito a ridefinire modi e sistemi di valorizzazione del potere informativo espresso dalle descrizioni archivistiche, favorendo considerazioni sulla natura stessa dei dati archivistici e sul loro riuso nell'ottica di un dialogo con altre fonti di dati, al fine di estendere le capacità semantiche ed espressive del patrimonio archivistico (Tomasi & Daquino, 2015).
I thesauri possono inoltre essere strumenti di raccordo tra il punto di vista dello specialista e la prospettiva dell'utente "comune", inteso come colui che entra in contatto con l'ambito disciplinare ma non lo padroneggia semanticamente.

3. CARATTERISTICHE DELLA PROOF OF CONCEPT

Il test è stato effettuato tenendo come thesaurus di riferimento il Nuovo Soggettario e su un ridotto insieme di venti termini, scelti perché fossero rappresentate le macrocategorie e le categorie in cui sono categorizzati i termini del thesaurus BNCF (Agenti:Organismi, Agenti:Organizzazioni, Agenti:Persone, Azioni:Attività, Azioni:Discipline, Azioni:Processi, Cose:Forme, Cose:Materia, Cose:Oggetti, Cose:Spazio, Cose:Strumenti, Cose:Strutture, Tempo):

- Accordatura
- Baritoni
- Canario
- Canaria
- Ciaccona
- Contrappunto
- Cori
- Esecuzione musicale
- Flauti
- Intermezzi
- Minuetto
- Monodia
- Litanie
- Lamenti
- Musica policorale
- Quarantore
- Scenografie
- Soprani
- Strumenti a corda
- Violini

Tra di essi sono stati scelti termini potenzialmente semplici (come Violini) e con caratteristiche che avrebbero potuto rendere l'analisi da parte del sistema per nulla scontata: termini su cui testare l'individuazione dei sinonimi (come Strumenti a corda), termini per i quali l'individuazione della categoria non fosse banale (come Accordatura e Contrappunto), termini polisemici per testare il funzionamento del sistema LLM su concetti più complessi (come Litanie, Monodia, Minuetto).
La sperimentazione è stata articolata in quattro fasi, ciascuna delle quali implementata attraverso uno specifico prompt: nella prima fase abbiamo chiesto agli LLMs oggetto dello studio di assegnare i termini campione ad una tassonomia data; nel secondo passaggio abbiamo chiesto di individuare eventuali sinonimi, nel terzo i termini gerarchicamente sovraordinati (BT) e quelli gerarchicamente sottordinati (NT) e quelli correlati logicamente (RT); infine, nel quarto, di organizzare i termini in una gerarchia[2].

[2] Il prompt completo è reperibile all'indirizzo: https://doi.org/10.6084/m9.figshare.32019999.

Questo approccio strutturato cerca di mimare, in forma semplificata e schematizzata, i processi logici seguiti dagli attori umani. In questa fase iniziale della sperimentazione è stato escluso ogni meccanismo di validazione e affinamento o di rinforzo avendo come obiettivo la verifica delle potenzialità zero-shot. Il risultato è un prototipo di thesaurus derivato dai termini iniziali, integrati con i termini individuati dai modelli e categorizzati in base alla tassonomia data[3].

4. VALUTAZIONE DEI RISULTATI

In ragione dell'esiguità del campione - il numero di termini proposti è volutamente esiguo perché si è ritenuto necessario riflettere sulla metodologia di valutazione prima di ampliare il bacino di test - il processo di valutazione ha preso in considerazione esclusivamente criteri qualitativi, rimandando una analisi basata su elementi quantitativi alla corposa implementazione del campione da sottoporre al test. I criteri di valutazione sono stati elaborati sulla base dell'aderenza del risultato ai descrittori inclusi nel Thesarus Nuovo Soggettario e tradotti in una griglia valutativa, definita a priori in modo avulso dagli strumenti utilizzati, costituita da:

- attribuzione della macrocategoria;
- attribuzione della categoria;
- individuazione del termine gerarchicamente sovraordinato o termine più generico (BT Brother Term);
- individuazione dei termini gerarchicamente sottordinati o termni più specifici (NT Narrow Terms);
- individuazione dei termini correlati (RT Related Terms);
- individuazione dei termini scartati a fronte della forma preferita (UF Use for).

Per valutare quindi in termini qualitativi i risultati della prova, prendiamo come elemento di analisi il termine Flauti in Nuovo Soggettario[4]. Microsoft Copilot ha sbagliato la categorizzazione (Cose:Strumenti), ha individuato correttamente il BT Strumenti a fiato e gli NT (Flauto traverso, Flauto dolce, Flauto barocco, Flauto basso), ha associato degli RT che non sembrano appropriati (Esecuzione musicale, Intermezzi).
Llama 3.3 ha individuato correttamente la categoria Cose:Oggetti e il BT Strumenti a fiato, non altrettanto correttamente gli NT (Ottavini, Piccoli flauti: questo secondo termine non trova corrispondenze nella tassonomia strumentale) e ha stabilito come RT Violini, termine sullo stesso piano gerarchico, ma è incomprensibile perché sia stato scelto come esclusivo rispetto ai molti altri termini potenzialmente associabili.
Gemini 2.5 Pro ha restituito la corretta categorizzazione Cose:Oggetti, il corretto BT Strumenti a fiato, degli NT parzialmente corretti (Flauto traverso, Flauto dolce, Flauto basso, Ottavino: corretti i primi tre, mentre l'ottavino è un tipo particolare di flauto traverso, per cui nel ramo tesaurale dovrebbe essere NT di Flauto traverso e non di Flauto) e degli RT non corretti (Ocarine, Pifferi, Siringhe) perché termini sinonimi.
Uno dei prossimi passi potrebbe essere l'inserimento di informazioni di ritorno sulla valutazione dei risultati
e successiva nuova valutazione di come esse vengano recepite dal sistema di AI.

5. CONCLUSIONI

A prescindere dagli specifici risultati restituiti dai singoli sistemi LLMs, possiamo notare che la categorizzazione è l'operazione meglio svolta, sulla base dell'opportuno prompt con cui i sistemi sono stati addestrati, così come la costruzione della gerarchia, ovvero l'individuazione del termine gerarchicamente sovraordinato a quello di partenza. Da migliorare l'individuazione degli RT (ovvero dei termini correlati logicamente che non hanno un rapporto gerarchico col termine di partenza), che i sistemi hanno invece interpretato secondo l'accezione più ricorrente di termine correlato nella lingua naturale, ovvero come sinonimo; si ipotizza sia da affinare il prompt di istruzione a questo proposito. Allo stesso tempo, dovremo considerare la necessità di definire una metodologia di valutazione dei risultati che consenta di gestire l'estensione del perimetro del caso di studio ed al tempo stesso di confrontare i

[3] Il dataset è disponibile all'indirizzo: https://doi.org/10.6084/m9.figshare.31289740.
[4] https://thes.bncf.firenze.sbn.it/termine.php?id=1712.

risultati ottenuti non solo utilizzando modelli diversi, ma anche diverse versioni dello stesso modello. Quest'ultimo aspetto è tutt'altro che secondario data l'incessante evoluzione accompagnata da una rapidissima obsolescenza dei modelli. Fattore che rischia di inficiare sia la riproducibilità dei risultati che la relativa costanza necessaria nel caso di processi lunghi ed iterativi come quelli di costruzione di strumenti complessi come i thesauri.
In conclusione, indicizzare la semantica delle risorse musicologiche significa dare una struttura solida a un patrimonio complesso, renderlo ricercabile in modi intelligenti e diversificati e garantire che la memoria musicale possa essere scoperta, studiata e tramandata: si tratta di un investimento strategico che permette di far emergere, valorizzare e utilizzare al massimo le risorse informative disponibili.
La costruzione di opportuni strumenti per l'indicizzazione semantica che possano agevolare il dialogo tra le fonti e tra gli istituti culturali è quindi cruciale: l'ausilio degli LLMs non è attualmente pensabile in sostituzione dell'intervento umano, ma le operazioni possono essere velocizzate e rese maggiormente sostenibili dal punto di vista della forza lavoro.

RINGRAZIAMENTI

Si ringraziano per il contributo nella realizzazione della proof of concept gli LLMs Copilot Think Deeper, Gemini 2.5 Pro, Llama 3.3.

BIBLIOGRAFIA

Brzustowicz, R. (2023). From ChatGPT to CatGPT: the implications of Artificial Intelligence on library cataloging. *Information technology and libraries, 42* (3), 1-22. https://ital.corejournals.org/index.php/ital/article/view/16295/11874.

Cheti, A. (2025). L'accesso per soggetto: dal catalogo a schede al catalogo online, con uno sguardo all'Intelligenza Artificiale. *AIB Studi, 64* (3), 321–352. https://doi.org/10.2426/aibstudi-14127.

Frické, M. (2024). *Artificial intelligence and librarianship: notes for teaching* (3rd ed.). SoftOption. https://softoption.us/sites/default/files/AIinLibrariesNotesForTeaching3rdEd.pdf.

Griffey, J. (Ed.). (2019). *Artificial Intelligence and machine learning in libraries*. American Library Association. https://journals.ala.org/index.php/ltr/issue/viewIssue/709/471.

Grillo, M. (2021). La costruzione di un thesaurus per la performance. In A.-M. Goulet, É. Oriol, & J. M. Dominguez (sous la direction de), *Spectacles et performances artistiques à Rome (1644-1740). Une analyse historique à partir des archives familiales* (pp. 85-95). Publications de l'École française de Rome. https://books.openedition.org/efr/16759.

Grillo, M., & Berti, M. (2024). From Scientific Research to a Difficult Definition, and Back Again: 'Magnificenza' as a Paradigm in the Development of a Thesaurus. In A.-M. Goulet, & M. Berti (Eds.), *Noble Magnificence: Culture of the Performing Arts in Rome, 1644-1740* (pp. 229-237). Brepols. https://www.torrossa.com/it/resources/an/5934656#.

Guarasci, R., & Guerrini, M. (2022). *Cos'è l'indicizzazione*. Editrice Bibliografica.

La Gorga, A., Testa, R., Verna, L. (2025). LLMs e retrieval augmented generation (RAG) per le biblioteche. *DigitCult: scientific journal on digital cultures, 10* (1), 59-73. https://www.tabedizioni.it/web/content/415814.

Meschini, F. (2025). I turbolenti anni Venti. Riflessioni sul ruolo delle biblioteche nell'attuale ecosistema informativo e su ciò che (auspicabilmente) verrà. *AIB Studi, 65* (2), 151–162. https://doi.org/10.2426/aibstudi-14087.

Roncaglia, G. (2023). *L'architetto e l'oracolo. Forme digitali del sapere da Wikipedia a ChatGPT*. Laterza.

Sparck Jones, K. (1999). Information retrieval and artificial intelligence. *Artificial Intelligence, 114* (1-2), 257-281.

Tomasi, F., & Daquino, M. (2015). Modellare ontologicamente il dominio archivistico in una prospettiva di integrazione disciplinare. *JLIS.it, 6* (3), 13–38. https://www.medra.org/servlet/MREngine?hdl=10.4403/jlis.it-11133.

Valacchi, F. (2023). Dai metadati ai dati, dai contesti ai contenuti: aumentare la descrizione archivistica. In E. Carbé, G. Lo Piccolo, A. Valenti, & F. Stella (a cura di), *La memoria digitale: forme del testo e organizzazione della conoscenza. Atti del XII Convegno Annuale AIUCD* (pp. 214-219). Università degli Studi di Siena. https://amsacta.unibo.it/id/eprint/7721/.

Per una formazione del filologo computazionale del futuro: Domain Specific Language e Large Language Model

CHRISTIAN D'AGATA[1], ANGELO MARIO DEL GROSSO[2], FEDERICO BOSCHETTI[2]
[1]UNIVERSITÀ DI CATANIA, ITALIA
[2]CNR-ISTITUTO DI LINGUISTICA COMPUTAZIONALE "A. ZAMPOLLI", ITALIA

ABSTRACT (ITALIANO)

L'articolo presenta la progettazione e la sperimentazione di un prototipo di syllabus universitario che integra Domain Specific Language (DSL) e Large Language Model (LLM), mostrando come tale integrazione favorisca forme scientifiche di editing digitale collaborativo, coniugando riflessione critica e uso consapevole dell'intelligenza artificiale generativa. In questo quadro, gli LLM sono in grado di supportare la progettazione dei DSL, l'estensione controllata delle annotazioni, la loro analisi e la traduzione verso formati strutturati e ambienti di pubblicazione come TEIPublisher.
Parole chiave: Large Language Model; Domain Specific Language; Natural Language Processing; Filologia e critica computazionale; Annotazione dei testi

ABSTRACT (ENGLISH)

Towards the Training of the Computational Philologist of the Future: Domain Specific Languages and Large Language Models. The article presents the design and experimentation of a university-level syllabus that integrates Domain Specific Languages (DSLs) and Large Language Models (LLMs). It shows how this integration enables collaborative forms of digital editing, combining critical reflection with the responsible use of generative artificial intelligence. In this framework, LLMs support the design of DSLs, the controlled extension of annotations, and their translation into structured formats and publishing environments such as TEIPublisher.
Keywords: Large Language Models; Domain Specific Languages; Natural Language Processing; Computational Philology and Criticism; Text Annotation

1. INTRODUZIONE. UN PROTOTIPO DI SYLLABUS TRA DSL E LLM

Una delle principali sfide delle *digital humanities* contemporanee consiste nel ripensare le proprie metodologie di ricerca e di insegnamento alla luce dell'introduzione dei Large Language Models (Ciotti, 2023; Roncaglia, 2023; Manca, 2025; Pollin et al., 2025). Su questo solco abbiamo immaginato, progettato e poi sperimentato un prototipo di syllabus universitario, concepito per integrare in modo progressivo Domain Specific Languages (DSL) (Parr, 2013; Mugelli & Taddei, 2020; Boschetti & Del Grosso, 2020; Zenzaro & Boschetti & Del Grosso, 2025) e Large Language Models (LLM). Il percorso didattico combina infatti un'introduzione ai fondamenti metodologici delle Digital Humanities (con un focus sul rapporto tra tradizioni di studio e standard di annotazione) e lo studio dei principi teorici degli LLM, così da accompagnare gli studenti dalla modellizzazione concettuale (Ciula & Marras, 2023) fino all'analisi di dati e alla produzione di output strutturati (ad es. in XML-TEI), fruiti in ambienti dedicati come TEIPublisher. In questa prospettiva, l'aula universitaria si configura come un laboratorio di apprendimento collaborativo, in cui studenti e docenti partecipano attivamente alla definizione dei modelli, alla sperimentazione degli strumenti e alla valutazione critica dei risultati, mettendo in evidenza il ruolo degli LLM come mediatori tra linguaggio naturale e modelli formali.

2. TRADIZIONI DI STUDIO E STANDARD DI ANNOTAZIONE: ISTANZE LOCALI E GLOBALI PER FILOLOGIA ED ERMENEUTICA

Una delle riflessioni principali che ispira il syllabus proposto è la tensione fra (I) aderenza agli standard, per usare risorse create dalla comunità globale e produrre nuove risorse riusabili; e (II) necessità di libertà quando si risponde a una domanda di ricerca complessa, per superare i vincoli imposti da modelli generici e recuperare gli approcci validi e originali di tradizioni di studio locali. Le discipline scientifiche sono espressione di comunità che condividono pratiche, strumenti concettuali e criteri di validazione. Nelle scienze empiriche ciò favorisce la convergenza: localismi terminologici e metodologici si riducono tramite standardizzazione, replicabilità e revisione tra pari. Nelle scienze storico-filologiche, invece, la

convergenza resta parziale: accanto a un nucleo di dati controllabili (fonti, attestazioni, cronologie, relazioni testuali) persistono strati interpretativi organizzati in scuole e tradizioni spesso compatibili, ma non riducibili a un unico quadro. La pluralità riguarda sia gli oggetti di studio (filologia classica, biblica, romanza, germanica, genetica, d'autore, tra le altre) sia le pratiche consolidate in contesti istituzionali e culturali diversi. Ogni tradizione costruisce un lessico e un proprio sistema di descrizioni, distinguendo ciò che è dato da ciò che è interpretazione del dato. In questo quadro le rappresentazioni digitali non sono contenitori neutri: rendono più o meno visibili certi aspetti del lavoro filologico (Buzzetti, 2023). Interoperabilità e riusabilità restano obiettivi importanti, ma richiedono equilibrio tra modellazione globale e fedeltà alle pratiche locali. Il syllabus propone di ripensare il passaggio dai "domain-specific languages" ai "linguaggi per domini specifici"[1], radicati in sottocomunità autorevoli, capaci di preservare distinzioni operative che rischiano di essere appiattite da schemi unificanti. Il DSL diventa così uno strato intermedio: garantisce un lessico adeguato alle varie discipline, rende esplicite le scelte di prospettiva, supporta tracciabilità e validazione interna, e consente integrazioni selettive verso standard (TEI, OntoLex-Lemon) preservando la specificità locale quando serve. L'obiettivo non è moltiplicare formati, ma distinguere ciò che va armonizzato da ciò che è opportuno mantenere come pluralità controllata, tramite mappature esplicite, documentate e reversibili.

3. COMPLEMENTARITÀ FRA DSL E LLM

Integrare DSL e LLM significa far dialogare il paradigma simbolico dei linguaggi formali (Voelter 2013) con quello connessionista della generalizzazione statistica (Lenci 2025). I DSL offrono precisione e controllabilità ma richiedono progettazione puntuale; gli LLM scalano su grandi corpora ma rischiano allucinazioni e ambiguità. Un primo campo di integrazione è allora la co-progettazione di Context-Free Grammars (CFG) (Parr 2013) per definire il DSL da usare: l'LLM assiste utenti non esperti nel modellare grammatiche formali, proponendo categorie, vincoli ed esempi, o estendendo grammatiche esistenti mantenendo coerenza sintattica. La CFG diventa così un oggetto condiviso, validabile, riusabile e versionabile (Mugelli & Taddei, 2020). Un secondo campo è l'estensione dell'annotazione: fornendo all'LLM annotazioni DSL come esempio, insieme ai vincoli grammaticali, si ottiene un prompting guidato che propaga lo schema interpretativo su porzioni più ampie di testo. L'LLM aumenta la copertura mentre il DSL mantiene il controllo tramite *parsing* e validazione. Un terzo campo è la traduzione in linguaggio naturale: l'LLM può esplicitare ciò che il DSL esprime in forma sintetica, trasformando etichette e strutture in spiegazioni calibrate per diversi destinatari, migliorando l'accessibilità. Infine, l'LLM può supportare l'interoperabilità progettando convertitori (ad esempio fogli XSLT) verso standard come XML-TEI o OntoLex-Lemon. Si ottiene così una filiera che bilancia precisione ed estensione.

4. EDITING DIGITALE CON GLI LLM: TEORIE E PRATICHE

Il secondo elemento cardine del syllabus consiste nel fornire sia le basi teoriche sia quelle pratiche relative agli attuali sistemi di Intelligenza Artificiale Generativa, in larga misura basati su modelli linguistici di grandi dimensioni (LLM) (Lenci et al., 2025). L'ecosistema dell'IA generativa (talvolta indicata con l'acronimo AGI, Artificial General Intelligence) si fonda su ricerche di linguistica computazionale volte a sviluppare modelli linguistici capaci di predire la parola successiva dato un contesto autoregressivo/causale (Jurafsky & Martin, 2025). L'architettura Transformer (Vaswani et al., 2017) è stata la svolta degli attuali modelli LLM. Si tratta di reti neurali profonde che producono rappresentazioni contestuali (embeddings) capaci di catturare dipendenze a lungo raggio, codificando

[1] Nella creazione di "domain-specific languages", o linguaggi specifici di dominio, si assume che il dominio sia già consolidato, con buone pratiche e formalismi propri che un DSL rende computazionalmente trattabili. Nella creazione di "linguaggi per domini specifici", invece, il dominio stesso (o, meglio: il sottodominio di un dominio più generale) è in via di definizione e i suoi formalismi sono co-progettati insieme al linguaggio: il formalismo dunque non precede il linguaggio, ma il linguaggio si co-definisce insieme al dominio specifico. Si pensi ad esempio alla filologia dei documenti nativamente digitali, sottodominio della filologia genetica. Mentre la filologia genetica gode già di formalismi su cui costruire DSLs, la filologia del born-digital può usare in tutto o in parte i costrutti di un nuovo linguaggio formale per aiutare a esplorarne i formalismi specifici. Un caso emblematico riguarda i documenti nativamente digitali e distribuiti, dove l'unità citabile non è data a priori, ma deve essere costruita formalmente. A differenza dei modelli codicologico-bibliografici o archivistici, in cui manoscritti o stadi redazionali come gli scartafacci o le bozze costituiscono entità identificabili, nei sistemi distribuiti la citazione richiede la definizione esplicita di versioni, porzioni testuali, operazioni e relazioni tra stati. In questo senso, i costrutti del linguaggio non si limitano a descrivere il dominio, ma contribuiscono a determinare che cosa possa essere considerato un oggetto filologico.

relazioni semantiche e sintattiche. Su questa base si collocano modelli con miliardi di parametri/pesi come GPT-3 2020, Gemini 2023, LLaMA 2023, Claude Sonnet 2024. La tecnologia degli LLM dà luogo alle cosiddette "abilità emergenti" (Wei et al., 2022), ossia capacità di affrontare efficacemente compiti specifici senza pre-addestramento esplicito. L'interazione tra studioso e modelli linguistici avviene soprattutto mediante in-context learning e prompting (Brown et al., 2020), dal zero-shot al few-shot fino a template strutturati con convenzioni dettate da linguaggi domain-specific (cfr. §2). Con prompt si intende l'input testuale fornito a un LLM per guidare risposta e/o compito; la sua scrittura segue principi e best practices ben studiate: chiarezza e specificità, background rilevante, pattern tramite esempi, raffinamento iterativo (approcci/istruzioni diverse), strutturazione grafica (parentesi, liste, simboli), ambiguità ridotta al minimo (Lenci et al., 2025). Domande di ricerca condivise dalla comunità attivano un processo di transfer/in-context learning che permette al sistema di replicare pattern su nuovi testi; ciò si presta alla didattica collaborativa, in cui docenti e studenti iterano sugli esempi e verificano, discutendo criticamente gli output sulla base dei prompt inseriti (Jurafsky & Martin, 2025). In più, gli iperparametri di decoding (temperature, sampling top-k e top-p) sono leve di creatività importanti (Holtzman et al., 2020). Per esempio, regolando temperature e sampling si bilanciano conservatività e creatività nelle annotazioni, favorendo la replica di pattern noti o l'esplorazione di interpretazioni alternative, e diventano uno strumento di riflessione sul rapporto tra norma interpretativa e innovazione nel testo generato. Consideriamo per esempio un template standard da utilizzare per strutturare i prompt:

[RUOLO], [TASK],
[CONTESTO], [ESEMPI], [TONO/STILE], [FORMATO], [VINCOLI]:
[RUOLO]: Sei un umanista digitale che si occupa di letteratura italiana contemporanea.
[TASK]: Codifica delle entità nominate in XML-TEI
[ESEMPI]:<persName ref="#abate_Vallet">abate Vallet</persName>
<placeName ref="#Salisburgo">Salisburgo</placeName>
[CONTESTO]: Ti fornisco il testo qui di seguito: <testo>
[FORMATO]: Output in XML-TEI
[VINCOLI]: Non aggiungere, cancellare, né modificare alcuna parola per nessun motivo

Questo template, il cui modello è stato utilizzato in occasione dell'Autumn School Etnatext per la codifica delle entità nominate secondo il modello TEI (Fig. 1a), è facilmente adattabile e consente, attraverso semplici modifiche a [TASK] e [FORMATO], di ottenere ad esempio la lemmatizzazione del testo in formato CoNLL-U (cfr, Sprugnoli & Redaelli, 2025) (Fig. 1b) o l'annotazione delle marche d'uso, realizzata nel nostro esempio con ChatGPT, capace di accedere alle risorse lessicografiche online come il GRADIT (De Mauro, 2000; Fig. 1c) se opportunamente guidato in fase di prompting.

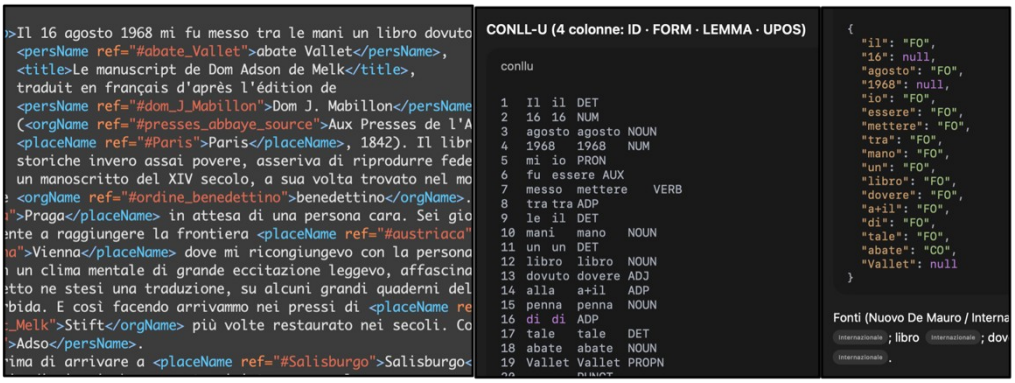


Figura 1. a) Codifica del testo (ChatGpt) b) Lemmatizzazione (ChatGpt) c) Marche d'uso (ChatGpt)

L'uso scientificamente guidato degli LLM, tramite strategie di prompting engineering, apre inoltre scenari operativi per compiti filologici complessi mediante pipeline ibride che combinano le capacità generative dei modelli con le peculiarità degli algoritmi specializzati, l'efficienza delle librerie software e l'utilità degli ambienti di pubblicazione standard. Per aumentare controllo, riproducibilità, riuso, replicabilità, efficacia ed efficienza delle indagini scientifiche, diviene prezioso interagire con gli LLM non soltanto per ottenere il risultato finale, ma soprattutto per generare procedure automatiche ready-to-use, da eseguire, in autonomia, per ottenere le analisi e le visualizzazioni sui dati oggetto di studio. Per esempio, con il

seguente prompt (di cui si mostrano solo le parti rilevanti) è possibile ottenere una procedura automatica per il riconoscimento di allitterazioni (Fig. 2):

[TASK]:Scrivi usando il linguaggio python un programma che estragga allitterazioni
[ESEMPI]:"E caddi come corpo morto cadde", "Il pietoso pastor pianse al suo pianto"
[OUTPUT]: <seg type="retorico" ana="#allitterazione" n="ripetizione della p">Il pietoso pastor pianse al suo pianto</seg>



```
segs: List[etree._Element] = []
g2p = PhonemizerEspeakBackend(lang="it")
report_format = "ipa"
txt = [
        "Nel mezzo del cammin di nostra vita",
        "mi ritrovai per una selva oscura,",
        "ché la diritta via era smarrita.",
        "Ahi quanto a dir qual era è cosa dura",
        "esta selva selvaggia e aspra e forte",
        "che nel pensier rinova la paura!"
    ]
for t in txt:
        if not isinstance(t, str) and not t.strip():
            continue
        seg = analyze_line(
            line=t,
            g2p=g2p,
            onset_mode="first_phone",
            min_freq=2,
            top_k=2,
            include_internal=True,
            internal_min_words=2,
            report_format=report_format,
        )
        if seg is not None:
            segs.append(seg)
print(build_tei(segs, title="Esperimento: Allitterazioni (IPA)"))
```

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title>Esperimento: Allitterazioni (IPA)</title>
      </titleStmt>
      <publicationStmt>
        <p>Output sperimentale generato automaticamente.</p>
      </publicationStmt>
      <sourceDesc>
        <p>Input testuale fornito dall'utente (un verso per riga).</p>
      </sourceDesc>
    </fileDesc>
  </teiHeader>
  <text>
    <body>
      <ab>
        <seg type="retorico" ana="#allitterazione" n="ripetizione dei suoni /d/ e /n/">Nel
        mezzo del cammin di nostra vita</seg>
        <seg type="source">mi ritrovai per una selva oscura,</seg>
        <seg type="source">ché la diritta via era smarrita.</seg>
        <seg type="retorico" ana="#allitterazione" n="ripetizione dei suoni /k/ e /d/">Ahi
        quanto a dir qual era è cosa dura</seg>
        <seg type="retorico" ana="#allitterazione" n="ripetizione del suono iniziale /s/
        ">esta selva selvaggia e aspra e forte</seg>
        <seg type="retorico" ana="#allitterazione" n="ripetizione dei suoni /p/ e /n/">che
        nel pensier rinova la paura!</seg>
      </ab>
    </body>
  </text>
</TEI>
```

Figura 2. Algoritmo e output XML generati per il riconoscimento delle allitterazioni

Un ulteriore caso d'uso interessante è la produzione di documenti XML-TEI, finalizzata alla pubblicazione di edizioni digitali in ambiente TEIPublisher (Boot, 2024) a partire da annotazioni DSL. I modelli linguistici assistono il filologo nella gestione del template HTML, del documento ODD (processing model) e del sorgente XML-TEI. Nell'esperimento, co-modellizzazione DSL+LLM delineano il profilo del filologo del futuro, capace di usare criticamente l'IA e di approfondire il rapporto tra intuizione ermeneutica e potenza rappresentativa ed elaborativa delle macchine.

5. CI SONO LLM IN QUESTA CLASSE?

Il syllabus è stato elaborato in occasione di Etnatext (29-31 ottobre 2025, https://etnatext.com/), autumn school organizzata dal Centro di Informatica Umanistica (CINUM) e dalla Scuola Superiore dell'Università di Catania, con il coinvolgimento di circa quaranta partecipanti tra dottorandi, studenti dei corsi di laurea magistrale e docenti di discipline letterarie. Il medesimo syllabus è stato successivamente riproposto, con i necessari adattamenti, all'interno di un seminario didattico di 18 ore presso l'Università di Catania[2]. Il presente paragrafo si concentra proprio su quest'ultima esperienza in quanto applicazione concreta del syllabus che ne conferma la sua riproducibilità in un contesto diverso. Il seminario è stato articolato in due moduli: il primo dedicato alla codifica del testo e ai principi metodologici delle Digital Humanities, il secondo incentrato sui Domain Specific Languages e sui Large Language Models. I 15 partecipanti provenivano dai corsi di Filologia moderna (7), Lettere moderne (4) e Scienze e lingue per la comunicazione (4). Il corpus di annotazione comprendeva testi della letteratura italiana contemporanea e un brano inedito, selezionato per mettere alla prova gli LLM di fronte a contenuti non presenti nei dati di addestramento né recuperabili tramite ricerche online: *La storia di fra Dolcino* da *Il nome della rosa* (Eco, 1980; 5 studenti), *Perché è sempre diverso* da *Il postino fedele* (Copioli, 2008; 3 studenti), *La presa di Roma* da *I Viceré* (De Roberto, 1894; 2 studenti) e l'inedito *Sole fulmine* (5). I modelli utilizzati sono stati ChatGPT (GPT-5), Gemini (modello Ragionamento con 3 PRO), Meta-AI (LlaMa 4), DeepSeek (v3.2) e Claude (Sonnet 4.5). La prima attività proposta, come introduzione alle strategie di prompting - tra i possibili contenuti del syllabus quali NER, lemmatizzazione, marche d'uso, analisi stilistico-retorica (cfr. §3) - ha riguardato il riconoscimento delle entità nominate, annotate secondo il modello di codifica XML-TEI. I risultati sono stati complessivamente positivi, anche nel caso di prompt zero-shot, con i modelli che in quest'ultimo caso hanno frequentemente restituito strutture di annotazione semplificate, caratterizzate

[2] Il seminario, intitolato «Leggere, annotare, interpretare i testi nel mondo digitale. Per una filologia e un'ermeneutica computazione», si è tenuto dal 3 al 16 dicembre 2025 con sei incontri di 3 ore ciascuno.

dall'uso di un numero limitato di tag e, spesso, dall'assenza di attributi. L'impiego di strategie one-shot e few-shot ha invece prodotto risultati più coerenti con il modello di codifica atteso. L'attività successiva si è invece concentrata sui DSL: dopo aver esplicitato la grammatica (CFG), preparata per l'occasione dal docente e ispirata dalla riflessione sui "linguaggi per domini specifici" (cfr. §2), è stato chiesto agli studenti di annotare liberamente il testo. Le annotazioni prevedevano l'impiego di parole chiave, introdotte dal simbolo del cancelletto, volte a caratterizzare l'interpretazione di una parola o di una frase secondo una prospettiva di ispirazione ermeneutica (Gadamer 1960) applicata alla letteratura italiana contemporanea. Ciascuna parola chiave poteva essere ulteriormente esplicitata attraverso un identificativo univoco, un titolo sintetico e una spiegazione in linguaggio naturale (corredata da riferimenti bibliografici). Un esempio di DSL proposto, per la poesia di Copioli, è raffigurato in Fig.3a. Attraverso il prompting gli annotatori hanno potuto poi estendere le proprie annotazioni al resto del testo. Ad esempio, la notazione precedente è stata ampliata con Claude (Fig. 3b) ed è stata poi tradotta dal medesimo LLM in linguaggio naturale (Fig. 3c).

A) ===INTERPRETAZIONE=== il fiume celeste: #spirituale --- id:spirituale Spirituale Il fiume celeste rispecchia l'io interiore. La parte spirituale e più profonda dell'esistenza, che ognuno di noi protegge e nasconde. Tutto ciò che arricchisce l'anima, creando un ponte tra corpo (fiume terrestre), mente (fiume iperuranio) e spirito. Il flusso di pensieri che abbiamo con noi stessi, che ci autoregola e alimenta la nostra coscienza. [bibl 2025] **B)** il fiume terrestre: #corpo il fiume celeste: #spirito il fiume infero: #inconscio il fiume iperuranio: #mente	**C)** «Il poeta costruisce una cosmologia quaternaria dell'essere umano attraverso quattro fiumi che corrispondono ad altrettante dimensioni esistenziali. Il fiume terrestre rappresenta il corpo, la dimensione fisica e materiale, l'ancoraggio sensibile dell'esperienza. Il fiume celeste simboleggia l'io interiore più profondo e spirituale, quella parte dell'anima che ognuno di noi protegge e nasconde, creando un ponte tra le diverse dimensioni dell'esistenza. Il fiume infero evoca le profondità psichiche e l'inconscio, ciò che giace sotto la soglia della coscienza razionale, completando la verticalità dell'essere. Infine, il fiume iperuranio rimanda al topos platonico delle idee, rappresentando la mente, il pensiero razionale e astratto che ordina e comprende l'esperienza».

Figura 3. a-b) Prompt DSL e risposta (Claude) c) Traduzione in linguaggio naturale del DSL (Claude)

Il risultato appare coerente con l'annotazione DSL iniziale dello studente e mostra come i modelli linguistici, ancora spesso imprecisi di fronte a richieste di annotazione generiche, siano potenzialmente efficaci nell'individuare fenomeni già modellizzati dall'annotatore. Le annotazioni co-sviluppate da studenti e LLM si sono concentrate, in particolare, sull'analisi stilistica e retorica, tematica e storica, fino all'individuazione di elementi di natura più squisitamente interpretativa. Per quanto riguarda invece la generazione autonoma di DSL da parte dei modelli a partire da specifici prompt, si nota una certa preferenza per linguaggi specifici di dominio piuttosto complessi. La maggior parte degli esperimenti ha mostrato infatti la costruzione di classificatori simbolici, strutturali e narratologici (ad esempio CONFLITTO_IDEOLOGICO {ASSE_1 = "Laicismo vs Clero" ASSE_2 = "Nazione vs Chiesa temporale"} INTERPRETAZIONE {TESI = "Il Risorgimento come spettacolo confuso e opportunistico" \ CRITICA = "Smontaggio del mito patriottico" ESITO = "Sostituzione di un potere con un altro, senza rinnovamento morale"}), prediligendo una ermeneutica profonda, ma i cui risultati devono essere validati e co-costruiti dall'annotatore. Infine gli LLM risultano ampiamente capaci anche di *parsare* le annotazioni e produrre output in formato XML (che possono poi essere resi TEI-compliant attraverso fogli di stile XSLT):

```
annotation: (head notaAnno)+;
head: TEQ type TEQ;
notaAnno: NEWLINE (keywordAnno NEWLINE)+ NOTESEP NEWLINE? standOff;
keywordAnno: keywordItem+;
keywordItem: ref+ SEP? ref? COLON keyword+;
standOff: list;
list: item+;
item: id NEWLINE label NEWLINE note NEWLINE?;
label: par+;
note: text+ bibl?;
text: par+ PUNCT? COMMA? DOT? par?;
bibl: L_BRAC par+ NUM? R_BRAC;
par: WORD+;
type: WORD+;
id: 'id:'WORD;
ref: WORD+;
keyword: HASHTAG;
```

```
<notaAnno>
    <keywordAnno>
        <keywordItem>
            <ref>Superbia</ref> : <keyword>#Superbia</keyword>
            <ref>Dolcino</ref> : <keyword>#Dolcino</keyword>
            <ref>Dante</ref> : <keyword>#Dante</keyword>
        </keywordItem>
    </keywordAnno>
    <standOff>
        <item>
            <id>id:SuperbiaDolcino</id>
            <label>La superbia di Dolcino</label>
            <note>La superbia di Dolcino si manifesta nella sua pretesa di essere
                l'unico vero apostolo di Dio e nella sua nomina a capo supremo del
                congregazione apostolica. Questo atteggiamento è simile a quello
                superbi descritti da Dante nella Divina Commedia, che sono puniti
                nella prima cornice del Purgatorio per la loro presunzione e sacc
        </item>
    </standOff>
</notaAnno>
```

Figura 4. a) Context-Free Grammar utilizzata b) Output XML di una annotazione DSL (LlaMA)

Infine, l'ultima attività prevista ha proposto una riflessione metacognitiva sul lavoro svolto, dalla quale è emerso come l'uso dei DSL sia stato valutato positivamente dai partecipanti (cfr. Fig. 5a) e preferito alla codifica in XML-TEI. I risultati indicano inoltre una preferenza generale per il prompting few-shot e con template (cfr. Fig. 5b), apprezzati per la loro efficacia a fronte di un livello di formalizzazione contenuto.

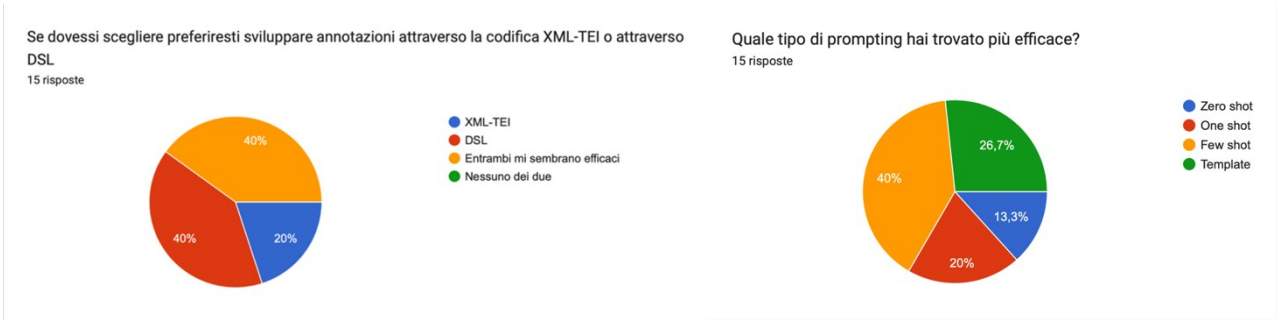


Figura 5. a) Risposte sondaggio su annotazione DSL e XML-TEI b) Risposte sondaggio sul prompting

6. CONCLUSIONE

Alla luce dei risultati incoraggianti, l'esperienza può evolvere in un syllabus per un curriculum strutturato di informatica umanistica, affiancabile a un'offerta più tradizionale (in corsi di Lettere o Filologia) ma centrata su formalismi, dati e algoritmi. L'esperimento mostra che classi inizialmente scettiche, per la rigidità dei DSL e per le allucinazioni degli LLM, hanno affrontato attività teoriche e pratiche con profitto, verificando l'efficacia dell'approccio integrato DSL+LLM. Gli LLM riducono la soglia d'accesso, mentre i DSL esplicitano e vincolano gli schemi interpretativi, rendendo trasparenti i presupposti dell'annotazione. Il passo successivo sarà: 1) sperimentare framework più avanzati di collaborazione uomo-macchina – come ad esempio il *machine-assisted quantitizing design* proposto da Karjus (2025) che permetterebbe di approfondire le possibilità di scalabilità delle annotazioni DSL su corpora più estesi; 2) integrare nel syllabus tecniche di RAG (Lewis et al., 2021) o fine-tuning; 3) adottare framework rigorosi di valutazione degli LLM[3] (Ye et al., 2025; Yang et al. 2026), così da consolidare una filologia computazionale critica, cooperativa e orientata al futuro.

BIBLIOGRAFIA

Boot, P. (2024). The TEI Processing Model: Introduction, limitations and potential extensions. In P. Lamers (Ed), *Declarative Amsterdam 2024*. https://doi.org/10.1075/da.2024.boot.tei-processing-model

[3] Occorrono metriche capaci di intercettare la coerenza argomentativa e la densità semantica proprie della prassi esegetica. Si rende pertanto necessaria l'integrazione di benchmark task-specific all'interno di workbench sperimentali modulari, dove l'impiego di tecniche di LLM-as-judge e framework di valutazione qualitativa formalizzati e rigorosi (rubriche) operi in sinergia con la validazione umana su gold standard esperti. Tale approccio permette di isolare analiticamente la qualità interpretativa della macchina, garantendo che la modellazione formale delle strutture ermeneutiche resti ancorata ai criteri di precisione filologica e sensibilità critica richiesti dalla disciplina.

Boschetti, F. & Del Grosso, A. M. (2020). L'annotazione di testi storico-letterari al tempo dei social media. *Italica Wratislaviensia* 11 (1), 65–99.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). *Language Models are Few-Shot Learners* (arXiv:2005.14165). arkiv. https://arxiv.org/abs/2005.14165

Buzzetti, D. (2023). Towards an operational approach to computational text analysis. In J. Nyhan, G. Rockwell, S. Sinclair, A. Ortolja-Baird (Eds), *On Making in the Digital Humanities: The scholarship of digital humanities development in honour of John Bradley*. UCL, 179-212. Press. https://library.oapen.org/handle/20.500.12657/62237

Ciotti, F. (2023). Minerva e il pappagallo: IA generativa e modelli linguistici nel laboratorio dell'umanista digitale. *Testo e Senso*, (26), 289–315. https://doi.org/10.58015/2036-2293/671

Ciula, A., & Marras, C. (2023). Modelli, metamodelli e modellizzazione nelle Digital Humanities. In F. Ciotti (Ed.), *Digital Humanities: Metodi, strumenti, saperi*. Carocci Editore, 51-65.

Copioli, R. (2008). *Il postino fedele*. Mondadori.

De Mauro, T. (2000). *Grande dizionario italiano dell'uso*, (Voll. 1-6). UTET (disponibile online: https://dizionario.internazionale.it/).

De Roberto, F. (1894). *I Viceré*. Galli.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805; Version 2). arXiv. https://arxiv.org/abs/1810.04805

Eco, U. (1980). Il nome della rosa. Bompiani.

Gadamer, H.-G. (1960). *Wahrheit und Methode: Grundzüge einer philosophischen Hermeneutik*. J.C.B. Mohr (Paul Siebeck).

Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2020). The Curious Case of Neural Text Degeneration. *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. https://openreview.net/forum?id=rygGQyrFvH

Jurafsky, D., & Martin, J. H. (2025). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (3rd Edition Draft)*. Stanford University (web draft). https://web.stanford.edu/~jurafsky/slp3/

Karjus, A. (2025). Machine-assisted quantitizing designs: augmenting humanities and social sciences with artificial intelligence. *Humanities and Social Sciences Communications* 12, articolo 277. https://doi.org/10.1057/s41599-025-04503-w

Lenci, A., Auriemma, S., & Miliani, M. (2025). *Linguistica computazionale: Natural Language Processing e Intelligenza Artificiale*. Hoepli.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2021). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. arXiv. https://arxiv.org/abs/2005.11401

Manca, M. (2025). Conloquia absentium. Le intelligenze generative tra prassi didattica e filologia del possibile. *ClassicoContemporaneo*, 11, 331-357.

Mugelli, G., Re, G., & Taddei, A. (2020). Annotazione digitale di testi antichi. Lingue antiche e Digital Humanities, tra ricerca e didattica. *Umanistica Digitale*, *5*(9), 35-60. https://doi.org/10.6092/issn.2532-8816/9962

Parr, T. (2013). *The Definitive ANTLR 4 Reference* (2nd ed.). Pragmatic Bookshelf.

Pollin, C., Fischer, F., Sahle, P., Scholger, M., & Vogeler, G. (2025). When it was 2024 – Generative AI in the Field of Digital Scholarly Editions. *Zeitschrift für digitale Geisteswissenschaften*, *10*. https://doi.org/10.17175/2025_008

Roncaglia, G. (2023). *L'architetto e l'oracolo. Forme digitali del sapere da Wikipedia a ChatGPT*. Laterza.

Sprugnoli, R., Redaelli, A. (2025). Annotating Manzoni: Challenges in the Annotation of Lemmas, POS and Features in "I Promessi Sposi", *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, 1084-1092.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17*, 6000–6010.

Voelter, M., Benz, S., Dietrich, C., Engelmann, B., Helander, M., Kats, L. C. L., Visser, E., & Wachsmuth, G. (2013). *DSL Engineering—Designing, Implementing and Using Domain-Specific Languages.* dslbook.org.

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). *A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT*. https://doi.org/10.48550/arXiv.2302.11382

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research*, *2022*.

Yang, X., Han, J., Bommasani, R., et al. (2026). *Reliable and Responsible Foundation Models: A Comprehensive Survey*. arXiv. https://arxiv.org/abs/2602.08145.

Ye, Y., Zhang, Z., Ma, T., Wang, Z., Li, Y. & Hou, S. (2025). *LLMs4All: A Review of Large Language Models Across Academic Disciplines* (arXiv:2509.19580). arXiv preprint. https://arxiv.org/abs/2509.19580.

Zenzaro, S., Boschetti, F., & Del Grosso, A. M. (2025). Making digital scholarly editions based on Domain Specific Languages. In J. O'Sullivan, M. Pidd, S. Whittle, B. Wessels, M. Kurzmeier, & Ó. Murphy (Eds), *Digital editing and publishing in the twenty-first century* (1st edn). Scottish Universities Press. https://doi.org/10.62637/sup.GHST9020.9

[2/B] Open Science, sostenibilità e pratiche aperte

Are Digital Humanities really committed to open? An exploratory study on the availability of methodological workflows and open peer review practices

SILVIO PERONI

DIGITAL HUMANITIES ADVANCED RESEARCH CENTRE (/DH.ARC), DEPARTMENT OF CLASSICAL PHILOLOGY AND ITALIAN STUDIES, UNIVERSITY OF BOLOGNA, BOLOGNA, ITALY - SILVIO.PERONI@UNIBO.IT - HTTPS://ORCID.ORG/0000-0003-0530-4305

ABSTRACT (ENGLISH)

Open Science has become a central framework for promoting transparency, accessibility, and inclusiveness in scholarly research. While the Digital Humanities (DH) community has long embraced openness in terms of research outputs, less attention seems to have been paid to the openness of the methodological and evaluative processes underlying knowledge production. This paper presents an exploratory study that investigates the current state of openness in DH research practices, focusing specifically on research data management documentation and peer review processes. In particular, this study addresses two research questions: (1) to what extent DH publications that describe data explicitly reference external documentation detailing data creation and management processes; and (2) how widely open peer review practices are adopted across DH conferences and journals. The results revealed a limited adoption of open methodological practices. Only a small fraction of the analysed articles provided explicit, reusable documentation of data creation workflows, and no references to data management plans or formal research data management documentation were found. An even more critical picture emerges from the analysis of peer review practices: the vast majority of DH venues continue to rely on traditional single- or double-blind review models, with open peer review adopted in only a few isolated cases.

Keywords: open science; data management plan; open peer review; methodologies

ABSTRACT (ITALIANO)

Le Digital Humanities sono davvero orientate all'open? Uno studio esplorativo sulla disponibilità dei workflow metodologici e sulle pratiche di peer review aperta. La Scienza Aperta è diventata un quadro di riferimento centrale per promuovere la trasparenza, l'accessibilità e l'inclusività nella ricerca scientifica. Sebbene la comunità delle Digital Humanities (DH) abbia da tempo adottato pratiche di apertura per quanto riguarda i prodotti della ricerca, sembra che sia stata prestata minore attenzione all'apertura dei processi metodologici e valutativi che sottendono la produzione della conoscenza. Questo contributo presenta uno studio esplorativo che indaga lo stato attuale dell'apertura nelle pratiche di ricerca nelle DH, concentrandosi in particolare sulla documentazione relativa alla gestione dei dati della ricerca e sui processi di peer review. In particolare, lo studio affronta due domande di ricerca: (1) in che misura le pubblicazioni DH che descrivono dei dati fanno esplicito riferimento a documentazione esterna che dettagli i processi di creazione e gestione dei dati; e (2) quanto siano diffuse le pratiche di open peer review tra conferenze e riviste DH. I risultati mostrano un'adozione limitata di pratiche metodologiche aperte. Solo una piccola parte degli articoli analizzati fornisce una documentazione esplicita e riutilizzabile dei workflow di creazione dei dati, mentre non è stato riscontrato alcun riferimento a piani di gestione dei dati o a documentazione formale di research data management. Un quadro ancora più critico emerge dall'analisi delle pratiche di peer review: la grande maggioranza delle sedi DH continua ad affidarsi a modelli tradizionali di revisione single-blind o double-blind, mentre le pratiche di open peer review risultano adottate solo in pochi casi isolati.

Parole chiave: open science; piano per la gestione dei dati; revisione tra pari aperta; metodologie

1. INTRODUCTION

A few years ago, UNESCO (2021) published its guidelines about Open Science, highlighting all the relevant dimensions that interests this "construct" and that involves all scientific disciplines and aspects of scholarly practices in (a) making *scientific knowledge* openly available, accessible and reusable for everyone, and (b) opening the *processes* of scientific knowledge creation, evaluation and communication

to societal actors beyond the traditional scientific community. The Digital Humanities (DH) domain has worked for years to open up the scientific knowledge produced by its research, approaching open access as a primary practice across its scholarly venues, from conferences to journals. In addition, more recently, several other initiatives have been launched to promote the publication of non-traditional research objects, such as datasets, software, corpora, and models. For instance, the *Journal of Open Humanities Data* (JOHD, https://openhumanitiesdata.metajnl.com/) has been proposed for publishing peer-reviewed articles describing humanities research objects or techniques with high potential for reuse across all subjects in the Humanities domain. Indeed, the publication of these non-traditional research objects has been demonstrated to have "a positive impact on both the metrics of research papers associated with them and on data reuse" (McGillivray et al., 2022).
From these premises, we can undoubtedly claim that the DH community has done (and continues to do) a lot for supporting the open sharing of scientific knowledge. However, a more blurred picture emerges when we consider the other side of the Open Science endeavour, namely, *how* (rather than *what*) a given research is conducted or evaluated. This aspect relates more to the *methodological process* that governs DH research rather than the outcomes it produces. Indeed, from the one hand, "we need to put more explicit attention on methodologies, and on the need for carefully documenting each step of a research workflow" (Barzaghi et al., 2025), including documenting research data management practices (Tóth-Czifra et al., 2023) via known documentation such as Data Management Plans (DMPs), since data "are of limited use if they are not sufficiently documented, and if they are not accompanied by a clear description of the research methodology" (Barzaghi et al., 2025). On the other hand, we also need to actively work to make research evaluation processes, such as *peer review*, more inclusive and, as a consequence, more transparent (Ross-Hellauer, 2017).
In this paper, we describe an exploratory study to address the aforementioned aspects and begin to understand the current state of DH research by focussing on the openness of its methodological process rather than its research outputs. In particular, we aim at having initial answers to the following research questions (RQ1-2), one related to the *knowledge production* practice and the other to the *knowledge evaluation* procedure:

1. How many publications which describe some data contain an explicit reference to external documentation detailing the process used to create and maintain such data, such as a DMP and/or a detailed workflow depicting the process of data creation?
2. How many DH venues adopt open peer reviewing practices?

To answer these two questions, we have devised two procedural workflows to gather relevant data from two distinct sets of resources. On the one hand, to address RQ1, we have downloaded and analysed, in a semi-automatic fashion, several articles describing data published in JOHD to identify references and mentions to external documentation detailing data management plans and methodological workflows adopted for data creation. On the other hand, to address RQ2, we have analysed several calls for papers and the editorial processes of DH conferences and journals to see how widely open reviewing is adopted within the community.
The rest of the paper is organised as follows. In Section 2, we introduce the materials used and the method developed to gather data to answer the research questions. In Section 3, we present the results obtained by analysing the data. In Section 4, we discuss the results and conclude the paper, sketching out some future work.

2. MATERIALS AND METHODS

To address the two RQs, we have developed two distinct workflows, available in (Peroni, 2026a) for reproducibility purposes, which are briefly introduced as follows. In addition, the related data obtained from running the workflow can be found in (Peroni, 2026c), while the related Data Management Plan can be found in (Peroni, 2026b).

Data Papers from the Journal of Open Humanities Data. The main goal of this activity is to understand if a specific Digital Humanities venue, i.e. JOHD, includes references or mentions to external material appropriately dedicated to either describing the management of the data a paper describes, such as a Data Management Plan (DMP), or detailing precisely the workflow used to produce these data, e.g.

published as a particular workflow document or as a computational notebook. In particular, we focus on specific kinds of articles published in JOHD, i.e. *Data Papers*, which are concise descriptions of a humanities research object with high reuse potential. These articles are particularly suitable for our activities since they describe only data, and thus they should have a higher possibility of having such documents referenced, given that the data are the focus of the article. It is worth clarifying, though, that such Data Papers usually include a methodological section that provides hints about the process of creating the data. However, the information in that section is often insufficient to fully reproduce the workflow steps required to recreate the same set of data.
To have a limited time window for the articles, we consider all the Data Papers included in volume 11 of JOHD, i.e. those published in 2025. We retrieve the basic metadata (DOI and title) for these Data Papers and download the corresponding XML and PDF files. Then, we parse all the XML versions of the articles to find out if their text contains any of the terms usually related with data management and analysis: *data management plan* and its acronym (*dmp*), *research data management* and its acronym (*rdm*), the broad concept of *data management* and, finally, two terms – i.e. *protocol* and *workflow* – that are commonly used to define processes devised for the creation and analysis of data. The rationale for using these terms for filtering relevant articles is that, instead of manually reviewing all downloaded articles, we focus on those that contain words that potentially indicate a mention of external documents defining DMPs, workflows, etc.
Finally, given the list of articles returned in the previous step, we read each article to see if it contains references or mentions to external documents describing the management or the creation of the data it introduces. In particular, we check if either (a) it contains a reference or a mention to such an external document, or (b) such a document is included in the data deposited and introduced in the article.

Reviewing processes in Digital Humanities venues. The main goal of this activity is to identify the practice of reviewing processes established in some DH venues, including conferences and journals. In particular, we want to check how many venues adopt open peer review processes instead of the classic ones that prescribe either the sole anonymity of reviewers (*single-blind* from now on) or the anonymity of both authors and reviewers (*double-blind* from now on).
Since the concept of open peer review can be decoupled along different dimensions depending on the possibilities a peer-reviewing process permits, we decided to adopt the taxonomy defined in (Ross-Hellauer, 2017), which comprises the following dimensions:

- open identities: authors and reviewers are aware of each other's identities;
- open reports: review reports are published alongside the relevant article;
- open participation: the wider community is able to contribute to the review process;
- open interaction: direct reciprocal discussion between author(s) and reviewers, and/or between reviewers, is allowed and encouraged;
- open pre-review manuscripts: manuscripts are made immediately available (e.g., via pre-print servers like arXiv) in advance of any formal peer review procedures;
- open final-version commenting: review or commenting on final "version of record" publications;
- open platforms (“decoupled review”): review is facilitated by a different organisational entity than the venue of publication.

We consider the following DH venues, of which websites have been analysed to find information about their call for papers and, in particular, the reviewing process adopted:

- the past five years of the annual conference organised by the *Associazione per l'Informatica Umanistica e la Cultura Digitale* (AIUCD, https://www.aiucd.it/);
- the past five years of the annual conference organised by the *Alliance of Digital Humanities Organizations* (ADHO, https://adho.org/);
- the DH journals in the list created by Spinaci et al. (2022) available at https://dhjournals.github.io/list/, in particular those that are (a) completely devoted to publishing DH articles, (b) still active, (c) have a URL associated, and (d) specify a standard article submission procedure.

All the retrieved information about the peer review process adopted by the various venues is extracted. When tracking open peer review processes, it is possible to consider more than one dimension per venue simultaneously if multiple dimensions are actually considered in the reviewing process. In addition, we also track the venues that do not clarify on their website the review process in use.

3. RESULTS

From JOHD, we downloaded 48 Data Papers, all published in 2025. From these, 17 articles matched at least one of the searched words, distributed as shown in the left graph in Figure 1. The terms *data management plan*, *DMP*, *research data management*, and *RDM* were not mentioned in any article. Out of these 17 articles, only 5 articles (around the 10% of the total) provided a full description of the protocol followed for creating the data introduced in the article, either as an external document or included in the data deposited in a relevant repository (e.g. Zenodo) [RQ1]. The 3 articles mentioning *data management* did not include any reference/mention to protocols, DMPs, or other relevant documentation.

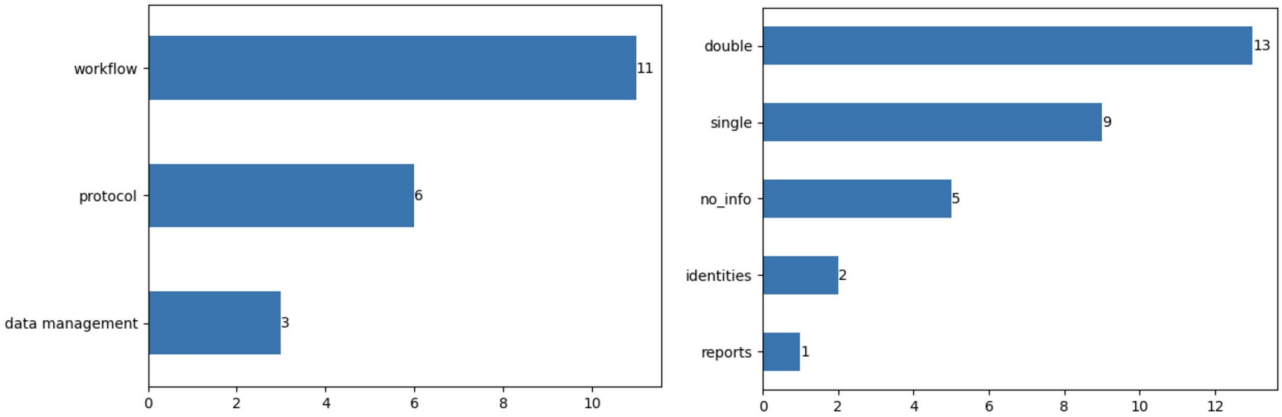


Figure 1. Two horizontal bar charts depicting some of the results obtained from the analysis. The left graph shows the distribution of the searched terms in the downloaded JOHD Data Papers, where each article may match more than one term. The right graph shows the different peer-reviewing processes adopted by the DH venues analysed, where "no_info" indicates those venues that did not have a clear reviewing process specified. Both graphs report only the dimensions that have been encountered at least once.

We have considered 29 distinct DH venues: 10 conference editions and 19 journals. The majority of the venues adopted a traditional reviewing process, with a preference for double-blind peer review (13 venues) over the single-blind peer review (9 venues), while 5 venues did not provide sufficient information to understand the peer review procedure used. Overall, only two venues adopted some form of open peer review in the past, i.e. DH 2023 (*open identities*) and DH 2025 (*open identities* and *open reports*) [RQ2]. The 5 AIUCD conference editions analysed, in contrast, had all adopted a double-blind peer review process.

4. DISCUSSIONS AND CONCLUSIONS

The results introduced by our exploratory study seem to suggest that there is still work to do by the DH community to properly address the openness of the Open Science area dedicated to the processes that regulate knowledge creation and evaluation. Indeed, looking with more detail at the results obtained by analysing the JOHD articles, it seems peculiar that, in all the Data Papers published in 2025, no mentions of the terms *data management plan* (and *DMP*) and *research data management* (*RDM*) have not been found in any of the 49 articles. While explicitly required by funders, such as the European Commission in all its funding programmes, it seems that the practice of including them as additional documentation to support the data researchers create and/or gather for answering specific questions is not systematically followed, even when looking at a venue, JOHD, explicitly dedicated to introducing data for Humanities research.

While DMPs were completely missing, we observed that a few articles at least provided an appropriate, detailed description of the procedural methodology used to generate the data described in the articles analysed. While not ideal, since it has not been systematically observed across all the articles considered, it seems that the community focuses more on the reproducibility of the approach used to generate the

data than on the administrative and managerial aspects of data before and after the research is completed. Another aspect that deserves future analysis in this regard is how detailed data creation workflows are organised to create data that are FAIR-by-design (Wilkinson et al., 2016), a fundamental practice to consider for reliable research data management, as highlighted by UNESCO (2022). In addition, recently several policy documents push for complementing the FAIR Principles with the CARE (*collective benefit*, *authority to control*, *responsibility*, *ethics*) principles (Carroll et al., 2020), which put "people and purpose, rather than data and technology" at the centre of RDM frameworks (Di Donato & Provost, 2025). Also, these aspects should be taken into strong consideration to address transparent and responsible RDM in DH projects.

The analysis of peer review process adoption in the considered venues revealed a worse scenario than the RDM practices mentioned above. With the exception of a few conference events that attempted to address at least the basic requirements of open peer review, the majority of DH venues do not currently support transparent, reliable processes for evaluating the scientific appropriateness of DH research contributions. While there have been ongoing discussions for years on the pros and cons of such a new peer review dynamics (Ross-Hellauer, 2017), all Open Science practices and norms released in the past years strongly support the open peer review process as the primary alternative to enable a broader setting of trust of the research endeavour. The hope is that this topic will spark constructive discussions within the community to understand how it can be implemented as part of modern policies for conducting transparent, reliable research.

ACKNOWLEDGEMENTS

This work has been funded by the European Union's Horizon Europe framework program under grant agreement No 101188018 (GRAPHIA).

REFERENCES

Barzaghi, S., Bordignon, A., Gualandi, B., & Peroni, S. (2025). Enlightening the Black Box of Humanities Research. Methodological Documentation as a Way to Transparency and Accountability of Digital Exhibitions. Umanistica Digitale, 9(20), 97–114. https://doi.org/10.6092/issn.2532-8816/21177

Carroll, S. R., Garba, I., Figueroa-Rodríguez, O. L., Holbrook, J., Lovett, R., Materechera, S., Parsons, M., Raseroka, K., Rodriguez-Lonebear, D., Rowe, R., Sara, R., Walker, J. D., Anderson, J., & Hudson, M. (2020). The CARE Principles for Indigenous Data Governance. Data Science Journal, 19, 43. https://doi.org/10.5334/dsj-2020-043

Di Donato, F., & Provost, L. (2025). Why isn't FAIR enough? Bringing together methods and values for Open Science uptake. Umanistica Digitale, 9(19), 17–46. https://doi.org/10.6092/issn.2532-8816/20976

McGillivray, B., Marongiu, P., Pedrazzini, N., Ribary, M., Wigdorowitz, M., & Zordan, E. (2022). Deep Impact: A Study on the Impact of Data Papers and Datasets in the Humanities and Social Sciences. Publications, 10(4), 39. https://doi.org/10.3390/publications10040039

Peroni, S. (2026a). Computational Workflows to Measure the Open Science Practices in the Digital Humanities Domain (Version v1.0.0) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.19758644

Peroni, S. (2026b). Data Management Plan for Measuring the Open Science Practices in the Digital Humanities Domain (Version 1) [Data Management Plan]. Zenodo. https://doi.org/10.5281/zenodo.19785333

Peroni, S. (2026c). Full-text data papers from JOHD and call for papers from DH venues (Version 1.0.0) [Dataset]. Zenodo. https://doi.org/10.5281/zenodo.19758167

Ross-Hellauer, T. (2017). What is open peer review? A systematic review. F1000Research, 6, 588. https://doi.org/10.12688/f1000research.11369.2

Spinaci, G., Colavizza, G., & Peroni, S. (2022). A map of Digital Humanities research across bibliographic data sources. Digital Scholarship in the Humanities, 37(4), 1254–1268. https://doi.org/10.1093/llc/fqac016

Tóth-Czifra, E., Błaszczyńska, M., Gelati, F., Admiraal, F., Blümm, M., Buelinckx, E., Chiquet, V., Gautschy, R., Gietz, P., Király, P., Vivas-Romero, M., Scholger, W., Szleszyński, B., & Wuttke, U.

(with Zalotyńska, A., & Taylor, L.). (2023). Research Data Management for Arts and Humanities: Integrating Voices of the Community (E. Tóth-Czifra, M. Błaszczyńska, & F. Gelati, Eds). Zenodo. https://doi.org/10.5281/zenodo.8059626

UNESCO. (2021). UNESCO Recommendation on Open Science (Programme and Meeting Document SC-PCB-SPP/2021/OS/UROS; p. 36). https://doi.org/10.54677/MNMH8546

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3, 160018. https://doi.org/10.1038/sdata.2016.18

Monitorare la transizione: l'Osservatorio sulle pratiche di Open Science nelle SSH come strumento di Public Engagement

PAMELA BARLETTA[1], MARTA CARADONNA[2], NICOLA GIAMPIETRO[3], ALICE ORRÙ[4], FEDERICO SILVESTRI[5]

[1] CNR – ILIESI, ITALIA, PAMELA.BARLETTA@CNR.IT,
[2] CNR – ILIESI, ITALIA, MARTA.CARADONNA@CNR.IT,
[3] CNR – ILIESI, ITALIA, ALICEORRU@CNR.IT, F
[4] CNR – ILIESI, ITALIA, FEDERICO.SILVESTRI@CNR.IT
[5] CREA, ITALIA, NICOLA.GIAMPIETRO@CREA.GOV.IT

ABSTRACT (ITALIANO)

Il contributo presenta il quadro teorico e metodologico dell'Osservatorio SSH, un'iniziativa volta allo sviluppo di strumenti analitici e digitali per lo studio delle pratiche editoriali scientifiche nelle Scienze Umane e Sociali (SSH) alla luce dei principi dell'Open Science (OS). A partire da un'analisi delle principali criticità del settore, il lavoro propone l'adozione di strumenti di monitoraggio del Diamond Open Access (DOA) e dell'applicazione dei principi FAIR, finalizzati a rendere osservabili, comparabili e riusabili i dati relativi alla comunicazione scientifica nelle SSH. Tali strumenti costituiscono vere e proprie infrastrutture di public engagement, capaci di favorire trasparenza, inclusione e partecipazione nei processi di produzione e disseminazione della conoscenza. In questa prospettiva, i dati della ricerca vengono interpretati come beni comuni digitali, accessibili e significativi anche per comunità non specialistiche. Il contributo si inserisce nel dibattito delle Digital Humanities (DH) proponendo una riflessione critica sul ruolo degli strumenti digitali e metodologici come mediatori tra ricerca accademica, pratiche editoriali e società.
Parole chiave: Scienza Aperta; Diamond Open Access; principi FAIR; Scienze Umane e Sociali; Public Engagement

ABSTRACT (ENGLISH)

Monitoring the transition: the Observatory on Open Science practices in the SSH as a tool for public engagement.
The paper presents the theoretical ad methodological framework of the SSH Observatory, an initiative aimed at the development of analytical and digital tools for studying scholarly editorial practices in Social Sciences and Humanities (SSH) in the context of Open Science (OS) principles. Moving from an assessment of main critical issues, the paper proposes the adoption of Diamond Open Access (DOA) and FAIR principles implementation monitoring tools, aiming at making SSH data observable, comparable and reusable. Such tools constitute genuine infrastructures for public engagement, capable of promoting transparency, inclusion and participation in the processes of knowledge production and dissemination. In this frame, research data are considered digital common goods, whose accessibility and significance also encompass non-specialist communities. The paper is set within the debate on Digital Humanities (DH), offering a critical reflection on the role of digital and methodological tools as mediators between academic research, publishing practices and society.
Keywords: Open Science; Diamond Open Access; FAIR principles; Social Sciences and Humanities; Public Engagement

1. INTRODUZIONE

Negli ultimi anni, l'Open Science (OS) e la pratica dell'Open Access (OA) si sono affermati come i principali paradigmi di trasformazione della ricerca scientifica, incidendo in modo significativo sui modelli di produzione, comunicazione e valutazione della conoscenza (Berlin Declaration, 2003) e diventando requisito centrale delle politiche europee (European Commission, 2021a: cap. 17) con l'obiettivo di

rendere i risultati della ricerca finanziata con fondi pubblici più trasparente, accessibile e riutilizzabile.[1] A livello europeo, l'adozione di politiche vincolanti sull'OA ha integrato requisiti quali la gestione dei dati secondo i principi FAIR e l'uso di infrastrutture interoperabili come EOSC[2] e OpenAIRE[3] (Manghi et al., 2019). Le Council Conclusions on Scholarly Publishing (Council EU, 2023) e iniziative come Plan S[4] hanno poi riconosciuto esplicitamente la crisi sistemica delle politiche OA basate sui modelli editoriali a pagamento per gli autori (APC), individuando nel Diamond Open Access (DOA) il paradigma più coerente con obiettivi di equità, sostenibilità e qualità della comunicazione scientifica. Nel settore delle Social Sciences and Humanities (SSH), tuttavia, l'implementazione dei principi della scienza aperta presenta caratteristiche e criticità specifiche. Le SSH sono contraddistinte da una forte eterogeneità disciplinare, da modelli editoriali spesso non commerciali e community-owned,[5] da un ruolo centrale delle lingue nazionali e da pratiche di comunicazione scientifica che non sempre si allineano ai modelli dominanti delle scienze dure.

All'interno di questo scenario, il presente contributo propone:

1) un'analisi delle pratiche di OS nelle SSH nel contesto italiano, attraverso la costruzione e l'esplorazione di un dataset originale di riviste in DOA;
2) la costituzione di un Osservatorio SSH come infrastruttura di public engagement, finalizzato allo sviluppo di strumenti analitici e digitali per lo studio delle pratiche editoriali SSH alla luce dei principi dell'OS;
3) un confronto sperimentale tra un controllo manuale della classificazione delle riviste SSH italiane e l'uso dell'IA per individuare eventuali incongruenze o criticità nei dati nell'ottica di valutare i limiti e le potenzialità dell'automazione in presenza di informazioni incomplete o ambigue.

L'obiettivo è triplice: da un lato, analizzare criticamente le condizioni strutturali dell'editoria scientifica italiana in relazione al contesto europeo; dall'altro, mostrare come l'adozione di strumenti di monitoraggio del DOA e dei principi FAIR (Findable, Accessible, Interoperable, and Reusable) possa condurre a rendere i dati della comunicazione scientifica osservabili, confrontabili e riusabili; dall'altro ancora, effettuare una ricognizione sistematica dell'editoria scientifica SSH, caratterizzata da elementi spesso impliciti o non strutturati (modelli di governance, condizioni di accesso e stato di attività delle riviste).

2. L'OSSERVATORIO SSH: MISSIONE E PROSPETTIVE

Sorto nel contesto della progressiva istituzionalizzazione dell'OS a livello europeo e nazionale, l'Osservatorio SSH, supportato dall'attività del Gruppo di Lavoro sull'Open Science (GdL OS),[6] si propone, al di là degli scopi tecnici e valutativi, di trasformare, nell'ottica del public engagement, dati frammentati e opachi in dati strutturati accessibili anche a comunità non specialistiche. Configurandosi come un ponte tra ricerca, policy dell'OS e comunità editoriali, l'Osservatorio SSH ambisce dunque a interpretare criticamente l'evoluzione normativa, politica e infrastrutturale del sistema della comunicazione scientifica negli ultimi due decenni, a produrre consapevolezza degli strumenti per la fairificazione delle SSH e a promuovere la discussione pubblica, il confronto critico e la riflessione collettiva.
In Italia la transizione verso l'OS si è sviluppata in modo disomogeneo e in ritardo rispetto al panorama europeo. A partire dalla Legge 112/2013[7] e dalle successive policy istituzionali di università ed enti pubblici di ricerca (EPR), si è progressivamente affermata una pluralità di iniziative OA, spesso prive di un coordinamento nazionale strutturale (in particolare nel settore delle SSH), con una frammentazione nei

[1] In particolare, il Programma Quadro FP7 e Horizon Europe hanno sancito come requisito di finanziamento il principio dell'accesso senza embargo ai risultati della ricerca. Cfr. Le indicazioni della Commissione Europea sull'OS: https://rea.ec.europa.eu/open-science_en (cons. 21/04/2026)
[2] https://eosc.eu (cons. 05/02/2026)
[3] https://www.openaire.eu/ (cons. 05/02/2026)
[4] https://www.coalition-s.org/why-plan-s/ (cons. 21/04/2026)
[5] Nel contesto della comunicazione scientifica, community-owned si riferisce a riviste, piattaforme o infrastrutture editoriali possedute e governate dalla comunità accademica (ricercatori, biblioteche, istituzioni), anziché da editori commerciali, con modelli decisionali partecipativi e orientati all'interesse collettivo della ricerca.
[6] Per i risultati della ricerca dei primi due anni di attività, cfr. Caradonna et al. (in preparazione).
[7] https://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:legge:2013-10-07;112 (cons. 05/02/2026)

modelli economici che editoriali, nelle infrastrutture tecniche e negli standard di qualità che incide negativamente sulla visibilità, sulla sostenibilità e sulla valutazione della ricerca (Galimberti, 2020). In sinergia con iniziative nazionali ed europee (European Commission, 2016), l'Osservatorio SSH si propone di mappare l'editoria scientifica italiana nelle SSH, monitorando, raccogliendo e integrando dati relativi a politiche editoriali, modelli di governance, infrastrutture tecniche e livelli di adozione degli standard FAIR e del DOA quale dispositivo di disseminazione pubblica della conoscenza. Intersecando dunque Digital Humanities (DH), OS e responsabilità sociale della ricerca, la missione non è esclusivamente descrittiva ma soprattutto inclusiva: i dati quali beni comuni digitali diventano utili ai decisori istituzionali e alle comunità accademiche, alle biblioteche, agli editori non commerciali e a tutto il pubblico interessato, nella prospettiva di un ecosistema editoriale più trasparente, equo e partecipato.

3. FARE OPEN SCIENCE NELLE SSH: CRITICITÀ E OSTACOLI

Parallelamente alla promozione politica europea di strategie per l'OA alle pubblicazioni scientifiche, la condivisione dei dati e l'adozione di standard comuni per la loro gestione e valutazione (European Commission, 2021b), nel contesto italiano l'implementazione dei principi dell'OS nelle SSH presenta delle criticità specifiche. Istituzionalmente, la mancanza di un coordinamento nazionale stabile comporta una marcata eterogeneità nelle pratiche editoriali e una condizione di precarietà strutturale delle riviste OA (risorse limitate, volontariato accademico, competenze non formalizzate). Persiste poi una tensione tra pratiche di OS e sistemi di valutazione della ricerca. Nelle SSH, i meccanismi di riconoscimento istituzionale continuano a privilegiare modelli editoriali tradizionali, rendendo marginale il ruolo delle riviste OA non commerciali e limitando gli incentivi all'adozione di pratiche pienamente aperte. L'apertura è spesso dichiarata a livello di policy, ma non sempre sostenuta da infrastrutture, strumenti e modelli organizzativi adeguati (cfr. Caso, 2020; Morriello, 2021; Pievatolo, 2026). L'OS nelle SSH richiede, più che un insieme di prescrizioni normative o tecniche, un ripensamento complessivo dei modelli editoriali, delle infrastrutture di supporto e delle modalità di valorizzazione della ricerca. Il ruolo del DOA diventa cruciale proprio nella tensione tra principi, pratiche e strumenti. Queste esigenze sono state confermate nella fase qualitativa di ricerca condotta mediante interviste semi-strutturate a tre opinion leader dell'OS in Italia, un'indagine dal carattere esplorativo che tuttavia restituisce già un quadro complessivo delle lacunosità della situazione nazionale.[8]

4. DIAMOND OPEN ACCESS: UNO STRUMENTO DI CONOSCENZA DEMOCRATICA

Nel panorama delle SSH, il DOA rappresenta uno dei modelli editoriali più coerenti con i principi fondativi dell'OS (Arasteh-Roodsary et al., 2025). L'assenza di costi sia per gli autori sia per i lettori, unita alla governance accademica e non profit delle riviste, consente di concepire la comunicazione scientifica come un'infrastruttura pubblica orientata alla produzione e alla circolazione della conoscenza come bene comune. D'altra parte, il potenziale democratico del DOA consta nell'eliminazione delle barriere economiche così come nell'adozione sistematica dei principi FAIR nelle pratiche editoriali. Senza una corretta identificazione persistente delle risorse, una metadatazione strutturata, licenze chiare e standard condivisi, l'OA rischia infatti di tradursi in una forma di accessibilità solo apparente, che non favorisce realmente l'accesso, la partecipazione e il riuso della conoscenza. Nel contesto italiano delle SSH, il DOA presenta spesso un'implementazione disomogenea: accanto a esperienze mature e ben integrate nelle infrastrutture europee (OPERAS,[9] DARIAH,[10] CLARIN,[11] E-RIHS,[12] OpenAIRE), persistono casi in cui l'assenza di standard FAIR limita la visibilità, la circolazione e la riusabilità dei contenuti (cfr. *infra*, 5.3). In questo senso, l'integrazione tra DOA e FAIR costituisce non un adempimento tecnico, ma un passaggio strategico per trasformare le riviste e i dati della ricerca in strumenti di public engagement, anche in un'ottica di citizen science (Cassella, 2022). L'adozione congiunta di DOA e FAIR consente infatti di

[8] La metodologia di analisi segue De Lillo, 2010 e Cardano, 2003. Le interviste si basano su tre "stimoli" di ricerca relativi, nel contesto italiano: 1) allo stato dell'arte dell'editoria aperta nelle SSH (di contro alle scienze dure); 2) alle prospettive future del ruolo degli editori in ambito di scienza aperta; 3) agli esempi istituzionali virtuosi di editoria aperta e alle loro caratteristiche fondamentali. La scelta degli opinion leader (Elena Giglia, UniTO; Fulvio Guatelli, UniFI-FUP; Stefano Bianco, INFN) ha voluto tenere conto di atenei/EPR rappresentativi della scienza aperta in Italia, dell'ambito tecnologico, dell'ambito professionale editoriale, e del confronto SSH vs. non-SSH.
[9] https://operas-eu.org/ (cons. 21/04/2026)
[10] https://www.dariah.eu/ (cons. 21/04/2026)
[11] https://www.clarin.eu/ (cons. 21/04/2026)
[12] https://www.e-rihs.it/ (cons. 21/04/2026)

rendere osservabili e riutilizzabili non solo i prodotti finali della ricerca, ma anche i processi editoriali e i dati che li sostengono. Ciò favorisce forme di trasparenza, inclusione e partecipazione che, oltre la comunità accademica, coinvolgono biblioteche, istituzioni culturali, decisori pubblici e cittadini interessati. In questa prospettiva, il DOA si configura come un dispositivo centrale per una OS orientata alla responsabilità sociale e alla democratizzazione del sapere nelle SSH (Cassella, 2023).

5. ANALIZZARE IL CONTESTO DELL'EDITORIA ITALIANA: DAL DATO "PER POCHI" AL DATO "PER TUTTI"

Costruire un dataset affidabile sull'editoria OA italiana nelle SSH si è rivelata, sin dalle prime fasi, un'operazione non banale, a causa del disallineamento di politiche, infrastrutture e pratiche editoriali tra atenei, EPR e comunità disciplinari e della variabilità della qualità della documentazione pubblica (siti web, metadati, registri, policy). Scarseggiano le piattaforme che promuovono l'OA, nonostante la gran parte degli EPR e degli atenei (statali e, più raramente, privati riconosciuti dal MUR) abbia varato proprie policy istituzionali per l'open science e l'OA, con la consapevolezza dell'opportunità di valorizzare le conoscenze attraverso la terza missione e il public engagement (es. CRUI, 2025). Questa difficoltà è un elemento strutturale del caso italiano: in particolare, ciò emerge mettendo in relazione il monitoraggio nazionale con standard e iniziative europee come DIAMAS[13] e l'infrastruttura di ricerca EDCH[14] (lanciata il 15 gennaio 2025), rendendo necessario individuare buone pratiche emergenti e strategie istituzionali nel panorama universitario e degli EPR (Angioni et al., 2025).

5.1 Dal dato "per pochi" al dato riusabile: cosa implica raccogliere dati sull'editoria

Nel corso dell'analisi, il "dato" non è stato concepito come mero risultato per la compilazione di un elenco di riviste. Piuttosto, il primo obiettivo posto è stato rendere osservabili elementi che presentano spesso le seguenti caratteristiche:

1. incompletezza o contraddittorietà (licenze dichiarate ma non applicate, informazioni presenti in homepage ma assenti negli articoli);
2. difficoltà di interpretazione senza contesto (ownership, governance, assenza/presenza di fee, apertura a tutti gli autori);
3. non verificabilità attraverso un'unica fonte (necessità di triangolare sito, registri, cataloghi e metadati).

Questo passaggio è cruciale anche in chiave FAIR: un dato "per pochi" è spesso un dato non replicabile, non trasmissibile perché legato (quasi esclusivamente) all'esperienza di chi lo raccoglie; un dato "per tutti" è invece un dato tracciabile, controllabile e aggiornabile, quindi utile come infrastruttura di public engagement. Il dato risulta così non solo alla portata del ricercatore, ma anche del cittadino, che diventa dunque partecipe del processo di costruzione del sapere scientifico (Cassella, 2023).

5.2 Il dataset manuale

Nel 2025 il GdL OS ha effettuato una ricognizione sistematica dell'editoria scientifica DOA in Italia, con particolare attenzione alle riviste delle SSH afferenti a università ed EPR nazionali. L'obiettivo finale era non censire, ma rendere osservabili e verificabili le condizioni reali di apertura, governance e sostenibilità editoriale, al di là delle informazioni dichiarative disponibili nei registri o nei siti web.
Per questo motivo, si è proceduto anzitutto alla costruzione manuale del dataset delle riviste italiane OA community-owned (cfr. *infra,* 5.3), secondo selezionati criteri di classificazione e di analisi sulla base di specifiche fonti.[15] Il passo successivo è stato confrontare il dataset generale con la sua replica attraverso ChatGPT[16] (cfr. *infra*, 5.4), al fine di verificare, dal punto di vista quantitativo e qualitativo, l'effettivo contributo umano rispetto all'automatismo delle macchine in tale tipologia di ricerca.

[13] https://diamasproject.eu (cons. 05/02/2026)
[14] https://edch.eu (cons. 05/02/2026)
[15] Segnatamente dai seguenti elenchi: atenei ed EPR italiani; University Press Italiane (UPI): https://www.universitypressitaliane.it/ (cons. 21/04/2026); riviste scientifiche e in Classe A ANVUR (per le SSH Aree 10 e 11, per filosofia 11/C1-5); report delle riviste italiane per il Diamond Discovery Hub (DDH); siti delle riviste; Catalogo italiano dei periodici (ACNP); Directory of Open Access Journals (DOAJ). Cfr. *infra*, 5.3.
[16] Versione GPT-4.x

5.3 Struttura e metodologia di costruzione

La costruzione del dataset ha visto quattro fasi: 1) rendicontazione degli atenei ed EPR e la scelta dei relativi criteri di classificazione; 2) selezione di atenei ed EPR relativi all'ambito delle SSH e delle relative riviste OA; 3) ulteriore selezione tra queste ultime, per interessi di ambito di ricerca, delle riviste OA di ambito filosofico; 4) costruzione del "dataset generale" delle riviste italiane DOA per il Diamond Discovery Hub (DDH).
Nella prima fase è stato prodotto un elenco di 124 record di atenei e EPR riconosciuti dal MUR, di cui 95 per le SSH. Di seguito, si è proceduto all'individuazione di atenei e EPR particolarmente impegnati nell'applicazione delle policies OA. Una prima classificazione grezza è stata per ateneo/EPR, casa editrice, titolo e ISSN della rivista pubblicata (in corso/cessata). Relativamente a questo elenco, sono stati stabiliti alcuni ulteriori parametri di ricerca: l'appartenenza ad aggregazioni di ateneo/EPR per l'OS; i tipi, le licenze (specifica di Creative Commons) e la policy OA utilizzati dall'ateneo/EPR; gli eventuali referenti e comitati per l'OS e l'OA; la pubblicazione o patrocinio di riviste/monografie; le aree disciplinari SSH coinvolte; eventuale numero totale di pubblicazioni in OA. Nella seconda fase, sono state rendicontate le riviste OA per i 95 atenei ed EPR italiani del campo delle SSH: in questo caso, come primo criterio generale è stato utilizzato quello della casa editrice di ateneo/EPR (o commerciale, che pubblica riviste patrocinate dall'ente). I criteri di selezione e classificazione delle riviste per ogni ateneo/EPR (raggruppate secondo appartenenza o affiliazione) sono stati i seguenti: peer review/no, l'indicizzazione in ANVUR/Classe A,[17] la licenza adottata/specificata, OA/no e tipo, la consistenza della rivista, i link di accesso alla rivista e alla piattaforma OA dell'ateneo/EPR, le tematiche inter- e intra-disciplinari di ambito filosofico. Nella terza fase è stata effettuata una scrematura qualitativa delle riviste OA di ambito filosofico attraverso: 1) la consultazione di eventuali informazioni con la formula "[nome Università/EPR] riviste filosofia open access" sui motori di ricerca; 2) la selezione dei risultati pertinenti, conferendo priorità, in primo luogo, al sito della rivista e/o dell'ateneo/EPR (es. policy OA) e, in secondo luogo, ai siti di riviste che ne presentano il patrocinio (OJS/no); 3) l'individuazione delle informazioni da reperire (presenza o no sui siti; ricavabili da ulteriori fonti, in caso segnalate e confrontate tra loro); l'ulteriore selezione dei risultati pertinenti a partire dal criterio DOA "community-owned" (selezionato, tra i sei, quale "guida" per l'indagine in oggetto). La successiva analisi quantitativa ha tenuto conto della media dei descrittori in DOAJ,[18] ACNP access,[19] homepage del sito della rivista e della classificazione ANVUR. Ne è emerso che gli atenei ed EPR italiani proprietari di riviste OA in ambito SSH risultano essere 13 nel Nord Italia (5 nel Nord-ovest e 8 nel Nord-est), 18 nel Centro Italia, 8 nel Sud Italia e 3 nelle Isole. Se si considera il solo ambito filosofico, essi scendono a 8 nel Nord Italia (2 nel Nord-ovest e 6 nel Nord-est), 9 nel Centro Italia, 6 nel Sud Italia e 2 nelle Isole. D'altra parte, le riviste italiane OA in corso di pubblicazione per l'ambito SSH, di cui atenei ed EPR sono proprietari, risultano essere in totale 142, di cui 46 con sede nel Nord Italia (24 nel Nord-ovest e 22 nel Nord-est), 66 nel Centro Italia e 15 sia nel Sud Italia che nelle Isole. Restringendo il calcolo all'ambito filosofico, le riviste attive diventano 61, di cui 29 con sede nel Nord Italia (17 nel Nord-ovest e 12 nel Nord-est), 17 nel Centro Italia, 11 nel Sud Italia e 4 nelle Isole.[20] Nella quarta fase, è stato strutturato un "dataset generale" delle riviste italiane DOA in ambito SSH integrando le informazioni raccolte nelle prime tre fasi e i risultati di controllo/verifica dei dati del primo elenco[21] fornito da DIAMAS per la costruzione del servizio Diamond Discovery Hub (DDH)[22] lanciato a ottobre 2025 come risultato del progetto CRAFT-OA.[23] Delle 462 riviste riportate, 391 sono risultate appartenenti all'ambito SSH contro le 71 relative all'ambito delle scienze dure, 3 sono risultate "discontinued" (equivalente a cessata), 30 non "community-owned" e 5 non "open to all authors".[24] La scelta di strutturare il dataset generale in base a questi ultimi due criteri sta nel fatto che, limitandoci al

[17] https://www.anvur.it/it/ricerca/riviste/elenchi-di-riviste-classificate (cons. 21/04/2026) (https://www.wikidata.org/wiki/Wikidata:Gruppo_Wikidata_per_Musei,_Archivi_e_Biblioteche)
[18] https://doaj.org/ (cons. 21/04/2026)
[19] https://acnpsearch.unibo.it/ (cons. 21/04/2026)
[20] L'ultimo aggiornamento di ricerca dei dati è del 27/01/2026. Cfr. Caradonna et al. (in preparazione). https://it.wikipedia.org/wiki/Progetto:Coordinamento/Universit%C3%A0/UniVR/Laboratori_Wikimedia_2023-2024
[21] Dei 462 record presenti nel documento "Report IT" (riguardante tutti gli ambiti di ricerca scientifica accademica), alla data di ultima consultazione (16/04/2026) sulla piattaforma risultano 416 riviste DOA con criteri Diamond "Fully Diamond" e Paese "Italia", di cui 309 con l'ulteriore criterio Lingua "Italiana".
[22] https://ddh.edch.eu/en (cons. 18/04/2026)
[23] https://www.craft-oa.eu/ (cons. 21/04/2026)
[24] Per i criteri operazionali DOA cfr. Armengou et al., 2024.

confronto per l'ambito filosofico, circa la metà (33 su 61) delle riviste di filosofia ricavate nella fase 3 non sono presenti nel dataset generale della fase 4, facendo emergere numerose discrepanze motivate sia dall'indeterminatezza di alcuni criteri (es. il fatto che, ai fini del DDH, "community-owned" riguardi non solo atenei/EPR, ma più ampiamente organizzazioni pubbliche o senza scopo di lucro), sia dal mancato adeguamento agli standard FAIR (cfr. *supra*, 4).
Il dataset generale manuale è metodologicamente rilevante e necessario poiché intercetta proprio ciò che spesso resta fuori dai metadati. Esso è stato costruito applicando indicatori qualitativi e strutturali (assenza di fee, ownership community-owned, apertura agli autori, stato di attività, allineamento a standard di base), verificati attraverso un processo di triangolazione delle fonti che ha incluso l'analisi diretta dei siti, il confronto con registri nazionali e internazionali e, nei casi ambigui, contatti diretti con redazioni ed editori.

5.4 Dal workflow umano al controllo assistito: primo esperimento di replica del dataset per il DDH con ChatGPT e risultati comparativi

A partire dagli stessi titoli e indicatori del dataset generale manuale, è stato sperimentato un tentativo di replica automatica tramite ChatGPT per testare l'IA come uno strumento di prima classificazione e supporto alla rilevazione di incongruenze. Considerando la medesima copertura (462 record coincidenti), le differenze tra i risultati emergono nella qualità delle classificazioni, soprattutto sugli indicatori più "interpretativi", che sono considerati in dettaglio di seguito:

1. "Open to all authors": alta convergenza, ma con falsi positivi. ChatGPT tende a classificare quasi tutte le riviste come aperte a tutti gli autori; rispetto al dataset manuale, emergono pochi scostamenti (circa 1% di casi in cui l'IA assegna "sì" dove il Gruppo aveva rilevato "no" o non conformità). Qui l'automazione funziona perché il criterio è spesso implicito nelle policy standard, ma resta il rischio di assunzioni non verificate quando la policy non è esplicita;
2. "Community-owned": molte sottostime dell'IA. Qui emerge una divergenza significativa: nei casi valutabili, la replica automatica presenta circa 22% di falsi negativi (riviste che il Gruppo classifica come community-owned, ma che ChatGPT marca come non community-owned). Questo scarto è coerente con un limite strutturale: l'ownership è spesso deducibile solo da indizi distribuiti (chi ospita ≠ chi possiede; editori universitari; società scientifiche; governance non esplicitata), e talvolta emerge solo tramite contatto diretto;
3. "Discontinued journal": errore sistematico nella replica automatica. Nel caso dello status di "rivista cessata", la replica con ChatGPT genera un tasso di falsi positivi estremamente elevato. Delimitando un arco temporale preciso dalla data dell'ultima pubblicazione, l'IA etichetta come "discontinued" quasi tutti i record (circa il 98% con un intervallo di 18 mesi), catalogando come cessate riviste che hanno pubblicato numeri nel periodo indicato ed apparentemente ignorando altri segnali di attività (es. Call for Papers aperte). Nonostante il dato sia, in linea di principio, di natura binaria, si tratta di un'informazione in genere non esplicitata e tantomeno presente in forma strutturata. Dunque, quando un indicatore richiede una definizione operativa netta e una verifica puntuale (es. "attiva/cessata"), l'IA può produrre errori sistematici se interpreta male il campo o se deduce da segnali deboli.

5.5 L'IA come strumento di controllo e monitoraggio: pro e contro

Sicuramente l'IA riduce i tempi di una prima ricognizione e aiuta a scalare su grandi quantità di titoli. La sua utilità sta nell'individuare rapidamente anomalie (es. siti irraggiungibili, metadati mancanti, incoerenze grossolane) e nel supportare aggiornamenti frequenti come "sensore" preliminare. L'IA tende poi a "colmare i vuoti" con inferenze: tuttavia, nelle SSH, proprio l'assenza è spesso un dato (policy non dichiarata, governance opaca, licenze non esplicitate). Riguardo ai criteri che richiedono interpretazione istituzionale (ownership, governance, sostenibilità), l'IA mostra il rischio di errori sistematici se il prompt o la mappatura degli indicatori non sono estremamente dettagliati (es. il caso "discontinued"). Il suo uso è promettente se inteso come secondo livello (controllo, triage, segnalazione) e non come sostituzione del lavoro umano: il dato "per tutti" nasce quando l'automazione è inserita in un workflow che conserva verifica, triangolazione e responsabilità del dato.

6. CONCLUSIONI

Il contributo mostra come la transizione verso l'OS nelle SSH non possa essere affrontata esclusivamente attraverso interventi normativi o dichiarazioni di principio, ma richieda strumenti capaci di rendere visibili, analizzabili e condivisibili le pratiche reali della comunicazione scientifica, esigenza particolarmente urgente nel contesto italiano, caratterizzato da forte frammentazione istituzionale ed eterogeneità editoriale. L'analisi di tale contesto ha evidenziato come il DOA rappresenti, nelle SSH, un modello coerente con una concezione della conoscenza come bene comune. Il confronto tra il dataset costruito manualmente e il tentativo di replica automatica tramite ChatGPT ha messo in luce opportunità e limiti dell'IA come strumento di supporto al monitoraggio. L'esperimento conferma che l'automazione può rafforzare, ma non sostituire, un flusso di lavoro fondato sulla triangolazione delle fonti e sulla responsabilità del dato, risultando al tempo stesso efficace nelle fasi preliminari di ricognizione, soprattutto nella gestione di grandi volumi di dati e nell'individuazione di anomalie. Dal punto di vista metodologico, il lavoro evidenzia come la raccolta manuale dei dati resti essenziale nei contesti in cui le informazioni sono incomplete, distribuite o implicite, in particolare per quanto riguarda aspetti quali governance editoriale, modelli di ownership e stato di attività delle riviste. L'Osservatorio SSH si configura in tal senso non solo come dispositivo tecnico, ma come infrastruttura di public engagement, capace di rendere accessibili e significativi dati altrimenti dispersi. Sul piano teorico e applicativo, si mostra, dunque, come la costruzione di dataset e strumenti di monitoraggio possa farsi pratica di mediazione tra ricerca, infrastrutture e società. In conclusione, i risultati evidenziano come il rafforzamento delle pratiche di OS nelle SSH sia strettamente legato allo sviluppo di modelli di monitoraggio integrati, in grado di coniugare strumenti digitali, rigore metodologico e responsabilità nella costruzione del dato.

BIBLIOGRAFIA

Angioni, S., Baglioni, M., Bardi, A., Manghi, P., Mannocci, A., & Pavone, G. (2025). Charting the landscape of Italian Diamond Open Access publishing. Zenodo. https://doi.org/10.5281/zenodo.17484911

Arasteh-Roodsary, S. L., Gaillard, V., Garbuglia, F., Mounier, P., Pölönen, J., Proudman, V., Rooryck, J., Saenen, B., & Stone, G. (2025). Diamond open access recommendations and guidelines for institutions, funders, sponsors, donors, and policymakers (1.0). Zenodo. https://doi.org/10.5281/zenodo.15518745

Armengou, C., Bargheer, M., Gingold, A., Holsinger, S., Laakso, M., Mitchell, D., Mounier, P., Pölönen, J., Rooryck, J., Ševkušić, M., Souyioultzoglou, I., & Varachkina, H. (2024). Operational Diamond OA Criteria for Journals. Zenodo. https://doi.org/10.5281/zenodo.12721408

Berlin Declaration [= Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities]. (2003). Berlin: Max Planck Society. https://openaccess.mpg.de/67605/berlin_declaration_engl.pdf

Caradonna, M., Barletta, P., Giampietro, N., Orrù, A., Silvestri, F. (in preparazione). La scienza aperta nelle SSH. Roma: CNR Edizioni.

Cardano, M. (2003). Tecniche di ricerca qualitativa. Percorsi di ricerca nelle scienze sociali. Roma: Carocci.

Caso, R. (2020). La rivoluzione incompiuta. La scienza aperta tra diritto d'autore e proprietà intellettuale. Milano: Ledizioni.

Cassella, M. (2022). La scienza di tutti. Biblioteche Oggi, 1. 10.3302/0392-8586-202201-006-1

Cassella, M. (2023). Citizen science, open science e digital humanities: relazioni e contaminazioni. Biblioteche Oggi, 1. 10.3302/0392-8586-202301-006-1

Council EU [= Council of the European Union]. (2023). Council conclusions on high-quality, transparent, open, trustworthy and equitable scholarly publishing. Official Journal of the European Union. https://data.consilium.europa.eu/doc/document/ST-9616-2023-INIT/en/pdf

CRUI. (2025). Il ruolo delle università per il progresso della società. Dalla Terza missione alla valorizzazione delle conoscenze. CRUI.IT. https://www.crui.it/documenti/54/New-category/1609/20250702-ruolo-uni-REV02-compresso.pdf

De Lillo, A. (2010) (a cura di). Il mondo della ricerca qualitativa. Torino: UTET.

European Commission. (2016). Open innovation, open science, open to the world: A vision for Europe. Publications Office of the European Union. https://doi.org/10.2777/061652

European Commission. (2021a). Horizon Europe programme guide. Version 5.1 (15 September 2025). European Union. https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/programme-guide_horizon_en.pdf

European Commission. (2021b). Towards a reform of the research assessment system. Scoping report. Publications Office of the European Union. https://op.europa.eu/publication-detail/-/publication/36ebb96c-50c5-11ec-91ac-01aa75ed71a1

Galimberti, P. (2020). Open Science and Evaluation. SCIRES-IT, 10, 65-70. http://dx.doi.org/10.2423/i22394303v10Sp65

Manghi, P., Artini, M., Atzori, C., Baglioni, M., Bardi, A., La Bruzzo, S., De Bonis, M., Dimitropoulos, H., Foufoulas, I., Iatropoulou, K., Manola, N., Martziou, S., & Principe, P. (2019). OpenAIRE: Advancing open science. The Grey Journal, 15 (3), 141-146. https://iris.cnr.it/bitstream/20.500.14243/361979/2/prod_416446-doc_146793.pdf

Morriello, R. (2021). Scienza aperta: una rivoluzione incompiuta. Considerazioni a margine del libro di Roberto Caso. Biblioteche Oggi, 39 (6). 10.3302/0392-8586-202106-006-1

Pievatolo, M. C. (2026). Di professori e di sgherri: scienza (aperta) e amministrazione universitaria. Il ponte, 41 (1), 6-12. https://doi.org/10.5281/zenodo.18164570

Nel 25° anniversario di Wikipedia: l'ecosistema Wikimedia e le Digital Humanities tra pratiche collaborative e sostenibilità della conoscenza

PIERGIOVANNA GROSSI
INDEPENDENT RESEARCHER, VERONA, ITALY[1] — ME@PIERGIOVANNAGROSSI.IT

ABSTRACT (ITALIANO)

Nel 2026, Wikipedia compie venticinque anni, in un contesto segnato, da un lato, dalla crescente adozione dell'intelligenza artificiale nella produzione e circolazione dei contenuti e, dall'altro, da sempre più marcate dinamiche di concentrazione e controllo dei dati. A partire da questo scenario, il contributo interpreta l'ecosistema Wikimedia come infrastruttura socio-tecnica di knowledge commons e come laboratorio per le Digital Humanities orientate al public engagement. Dopo un inquadramento storico (dalla revisione esperta di Nupedia alla partecipazione distribuita resa possibile dal software wiki), vengono discusse le affinità tra Wikimedia e prospettive critiche delle DH, in cui la conoscenza è intesa come processo pubblico, negoziato e trasparente. La relazione è letta in chiave bidirezionale: i progetti Wikimedia forniscono piattaforme, comunità e dati aperti per ricerca e didattica; le DH offrono cornici teorico-metodologiche per rafforzare qualità, sostenibilità ed equità delle pratiche collaborative. Attraverso quattro ambiti esemplificativi — didattica e citizen science, collaborazioni GLAM (Galleries, Libraries, Archives and Museums) e patrimonio culturale, diversità linguistica e inclusione, dati aperti e principi FAIR in Wikidata — si mostra come la co-costruzione del sapere si traduca in pratiche di cura, validazione, mappature partecipate e riuso. In conclusione, Wikimedia è proposto come spazio di citizen humanities capace di sostenere la Terza Missione accademica, rendendo la ricerca più accessibile, verificabile e socialmente rilevante anche nell'era dell'automazione.
Parole chiave: Citizen humanities; Public engagement; Wikimedia/Wikipedia; Wikidata; GLAM.

ABSTRACT (ENGLISH)

On the Twenty-Fifth Anniversary of Wikipedia: The Wikimedia Ecosystem and Digital Humanities between Collaborative Practices and the Sustainability of Knowledge.
In 2026, Wikipedia marks twenty-fifth anniversary in a context shaped by the pervasive adoption of artificial intelligence in content production and circulation and by increasingly pronounced dynamics of data concentration and control. Against this backdrop, the paper examines the Wikimedia ecosystem as a socio-technical infrastructure of knowledge commons and as a laboratory for Digital Humanities oriented toward public engagement. After a historical outline (from Nupedia's expert-driven model to the distributed participation enabled by wiki software), the contribution highlights the affinity between Wikimedia projects and critical DH approaches, where knowledge is conceived as a public, negotiated, and transparent process. The relationship is framed as bidirectional: Wikimedia provides platforms, communities, and open resources for research and teaching, while Digital Humanities offer theoretical and methodological tools to strengthen the quality, sustainability, and equity of collaborative practices. Four exemplary domains are discussed — education and citizen science, GLAM (Galleries, Libraries, Archives, Museums) partnerships and cultural heritage, linguistic diversity and inclusion, and open data with FAIR principles in Wikidata — showing how knowledge co-construction materializes through care, validation, participatory mapping, and reuse. The paper ultimately proposes Wikimedia as a citizen humanities space that can reinforce academia's Third Mission by making research more accessible, verifiable, and socially relevant in the age of automation.
Keywords: Citizen humanities; Public engagement; Wikimedia/Wikipedia; Wikidata; GLAM.

[1] Il presente contributo si inserisce nel quadro delle mie collaborazioni con Wikimedia Italia e si basa in parte su esperienze progettuali sviluppate in collaborazione con istituzioni culturali e accademiche. In particolare, il Dipartimento di Lingue e Letterature Straniere dell'Università di Verona ha contribuito in modo significativo a orientare le prospettive qui discusse sul rapporto tra ecosistema Wikimedia e Digital Humanities.

1. INTRODUZIONE

Nel 2026, Wikipedia celebra il suo venticinquesimo anniversario, un traguardo che assume un significato particolarmente rilevante nel contesto contemporaneo, segnato dalla crescente centralità dell'intelligenza artificiale (AI) e dei Large Language Models (LLM) nella produzione e distribuzione di contenuti, e da processi di concentrazione e controllo dei dati da parte di attori pubblici e privati. In questo scenario, l'importanza di piattaforme che fondano la loro esistenza sull'aggregazione, sulla negoziazione e sulla cura collettiva dei contenuti emerge con rinnovata urgenza.
Wikipedia e i progetti della Wikimedia Foundation costituiscono uno degli esempi più estesi e duraturi di knowledge commons, basati su pratiche di collaborazione diffusa, revisione tra pari e produzione di sapere accessibile senza barriere economiche o tecnologiche. Questi progetti dimostrano come la costruzione della conoscenza — pur sostenuta da tecnologie digitali — resti un processo fondamentalmente umano, radicato nel contributo volontario di migliaia di persone in tutto il mondo.
Cogliendo l'occasione dell'anniversario, in questo contributo discuto il legame tra l'ecosistema Wikimedia e le Digital Humanities (DH), mostrando come tali spazi collaborativi siano impiegati in molteplici ambiti delle DH e, reciprocamente, come le prospettive teoriche e metodologiche delle DH possano rafforzarne valore, qualità e sostenibilità.
Chiarisco fin d'ora che quanto qui propongo è soltanto uno dei molti possibili accessi al vasto ecosistema Wikimedia. Ho privilegiato esempi direttamente osservati o conosciuti attraverso collaborazioni negli anni, senza alcuna pretesa di esaustività e senza voler sminuire il valore degli innumerevoli contributi, studi e ricerche di eccellenza rimasti esclusi.

2. ORIGINI E DIMENSIONE PARTECIPATIVA DEI PROGETTI WIKIMEDIA

Wikipedia nasce nel 2001 come evoluzione di Nupedia, progetto enciclopedico fondato su un modello di revisione esperta e centralizzata che si rivelò poco sostenibile in termini di produzione di contenuti, dato che in circa tre anni di attività riuscì a pubblicare soltanto 24 voci completate (Sanger, 2005). L'apertura di Wikipedia segnò una svolta radicale: nei primi mesi del 2001 furono create alcune centinaia di voci, che divennero decine di migliaia nel giro di un anno, inaugurando un modello di produzione della conoscenza fondato sulla partecipazione aperta e sulla collaborazione distribuita. A distanza di 25 anni, Wikipedia conta oltre 66 milioni di voci a livello mondiale, distribuite in più di 350 lingue (Wikipedia: Statistics, n.d.).
Nel tempo, a Wikipedia si è affiancato un insieme articolato di progetti complementari — tra cui Wikidata, Wikimedia Commons, Wikisource, ecc. — che gestiscono differenti tipi di contenuti e dati e costituiscono un ecosistema integrato di sapere aperto e interconnesso. Wikidata, avviato nel 2012, è oggi il più grande database semantico aperto esistente, con oltre 120 milioni di elementi strutturati, mentre Wikimedia Commons ospita più di 130 milioni di file multimediali liberamente riutilizzabili. Queste risorse risiedono sui server della Wikimedia Foundation e sono mantenute attraverso il lavoro di circa 300.000 volontari, singoli o organizzati in comunità locali o tematiche, distribuiti a livello globale.[2]
Dal punto di vista delle DH, tali progetti non rappresentano soltanto archivi digitali, ma pratiche sociali e infrastrutture operative di crowdsourcing, fondate su meccanismi di consenso comunitario, revisione tra pari e cura collettiva delle risorse. È questa dimensione partecipativa che rende i progetti Wikimedia particolarmente affini alle DH, intese non come semplice applicazione di strumenti digitali alle discipline umanistiche, ma come ambito critico di riflessione sulle trasformazioni epistemiche, sociali e politiche introdotte dal digitale (Reagle, 2010; Fiormonte, 2017). In questa prospettiva, Wikimedia può essere letto come uno dei più estesi esperimenti di citizen humanities e come un laboratorio in cui la sostenibilità del sapere è continuamente negoziata tra infrastruttura tecnica e responsabilità comunitaria.

3. WIKIMEDIA E DIGITAL HUMANITIES: UNA RELAZIONE BIDIREZIONALE

Le DH, come campo di studi che integra tecnologie digitali e discipline umanistiche, promuovono approcci collaborativi, transdisciplinari e orientati all'apertura nella ricerca, nell'insegnamento e nella divulgazione. In questa prospettiva, esse non si esauriscono nell'adozione di strumenti digitali, ma includono una riflessione critica sulle trasformazioni introdotte dal digitale nelle pratiche epistemiche e nelle forme di

[2] Oltre 260000 (https://www.wikimedia.org/ (cons. 30/01/26)). Gli editor attivi globali, cioè coloro che ogni mese effettuano almeno 5 modifiche, sono circa 300.000 (Wikimedia: Statistics, n.d.).

circolazione della conoscenza ("Digital humanities", 2026; Reagle, 2010; Fiormonte, 2017).
Coerentemente, i progetti Wikimedia realizzano pratiche aperte di produzione collettiva di contenuti e dati, in cui infrastrutture tecniche e dinamiche sociali si intrecciano per consentire a pubblici eterogenei di partecipare attivamente alla costruzione del sapere.
Questa visione condivisa di "sapere come processo" configura una relazione bidirezionale tra Wikimedia e DH: da un lato, Wikipedia, Wikidata e gli altri progetti offrono risorse, piattaforme e comunità che ampliano le possibilità di ricerca, cooperazione scientifica e public engagement; dall'altro, le DH forniscono cornici teoriche e metodologiche per interpretare criticamente tali pratiche collaborative, potenziandone l'impatto culturale e sociale e rendendone leggibili i presupposti epistemologici. La letteratura recente mostra, ad esempio, come Wikidata sia impiegato non soltanto come fonte per l'arricchimento e la modellazione semantica in progetti umanistici, ma anche come ambiente di pubblicazione e interrogazione di conoscenze strutturate in dialogo con comunità di contributori (Zhao, 2023). In questa prospettiva, Wikimedia non è un semplice deposito di contenuti, bensì uno spazio di co-produzione e negoziazione del sapere: dalla curatela dei dati alla definizione di metadati, fino all'uso di strumenti semantici e Linked Open Data per connettere discipline, fonti e comunità diverse. Tale bidirezionalità si riflette anche nelle comunità di pratica delle DH, che attraverso piattaforme collaborative favoriscono la condivisione di risorse, casi d'uso e competenze tra studiosi, rafforzando un modello di conoscenza aperto e cooperativo.[3]

3.1 DIDATTICA, CITIZEN SCIENCE E CO-COSTRUZIONE DEL SAPERE

L'utilizzo dei progetti Wikimedia nei contesti educativi costituisce una delle modalità più consolidate di integrazione tra DH e pratiche di public engagement. Wikipedia e Wikidata sono sempre più impiegati come ambienti di apprendimento attivo, in cui studenti e studentesse partecipano direttamente ai processi di produzione di contenuti, analisi critica delle fonti e modellazione della conoscenza.
In Italia, i progetti didattici in ambito universitario sono consultabili all'interno del Progetto Coordinamento/Università,[4] che favorisce l'uso strutturato di Wikipedia all'interno dei corsi, promuovendo competenze di scrittura collaborativa, valutazione delle fonti e cittadinanza digitale attraverso l'interazione con la comunità wikipediana. Analogamente, il Progetto Coordinamento/Scuole[5] supporta l'inserimento di attività basate su Wikipedia nella didattica scolastica, con l'obiettivo di sviluppare pensiero critico e consapevolezza dei meccanismi di produzione dell'informazione.
Queste pratiche, lette attraverso la lente delle DH, sono pienamente compatibili con i principi della citizen science: la produzione di contenuti e dati si configura come attività condivisa, verificabile e pubblica, in cui la trasparenza del processo (versioning, discussioni, policy) è parte integrante della qualità del risultato.
La ricerca recente evidenzia inoltre il potenziale di Wikidata come piattaforma didattica per l'esplorazione del Semantic Web: studi empirici mostrano come l'uso di dati strutturati favorisca forme di apprendimento basate sull'analisi e la modellazione della conoscenza (Evenstein Sigalov & Nachmias, 2023). In questo quadro si collocano numerose esperienze accademiche che integrano attività sui progetti Wikimedia nei percorsi formativi[6] e sperimentazioni nella scuola secondaria documentate dalla letteratura, che mostrano come l'uso di piattaforme wiki possa rafforzare la partecipazione e la collaborazione tra studenti (Fulton, 2019; Gallo & Petrucco, 2021; Monella, 2020, Capitolo 11; Monella et al., 2025, con bibl. precedente).

3.2 PATRIMONIO CULTURALE, GLAM E PARTECIPAZIONE PUBBLICA

Nel settore della valorizzazione del patrimonio culturale, i progetti Wikimedia svolgono un ruolo sempre più rilevante come infrastrutture di mediazione tra ricerca specialistica e pubblico. Attraverso collaborazioni strutturate tra comunità di volontari e istituzioni culturali — musei, archivi, biblioteche — Wikipedia e i progetti Wikimedia si configurano come spazi di mediazione tra ricerca esperta e pubblico più vasto, ampliando la portata sociale delle pratiche proprie delle DH: ad esempio, in collaborazione con

[3] Si vedano, ad esempio: Digital Classicist (https://wiki.digitalclassicist.org/Main_Page) e il Gruppo Wikidata per Musei, Archivi e Biblioteche (https://www.wikidata.org/wiki/Wikidata:Gruppo_Wikidata_per_Musei,_Archivi_e_Biblioteche)
[4] https://it.wikipedia.org/wiki/Progetto:Coordinamento/Università.
[5] https://it.wikipedia.org/wiki/Progetto:Coordinamento/Scuole.
[6] Due esempi tra i molti: il Master in Digital Humanities dell'Università di Milano https://it.wikipedia.org/wiki/Progetto:Coordinamento/Universit%C3%A0/UniMi/Master_Digital_Humanities_2025 e i laboratori di traduzione dell'Università di Verona.

il settore culturale nel 2019-2020 i contenuti culturali rilasciati sui progetti Wikimedia hanno raccolto oltre 1,2 miliardi di visualizzazioni, segno della significativa diffusione dei dati specialistici verso pubblici più vasti.[7]
In questo contesto si collocano le iniziative riunite sotto il nome di Progetto GLAM (Galleries, Libraries, Archives and Museums), che promuove la collaborazione tra il mondo Wikimedia e le istituzioni culturali per la condivisione di contenuti, dati e competenze.[8] In Italia, i progetti GLAM hanno dato vita a esperienze significative di documentazione partecipata e disseminazione del patrimonio, come nel caso della collaborazione con il Museo Egizio di Torino, che ha portato alla produzione e al miglioramento di voci enciclopediche, al rilascio di immagini su Wikimedia Commons e all'integrazione di dati strutturati su Wikidata, favorendo una maggiore visibilità internazionale delle collezioni.[9]
Numerose altre esperienze internazionali analoghe, tra cui quella dello Smithsonian Institution, confermano il valore delle piattaforme Wikimedia come infrastrutture per conservazione digitale e riuso di dati e immagini a supporto di ricerca, didattica e fruizione pubblica.[10]
Dal punto di vista delle DH, tali pratiche sono un esempio operativo di co-creazione e di "mappatura partecipata" delle collezioni: i dati vengono descritti, georeferenziati, collegati e resi interrogabili in forma aperta, con un impatto diretto su ricerca, didattica e fruizione pubblica.
Diversi studi hanno evidenziato come la presenza e la qualità delle informazioni culturali su Wikipedia producano effetti positivi misurabili sulla fruizione del patrimonio e sul turismo culturale. Analisi economiche mostrano che Wikipedia incide significativamente sulla visibilità e sull'interesse verso luoghi e istituzioni culturali, confermandone il ruolo come infrastruttura informativa di riferimento (Hinnosaar et al., 2021). Dal punto di vista delle DH, ricerche empiriche hanno messo in luce come iniziative collaborative basate su Wikipedia contribuiscano ad amplificare l'outreach scientifico e a rafforzare il rapporto tra ricerca, territorio e pubblico, in linea con i principi delle citizen humanities (Grossi & Plata, 2024).

3.3 DIVERSITÀ LINGUISTICA E INCLUSIONE

La pluralità delle edizioni linguistiche di Wikipedia e degli altri progetti Wikimedia costituisce un ambito chiave per l'analisi della diversità culturale e linguistica nel dominio digitale, poiché consente la documentazione di lingue minoritarie, dialetti regionali e lingue antiche attraverso comunità di parlanti e contributori. Wikipedia può essere considerata un laboratorio di produzione culturale in cui la distribuzione dei contenuti riflette dinamiche sociali, culturali e geografiche del sapere digitale (Miquel-Ribé & Laniado, 2019; Taylor et al., 2025). In questa prospettiva, progetti quali Wikipedia in latino, in veneto o in esperanto[11] mostrano come la piattaforma possa sostenere, con modalità differenti, sia processi di patrimonializzazione di lingue storiche, sia pratiche di riconoscimento e visibilità di varietà regionali marginalizzate nello spazio pubblico digitale.
La letteratura evidenzia però l'esistenza di gap di contenuto tra le edizioni linguistiche, legati a disparità nelle risorse delle comunità di editori (Wikipedia Diversity Observatory, n.d.; Miquel-Ribé & Laniado, 2020) e a strutture sociali che influenzano quali prospettive e quali gruppi culturali trovino rappresentazione nell'ecosistema digitale (Galina et al., 2024).
Dal punto di vista metodologico, l'analisi comparativa dei contenuti multilingui e dei pattern di partecipazione su Wikipedia consente di comprendere come le comunità linguistiche plasmino il sapere digitale e di misurare i gap tra edizioni, anche attraverso strumenti come topic modeling cross-linguistico e dataset dedicati alla diversità culturale. L'integrazione di metodi propri delle DH — analisi di corpora, visualizzazione e mappatura delle reti di contributori— permette così non solo di descrivere, ma anche di intervenire criticamente sui processi di costruzione della conoscenza nei progetti Wikimedia. In questa prospettiva, edizioni minoritarie, dialetti e lingue antiche emergono come un terreno di co-costruzione

[7] *Wikimedia UK,* n.d.*;* Wikimedia Foundation, 2022.
[8] https://it.wikipedia.org/wiki/Progetto:GLAM/Progetti
[9] https://it.wikipedia.org/wiki/Progetto:GLAM/Museo_Egizio_Torino. Per i progetti GLAM italiani si vedano: Progetto:GLAM/Progetti (nota 8) e Wikidata: Gruppo Wikidata per Musei, Archivi e Biblioteche (nota 3). Ivi, nella sessione "Pubblicazioni" sono raccolti studi metodologici e analitici sul tema.
[10] Smithsonian Institution: https://en.wikipedia.org/wiki/Wikipedia:GLAM/Smithsonian_Institution, https://commons.wikimedia.org/wiki/Category:Smithsonian_Institution
[11] https://it.wikipedia.org/wiki/Wikipedia_in_latino; https://vec.wikipedia.org/wiki/Wikipedia:Prinsipio; https://eo.wikipedia.org/wiki/Vikipedio:%C4%88efpa%C4%9Do

epistemica tra comunità, tecnologie aperte e pratiche critiche di produzione del sapere (Miquel-Ribé & Laniado, 2019; Miquel-Ribé & Laniado, 2020; Wikipedia Diversity Observatory, n.d.).

3.4 DATI APERTI, FAIR E WIKIDATA

Wikidata è uno dei principali depositi di dati strutturati liberamente accessibili per ricerca, analisi e integrazione con altri dataset aperti.[12] Con oltre 120 milioni di elementi collegati a concetti, entità, luoghi e relazioni, Wikidata non solo alimenta Wikipedia, ma agisce come nodo fondamentale nel web dei Linked Open Data: consente query tramite il linguaggio SPARQL e supporta l'arricchimento semantico di collezioni culturali e dataset umanistici.[13] In questa prospettiva, Wikidata si colloca nel quadro dell'Open Access e dei principi FAIR, poiché rende i dati accessibili e riusabili nel tempo attraverso pratiche condivise di descrizione e collegamento, estendendosi oggi anche a contesti di AI verificabile (Wikidata Embedding Project).[14]
L'adozione di Wikidata nelle DH ha reso praticabili nuove domande metodologiche e analitiche, soprattutto in relazione alla modellazione e all'esplorazione di eventi storici e alle dimensioni temporali delle collezioni digitali.[15] Studi recenti mostrano come la visualizzazione di dati cronologici e relazionali migliori l'accessibilità e la comprensione di fenomeni culturali complessi (Sugimoto et al., 2025). Tuttavia, la stessa apertura che abilita tali possibilità rende visibili anche squilibri: analisi quantitative sul gender gap, ad esempio nel caso degli scrittori francesi, evidenziano la necessità di approcci critici e partecipativi per rafforzare equità e inclusività delle basi di dati (Conroy, 2023).
Alla luce di questi sviluppi, le DH non si limitano più a utilizzare dati da Wikidata, ma promuovono pratiche attive di cura, arricchimento e pubblicazione, coerenti con una concezione del dato come costruzione culturale e non mero oggetto tecnico. L'integrazione di Wikidata nei progetti umanistici consente di esplorare prospettive di ricerca e di diffondere conoscenza attraverso iniziative collettive (Zhao, 2023). In questo quadro, Wikidata non è solo un repository, ma spazio di produzione condivisa, in cui le opportunità metodologiche richiedono una riflessione su bias, rappresentazione e partecipazione comunitaria. [16]

4. ALCUNE RIFLESSIONI CONCLUSIVE

Wikipedia si è evoluta in 25 anni nella più grande enciclopedia libera, con oltre 66 milioni di articoli in più di 350 lingue e circa 15 miliardi di visualizzazioni mensili, adattandosi a profondi cambiamenti tecnologici. In un ambito caratterizzato da cicli di innovazione estremamente rapidi, si può affermare che abbia attraversato e dominato un'intera era digitale. Ci si interroga quindi su quali fattori abbiano consentito a Wikipedia e ai progetti Wikimedia di durare così a lungo e di crescere con una scala e una velocità difficilmente prevedibili 25 anni fa. Le risposte sono molteplici, ma dalle riflessioni sviluppate in questo contributo emerge un elemento centrale: la dimensione umana e comunitaria e un processo di co-creazione del sapere fondato su negoziazione, trasparenza e responsabilità condivisa.
È proprio questa infrastruttura sociale che consente di interpretare l'ecosistema Wikimedia come il più ampio e duraturo progetto di DH oggi esistente: un'infrastruttura socio-tecnica globale in cui tecnologie digitali, pratiche collaborative e produzione culturale si intrecciano quotidianamente attraverso il contributo umano. La relazione tra Wikimedia e DH è profondamente bidirezionale e trova il suo fondamento comune nell'"human being": il soggetto attivo che partecipa alla costruzione, negoziazione e manutenzione del sapere. Wikipedia, Wikidata e i progetti fratelli rendono visibili, su scala senza precedenti, pratiche di scrittura collettiva, curatela dei dati e condivisione aperta che avvicinano ricerca, didattica e public engagement, configurandosi come vere e proprie infrastrutture di citizen humanities.
Al compimento dei 25 anni, tuttavia, l'ecosistema Wikimedia si confronta con sfide rilevanti: l'invecchiamento delle comunità, il calo del turnover[17] e la pressione esercitata dall'AI e dai modelli di consumo automatizzato della conoscenza. Si tratta di una sfida non solo tecnologica, ma culturale ed

[12] https://en.wikipedia.org/wiki/Wikidata
[13] https://www.wikidata.org/wiki/Wikidata:Main_Page; https://query.wikidata.org/
[14] Wikimedia Deutschland, 2025; https://www.wikidata.org/wiki/Wikidata:Embedding_Project.
[15] Per l'ambito italiano: le attività e pubblicazioni del Gruppo Wikidata per Musei Archivi e Biblioteche (nota 3 e 9).
[16] Numerosi studi evidenziano le criticità e le potenzialità di Wikidata. Una selezione, con bibliografia precedente: Farda-Sarbas & Müller-Birn, 2019; Heftberger & Duchesne, 2020; Baker & Mahal, 2024.
[17] È attestato un calo di contributori del 36% nel decennio 2016-2025 *(Wikinotizie,* 2026; *Wikimedia Foundation.* 2025).

epistemica.[18] In questo quadro, le DH risultano strategiche: offrono strumenti teorici e metodologici per governare l'evoluzione senza smarrire i principi fondanti di apertura, equità e sostenibilità, bilanciando il conservatorismo necessario per preservare libertà e trasparenza e l'innovazione imprescindibile per adattarsi ai cambiamenti tecnologici. Rafforzare il legame tra Wikimedia e DH significa, dunque, investire in modelli di conoscenza aperti, critici e partecipati, in cui l'innovazione tecnologica rimane al servizio dell'umano e la dimensione collettiva del sapere continua a rappresentarne il principale punto di forza.

RINGRAZIAMENTI

Ringrazio Wikimedia Italia per il supporto alle attività di ricerca e sperimentazione. Ringrazio inoltre Ferdinando Traversa e Paolo Cavallini per la rilettura del testo e per l'utile confronto. Resta naturalmente a mio carico ogni responsabilità sui contenuti.

BIBLIOGRAFIA

Baker, J., & Mahal, A. (2024). "I have always found the whole area a minefield": Wikidata, historical lives, and knowledge infrastructure. *International Journal of Digital Humanities*, *6*, 217–236. https://doi.org/10.1007/s42803-024-00090-5

Conroy, M. (2023). Quantifying the gap: The gender gap in French writers' Wikidata. *Journal of Cultural Analytics*, *8*(2). https://doi.org/10.22148/001c.74068

Digital humanities. (2026, 31 gennaio). In *Wikipedia*. https://en.wikipedia.org/wiki/Digital_humanities

Evenstein Sigalov, S., & Nachmias, R. (2023). Investigating the potential of the semantic web for education: Exploring Wikidata as a learning platform. *Education and Information Technologies, 28*(10), 12565–12614. https://doi.org/10.1007/s10639-023-11664-1

Farda-Sarbas, M., & Müller-Birn, C. (2019). *Wikidata from a research perspective: A systematic mapping study of Wikidata*. arXiv. https://doi.org/10.48550/arXiv.1908.11153

Fiormonte, D. (2017). Digital Humanities and the Geopolitics of Knowledge. *Digital Studies/Le champ numérique*, *7*(1), 1–18. https://doi.org/10.16995/dscn.274

Fulton, C. (2019). The use of collaborative open-access publishing via Wikipedia in university education to embed digital citizenship skills. *Netcom. Réseaux, communication et territoires*, *33*(1/2). https://doi.org/10.4000/netcom.3893

Galina, P., Saisó, A., & Tomlinson, J. (2024). Towards language sensitivity and diversity in the digital humanities. *Digital Studies / Le Champ Numérique*, *14*(1). https://doi.org/10.16995/dscn.10291

Gallo, V., & Petrucco, C. (Eds.). (2021). *Wikipedia in Academia*. Padova University Press. https://www.padovauniversitypress.it/it/publications/9788869381980

Grossi, P., & Plata, S. (2024). Amplifying archaeological outreach: The impact of Wikipedia and collaborative initiatives on cultural tourism. *Archeologia e Calcolatori, 35*(2), 19–28. https://doi.org/10.19282/ac.35.2.2024.03

Heftberger, A., & Duchesne, P. (2020, giugno). *Cataloguing practices in the age of Linked Open Data: Wikidata and Wikibase for film archives*. FIAF. https://www.fiafnet.org/pages/E-Resources/Cataloguing-Practices-Linked-Open-Data.html

Hinnosaar, M., Hinnosaar, T., Kummer, M., & Slivko, O. (2021). Wikipedia matters. *Journal of Economics & Management Strategy, 32*(3), 1–13. https://doi.org/10.1111/jems.12421

Jemielniak, D. (2026, 30 gennaio). *At age 25, Wikipedia refuses to evolve: The digital commons champion faces a crisis of its own making*. IEEE Spectrum. https://spectrum.ieee.org/wikipedia-at-25

Miquel-Ribé, M., & Laniado, D. (2019). Wikipedia cultural diversity dataset: A complete cartography for 300 language editions. *Proceedings of the International AAAI Conference on Web and Social Media, 13*(01), 620–629. https://doi.org/10.1609/icwsm.v13i01.3260

Miquel-Ribé, M., & Laniado, D. (2020). The Wikipedia Diversity Observatory: A project to identify and bridge content gaps in Wikipedia. *Proceedings of the 16th International Symposium on Open Collaboration*, Article 14, 1–4. https://doi.org/10.1145/3412569.3412866

[18] Per una analisi critica di sintesi si veda: Jemielniak, 2026.

Monella, P. (2020). *Metodi digitali per l'insegnamento classico e umanistico*. EDUCatt. https://www.paolomonella.it/metodi.pdf

Monella, P., Fiorino, A., Saieva, F., & Sorci, A. (2025). Piattaforme wiki per l'insegnamento umanistico: Sperimentazioni in corso nel Liceo 'De Cosmi' di Palermo. *Umanistica Digitale*, *9*(20), 601–610. https://doi.org/10.6092/issn.2532-8816/21227

Reagle, J. (2010). *Good faith collaboration: The culture of Wikipedia*. MIT Press. https://reagle.org/joseph/2010/gfc/reagle-2010-good-faith-collaboration.pdf

Sanger, L. (2005). *The Early History of Nupedia and Wikipedia: A Memoir.* Slashdot. https://features.slashdot.org/story/05/04/18/164213/the-early-history-of-nupedia-and-wikipedia-a-memoir

Sugimoto, G., de Boer, V., & van Ossenbruggen, J. (2025). Wikidata visualization for event and temporal data exploration in digital humanities and cultural heritage. *Semantic Web*, *16*(1), 1–25. https://doi.org/10.3233/SW-243578

Taylor, M., Proven, R., & Areia, C. (2025). *Evaluating the diversity of scientific discourse on twenty-one multilingual Wikipedias using citation analysis*. arXiv. https://doi.org/10.48550/arXiv.2501.09666

Wikimedia Deutschland. (2025). *The challenge of generative AI: Opportunities and risks for free knowledge*. Wikimedia Deutschland. https://www.wikimedia.de/the-challenge-of-generative-ai/

Wikimedia Foundation. (2022). *GLAM-Wiki project: Supporting galleries, libraries, archives, and museums to produce open-access, freely reusable content*. Wikimedia Outreach. https://outreach.wikimedia.org/wiki/GLAM

Wikimedia Foundation. (2025). *Annual Plan/2025-2026: Strategic priorities for a shifting digital landscape*. Meta-Wiki. https://meta.wikimedia.org/wiki/Wikimedia_Foundation_Annual_Plan/2025-2026

Wikimedia: Statistics, (n.d.). In *Wikimedia Statistics*, Estratto il 30 gennaio 2026, da https://stats.wikimedia.org/#/all-projects

Wikimedia UK. (n.d.). In *Cultural partnerships.* https://wikimedia.org.uk/wiki/Cultural_partnerships

Wikinotizie. (2026, 18 gennaio). *Wikipedia festeggia 25 anni di sapere condiviso e libero, ma il declino è un rischio concreto*. https://it.wikinews.org/wiki/Wikipedia_festeggia_25_anni_di_sapere_condiviso_e_libero,_ma_il_declino_è_un_rischio_concreto

Wikipedia Diversity Observatory. (n.d.). *Main Page*. Meta-Wiki. Estratto il 30 gennaio 2026, da https://meta.wikimedia.org/wiki/Wikipedia_Diversity_Observatory

Wikipedia:Statistics. (n.d.). In *Wikipedia*. Estratto il 30 gennaio 2026, da https://en.wikipedia.org/wiki/Wikipedia:Statistics

Zhao, F. (2023). A systematic review of Wikidata in Digital Humanities projects. *Digital Scholarship in the Humanities*, *38*(2), 852–874. https://doi.org/10.1093/llc/fqac083

Una pausa e tre domande. La crisi della memoria nell'era dell'accesso

ENRICA SALVATORI[1], FEDERICO VALACCHI[2]

[1]UNIVERSITÀ DI PISA, AIPH – ENRICA.SALVATORI@UNIPI.IT;
[2]UNIVERSITÀ DI MACERATA, AIDUSA - FEDERICO.VALACCHI@UNIMC.IT

ABSTRACT (ITALIANO)

Il contributo propone una riflessione critica sulla crisi della memoria nell'era digitale, caratterizzata dal paradosso tra una produzione informativa senza precedenti e la crescente fragilità dei processi di conservazione e trasmissione. Gli autori mettono in discussione l'idea della digitalizzazione come risposta automatica ai problemi della tutela e dell'accesso alla memoria, evidenziandone invece la natura selettiva, interpretativa e profondamente politica. Attraverso tre domande chiave – perché digitalizziamo, chi decide cosa digitalizzare e conservare, come utilizziamo le fonti digitalizzate e digitali – il saggio analizza le trasformazioni in atto nelle pratiche archivistiche, storiche e di *Public History.* Particolare attenzione è dedicata alla tensione tra archivi istituzionali e archivi distribuiti e partecipativi, nonché agli effetti dell'intelligenza artificiale sull'uso, sulla mediazione e sull'interpretazione delle fonti storiche. Il contributo auspica una "pausa" di riflessione condivisa e interdisciplinare, capace di coinvolgere storici, archivisti, informatici e comunità, per orientare scelte più consapevoli, sostenibili e responsabili nella costruzione della memoria digitale contemporanea.
Parole chiave: memoria digitale; digitalizzazione; archivi; Digital Humanities; intelligenza artificiale

ABSTRACT (ENGLISH)

The paper offers a critical reflection on the crisis of memory in the digital age, characterized by the paradox between unprecedented information production and the growing fragility of preservation and transmission processes. The authors question the idea of digitization as an automatic solution to the challenges of safeguarding and accessing memory, emphasizing instead its selective, interpretative, and inherently political nature. Through three key questions—why we digitize, who decides what should be digitized and preserved, and how digitized and born-digital sources are used—the paper examines ongoing transformations in archival practices, historical research, and public history. Particular attention is given to the tension between institutional archives and distributed, participatory archives, as well as to the impact of artificial intelligence on the use, mediation, and interpretation of historical sources. The contribution calls for a shared and interdisciplinary "pause for reflection," involving historians, archivists, computer scientists, and communities, in order to foster more conscious, sustainable, and responsible approaches to the construction of contemporary digital memory.
Keywords: digital memory; digitization; archives; Digital Humanities; artificial intelligence

1. INTRODUZIONE

Stiamo vivendo un paradosso: produciamo memoria e informazione ai limiti dell'ipermnesia, ma dobbiamo fare concretamente i conti con i rischi di un potenziale oblio di noi stessi e con un incombente e delirio di smemoratezza. Ogni giorno, l'umanità produce 2,5 quintilioni di byte di dati[1]: fotografie, video, testi, scambi comunicativi. L'illusione meccanica che tutto questo possa sedimentarsi senza colpo ferire in un cloud scontato ed eterno si infrange ogni giorno contro la pratica di chi lavora con le fonti e la condivisione della memoria. L'era digitale rischia di essere ricordata dagli storici futuri come un'epoca oscura, la Digital Dark Age di Kurt Bollacker, un periodo di cui rimarranno pochissime tracce dirette (Bollacker, 2010). Prima ancora, però, rischiamo di diventare - se già non lo siamo - i protagonisti di un'epoca in cui si è guardato irrazionalmente alla digitalizzazione del supporto analogico come a una panacea per mali diversi: il deterioramento del materiale, la difficoltà di accesso alle fonti o ai beni

[1] Data Never Sleeps di Domo Inc. (2023). La cifra include dati generati da social media, transazioni digitali, sensori IoT e contenuti multimediali. https://www.domo.com/it/data-never-sleeps.

culturali in senso lato, la volatilità dell'informazione. Innumerevoli case studies entro le DH e le scienze documentarie si sono effettivamente dati obiettivi in queste direzioni e hanno cercato e stanno cercando rimedi più o meno "sistemici" per affrontare queste grandi problematiche. Sul versante archivistico un tentativo di approccio sistemico, per quanto probabilmente perfettibile è quello di Archivio Digitale, condotto da ICAR (https://archiviodigitale-icar.cultura.gov.it/).
La realtà in cui viviamo è però particolarmente complessa e articolata, caratterizzata da una crescente produzione quotidiana di documenti informatici, cui si aggiungono processi massicci di digitalizzazione di archivi analogici pregressi - spesso favoriti da finanziamenti pubblici - e la costruzione di archivi inventati partecipativi, tanto innovativi quanto effimeri e potenzialmente esposti al rischio di decontestualizzazioni e di produzione di *fake news* distorsive e fuori controllo (Valacchi, 2024).
Non si tratta solo di denunciare la fragilità del supporto digitale: anche una pietra, un vaso di ceramica, una pergamena e un libro di carta sono oggetti fragili, che necessitano di cura per essere conservati e consentir loro la trasmissione del contenuto che veicolano. Tuttavia, l'umanità ha dedicato nei secoli molte energie e risorse per creare un sistema adatto non solo alla loro conservazione, ma anche alla lettura, decodifica, trasmissione, interpretazione e condivisione di quanto trasmettevano. In sostanza, nel corso del tempo intellettuali, classi dirigenti e istituzioni hanno investito ingenti risorse etiche, tecniche e finanziarie nella conservazione, definendo progressivamente le politiche e le azioni della memoria in termini di selezione, conservazione e trasmissione di archivi, intesi come strutture di supporto al loro potere e di rappresentazione delle loro società. Il *digital turn*, la nascita del *cloud* e l'apparente facilità nella creazione e accumulo di fonti, metafonti e nuove fonti, ha fatto completamente saltare queste dinamiche, creando tuttavia nuovi e differenti "poteri". Riteniamo necessario che, su questa congiuntura epocale, il mondo delle DH e delle discipline documentarie si prenda una pausa di riflessione per porsi alcune cruciali domande.

2. LA PAUSA

Proprio in qualità di umanisti digitali e di professionisti dell'informazione che in questi anni hanno guardato con grande interesse all'onda lunga del cambiamento e contemporaneamente hanno esercitato una critica attenta e costante sui mutamenti che il *digital turn* apporta alla società, riteniamo che sulla questione della produzione degli archivi informatici contemporanei e della costruzione di fonti storiche digitalizzate e digitalizzabili si debba prendere tempo per riflettere su quanto sta accadendo. La pausa di riflessione si ritiene necessaria non tanto al fine di fornire "soluzioni", quanto per far crescere il livello di consapevolezza in tutte le realtà istituzionali o sociali impegnate nella preservazione della memoria. Il momento è adatto, a livello nazionale, europeo e mondiale, perché sono giunte a un discreto grado di maturazione le discussioni riguardanti i cambiamenti che il digitale ha apportato nel mondo dell'archivistica e della storia (con particolare riguardo alla *Public History*), intese qui non come discipline accademiche, ma come campi del sapere fondamentali e complementari, tese a garantire fonti autentiche e ordinate e a favorire la loro interpretazione critica da parte di diversi pubblici. Illustriamo quindi le domande che crediamo debbano essere poste in tavoli di discussione da aprire in diversi contesti e a diverse scale territoriali, sulla base di una griglia concettuale da costruire insieme a storici, archivisti, filosofi, antropologi, letterati, esperti di museologia e di biblioteche e informatici.

3. PERCHÉ DIGITALIZZIAMO?

Attualmente numerosi istituti di conservazione hanno promosso campagne di digitalizzazione dei propri fondi analogici, seguendo criteri e obiettivi non sempre chiari o poco chiari o, e non ponendosi, o ponendosi poco, il problema delle possibili ricadute realmente innovative della digitalizzazione, soprattutto in direzione di un potenziamento delle opportunità euristiche. Un primo criterio è stato quello conservativo, ossia della salvaguardia della materia deteriorabile, a prescindere dal contenuto veicolato. In sostanza si è accettata e sposata la scelta di rilevanza della fonte operata nel passato e si è guardato alla digitalizzazione come intervento necessario di salvaguardia, svincolando il contenuto informativo dal deterioramento del supporto materiale per garantirne la sopravvivenza. Tuttavia, sappiamo che la digitalizzazione non si riduce a un mero trasferimento lineare di informazione tra due supporti diversi, ma si sostanzia in un processo filologico e rappresentativo che si traduce in un complesso intervento interpretativo e ricostruttivo, diverso e specifico a seconda del medium digitalizzato e delle esigenze specifiche nel progetto culturale sotteso. Le conseguenze profonde di tale complessità non sono ancora valutate nel quadro di una digitalizzazione massiva e poco riflessiva che agisce spesso in maniera poco

consapevole (Barbuti, 2022). L'altro criterio imperante è stato quello dell'antichità e della presunta rilevanza storica e istituzionale: si è cominciato cioè a digitalizzare in maniera sistematica dai pezzi più antichi (esempio classico il fondo *Diplomatico* di molti archivi) e dai fondi che sono espressione di istituzioni valutate come "importanti". Anche in questo caso si è spesso fatto riferimento a criteri meccanici, presupponendo più o meno consapevolmente la continuità di una digitalizzazione integrale e sistematica di tutte le fonti che non potrà mai avvenire e che non ha alcun senso fare. Al fine di collocare la digitalizzazione selettiva degli archivi e degli oggetti informativi nel quadro di un effettivo potenziamento delle azioni di uso e interpretazione delle fonti è perciò urgente che ci si chieda perché digitalizziamo, sforzandoci poi di definire i destinatari delle azioni digitali e di individuare in maniera consequenziale su quali sistemi di fonti intervenire.

4. CHI DECIDE COSA DIGITALIZZARE E CONSERVARE?

La digitalizzazione può determinare ricontestualizzazioni e crea di fatto nuove fonti (metafonti) che tramite il Web vengono redistribuite in contesti di fruizione potenzialmente diversi, tanto incontrollabili quanto effimeri. In questo scenario di fondo registriamo oggi la presenza di due grandi macrocategorie di archivio: quello tradizionale, espressione del potere e di una lettura istituzionale del passato, e quello distribuito, collettivo e più o meno spontaneo. Il primo è in forte crisi, ed è alle prese con le radicali trasformazioni delle modalità di produzione, gestione, conservazione e trasmissione che ne riconfigurano posizione e ruolo, ma è ancora sostenuto da un passato solido e da una percezione diffusa di un ineludibile grado di utilità. Il secondo è in continua e spasmodica fioritura, ma manca molto spesso di competenze, metodo e procedure di salvaguardia e disseminazione. Anche gli archivi spontanei e distribuiti sono ovviamente espressione di un potere, ma di carattere e taglia diversi a seconda dei contesti, perché espressi dalla società del presente in costante cambiamento ed evoluzione dinamica. La prima tipologia ha bisogno di chiedersi se la scelta su cosa digitalizzare possa essere fatta anche insieme ai diversi pubblici e alle comunità da loro servite[2]; i secondi hanno estremamente bisogno di infrastrutture, scelte procedurali e competenze che trasformino le loro collezioni disordinate ed estemporanee in archivi veri e propri, dove i documenti sono organizzati, conservati e resi effettivamente consultabili[3]. In entrambi i casi storici/storiche e archivisti/archiviste devono collaborare strettamente per dare ordine, senso e durata alla memoria documentata. Il ciclo della conservazione/fruizione digitale è legato in maniera inestricabile non solo all'accessibilità e alla diffusione delle risorse, ma all'interfaccia di utilizzo e alle pratiche che si attivano per un godimento consapevole. La transizione dal supporto materiale al dato numerico sta ridefinendo la natura stessa delle istituzioni culturali: musei e archivi di ogni ordine e grado evolvono da luoghi di custodia esclusiva a nodi nevralgici di una rete di distribuzione globale e di costruzione di nuove narrazioni (Ravveduto & Salvatori 2023; Salvatori 2025). Dobbiamo discutere con le comunità e gli enti sull'assunzione di responsabilità riguardo alla scelta di cosa tenere e redistribuire.

5. COME PENSIAMO DI UTILIZZARE LE FONTI DIGITALIZZATE E DIGITALI?

Le DH in questi anni hanno fatto passi giganteschi nell'elaborazione di standard per l'archiviazione e la metadatazione delle fonti e nella messa a punto di sistemi di connessione semantica tra gli oggetti. Al tempo stesso gli standard archivistici, come dimostra *Record in Contexts*, hanno recepito l'esigenza di nuove modalità di descrizione e rappresentazione del materiale documentario nel suo insieme[4]. Buone pratiche e standard di metadatazione, tuttavia, sia dal punto di vista archivistico che storico, per quanto puntuali e quando applicati, sono necessari ma non sufficienti a rendere le fonti (analogiche, digitalizzate e native digitali) un autentico patrimonio comune. Se la scelta di cosa, come e perché digitalizzare deve passare attraverso dei tavoli di discussione a più voci, allo stesso modo deve essere favorita la consapevolezza sul loro possibile uso. Sappiamo che attualmente i sistemi di AI generativa interrogati

[2] Un esempio interessante a questo proposito è quello offerto dall'Archivio comunale di Vicopisano nel progetto Archivico Digitale, dove alla scelta concordata con la Soprintendenza archivistica di digitalizzare due pezzi a rischio dopo averli restaurati, si è deciso di digitalizzare fonti ritenute potenzialmente interessanti per la comunità locale, creando contemporaneamente percorsi tematici di fruizione e accesso all'inventario generale (https://archivicodigitale.labcd.unipi.it/).
[3] Altro esempio emblematico l'Archivio degli Iblei, un "archivio digitale partecipato" e progetto di Public History di indubbio successo, che tuttavia non adotta standard di archiviazione (https://www.archiviodegliiblei.it/).
[4] Records in Contexts (RiC) è uno standard ICA per descrivere i documenti che favorisce a un approccio più contestualizzato e dinamico: https://www.ica.org/ica-network/expert-groups/egad/records-in-contexts-ric/.

sulle questioni storiografiche intercettano e usano solo parzialmente le fonti primarie e secondarie: attingono invece a piene mani a una letteratura "terziaria" (Wikipedia, siti web) e alla versione digitale della letteratura scientifica realmente utilizzabile (De Ninno & Lacriola, 2025). L'uso dell'AI nel passaggio dalla fonte al dato e dal dato all'informazione è ancora appannaggio di pochissimi. Si sta replicando quindi sul piano dell'uso delle fonti la creazione di un sistema di potere in cui l'interpretazione "scientifica" - finché sarà ritenuta utile - è prerogativa di pochi, mentre dilaga una restituzione popolare che fatalmente attinge a visioni e linee storiografiche superate, a fonti selezionate a priori, oltre che facilmente manipolabili. È tempo che sulla questione si faccia un ragionamento più strutturato, di taglio non solo storico e archivistico, ma anche etico, filosofico e persino antropologico, per costruire intorno alla digitalizzazione del passato e alla conservazione delle fonti del presente un contesto solido, propedeutico alla co-progettazione di soluzioni affidabili e condivise. Sembra necessario, in definitiva, un approccio realmente cross domain alla costruzione delle diverse risorse digitali e digitalizzate, un approccio che muova da una riflessione condivisa a partire non tanto dall'opportunità della digitalizzazione, quanto dal ruolo che essa può avere nel potenziare le strategie euristiche e nell'orientare un uso consapevole, selettivo e critico della ipermnesia contemporanea.

BIBLIOGRAFIA

Barbuti, N. (2022). *La digitalizzazione dei beni documentali. Metodi, tecniche, buone prassi* (Vol. 38). Editrice bibliografica.

Bollacker, K. D. (2010). Computing Science: Avoiding a Digital Dark Age. *American Scientist*, *98*(2), 106–110. https://www.jstor.org/stable/27859471

De Ninno, F., & Lacriola, M. (2025). Mussolini and ChatGPT. Examining the Risks of A.I. writing Historical Narratives on Fascism. *Journal of Modern Italian Studies*, *30*(2), 187–209. https://doi.org/10.1080/1354571X.2025.2457250

Ravveduto, M., & Salvatori, E. (2023). Storia digitale e digital public history: Le novità di un antico mestiere. In F. Ciotti (A c. Di), *Digital Humanities: Metodi, strumenti, saperi* (pp. 229–254). Carocci.

Salvatori, E. (2025). Per uno statuto scientifico della Public History. *Ricerche Storiche*, *LV* (1), 127–150.

Valacchi, Federico. (2024). *L'archivio aumentato. Tempi e modi di una digitalizzazione critica*. Editrice Bibliografica.

[2/C] Modelli digitali per gestire e condividere fonti multidisciplinari

Digital Repatriation in Practice: Sharing Ethnographic Photography with Communities of Origin at the Náprstek Museum

DANIELA ŠNAPKOVÁ
CENTRE FOR IBERO-AMERICAN STUDIES, FACULTY OF ARTS, CHARLES UNIVERSITY

ABSTRACT (ENGLISH)

This paper presents a case study of digital repatriation of ethnographic photographs from the photographic collection of Václav Šolc held at the Náprstek Museum. Drawing on contemporary approaches to ethnographic photography and participatory museology, the study examines digital repatriation not simply as a process of digitisation and online access, but as a relational and ethically situated practice shaped through engagement with communities of origin. The paper focuses on photographs produced among Aymara, Mapuche, and Kawésqar communities in Chile during the 1960s and 1970s, and critically reflects on the capacities and limitations of digital tools in reconnecting museum collections with communities of origin.
Methodologically, the research combines archival and collection-based analysis with ethnographic fieldwork. It involves the study of museum documentation and online databases alongside in-person encounters, interviews, and outreach via online platforms which functioned as part of the digital infrastructure of repatriation and public engagement. Community-based identification of individuals portrayed in the photographs enabled the correction and enrichment of museum metadata through lived memory and intergenerational knowledge.
The findings demonstrate that while digitisation is a crucial prerequisite for access, it does not in itself constitute meaningful digital repatriation. Structural barriers including linguistic limitations, technical complexity, institutional restrictions on image sharing, and geographic isolation significantly shape community engagement. The paper argues that the value of digitised ethnographic photography lies in its present-day relational use, and that digital repatriation requires sustained human mediation, ethical reflexivity, and collaborative institutional practices beyond technological solutions alone.
Keywords: ethnographic photography; digital repatriation; communities of origin

ABSTRACT (ITALIANO)

Il rimpatrio digitale nella pratica: condivisione delle fotografie etnografiche con le comunità di origine al Museo Náprstek
Questo articolo presenta un caso di studio sul rimpatrio digitale di fotografie etnografiche provenienti dalla collezione fotografica di Václav Šolc conservata al Museo Náprstek. Attingendo agli approcci contemporanei alla fotografia etnografica e alla museologia partecipativa, lo studio esamina il rimpatrio digitale non semplicemente come un processo di digitalizzazione e accesso online, ma come una pratica relazionale ed eticamente situata, plasmata dal coinvolgimento delle comunità di origine. Il documento si concentra sulle fotografie prodotte tra le comunità Aymara, Mapuche e Kawésqar in Cile durante gli anni '60 e '70 e riflette criticamente sulle capacità e i limiti degli strumenti digitali nel ricollegare le collezioni museali alle comunità di origine.
Dal punto di vista metodologico, la ricerca combina l'analisi basata su archivi e collezioni con il lavoro etnografico sul campo e la comunicazione mediata dal digitale. Comprende lo studio della documentazione museale e delle banche dati online, insieme a incontri di persona, interviste e attività di sensibilizzazione tramite piattaforme online che hanno funzionato come parte dell'infrastruttura digitale di rimpatrio e coinvolgimento del pubblico. L'identificazione da parte della comunità delle persone ritratte nelle fotografie ha permesso la correzione e l'arricchimento dei metadati museali attraverso la memoria vissuta e la conoscenza intergenerazionale.
I risultati dimostrano che, sebbene la digitalizzazione sia un prerequisito fondamentale per l'accesso, di per sé non costituisce un rimpatrio digitale significativo. Le barriere strutturali, tra cui i limiti linguistici, la complessità tecnica, le restrizioni istituzionali sulla condivisione delle immagini e l'isolamento geografico, influenzano in modo significativo il coinvolgimento della comunità. Il documento sostiene che il valore

della fotografia etnografica digitalizzata risiede nel suo uso relazionale attuale e che il rimpatrio digitale richiede una mediazione umana costante, una riflessività etica e pratiche istituzionali collaborative che vanno oltre le sole soluzioni tecnologiche.
Parole chiave: fotografia etnografica; rimpatrio digitale; comunità di origine

PROPOSAL

This proposal presents a case study of digital repatriation of ethnographic photographs from the photographic collection of Václav Šolc held at the Náprstek Museum (part of the National Museum of the Czech Republic), examined through the lenses of contemporary approaches to ethnographic photography, participatory museology, and digital public engagement. Focusing on the digitisation, online accessibility, and community sharing of photographs taken among indigenous communities in Chile, the paper explores digital repatriation not only as a technical process, but as a relational, ethically charged practice that unfolds through direct engagement with communities of origin. The contribution addresses the possibilities and limitations of digital tools in reconnecting museum collections with source communities, while critically reflecting on the roles of curators and researchers as mediators between institutions, digital infrastructures, and indigenous communities.
In recent years, the digitisation of ethnographic collections has become one of the central concerns within museum studies and visual anthropology, raising questions about access, ownership, ethics, and public engagement. While digitisation has often been framed as a means of democratising access to cultural heritage, digital availability alone does not constitute meaningful repatriation or participation (Kjellman, 2025: 283). The concept of digital repatriation has emerged in response to these issues, highlighting the importance of collaboration with communities of origin, contextualisation of visual materials, and recognition of photographs as living heritage with ongoing social lives (Edwards, 2015: 246).
Ethnographic photography, in particular, poses specific challenges due to its historical entanglement with colonial power relations, uneven authorship, and frequent absence of names or community-based knowledge. Within this context, participatory museology calls for approaches that move beyond representation toward
co-creation, shared authority, and sustained relationships between institutions and communities.
This paper focuses on Šolc's photographic collections created among indigenous communities in Chile. The primary goals of this research are: (1) to analyse Václav Šolc's ethnographic photographic collections from Aymara and Kawésqar communities in Chile in light of contemporary approaches to ethnographic photography; (2) to examine how digitised museum photographs are perceived, interpreted, and valued by communities of origin; (3) to identify individuals portrayed in the photographs and enrich existing museum metadata through community knowledge; and (4) to critically assess the practical, ethical, and infrastructural challenges of digital repatriation and public engagement from the position of a researcher working in collaboration with, but not institutionally embedded within, a museum. By doing so, the research seeks to contribute to broader discussions on digital access, participatory practices, and the responsibilities of cultural heritage institutions in the digital age.
Methodologically, this research combines archival and collection-based research with ethnographic fieldwork and digital engagement practices. The study involved systematic analysis of inventory documentation, archival materials, and the Náprstek Museum's online ethnographic photo-collection database, alongside critical evaluation of its accessibility, language barriers, and usability for non-specialist audiences. Fieldwork methods included in-person meetings, semi-structured interviews, and informal conversations with members of communities of origin, as well as the use of digital platforms such as Facebook and WhatsApp as tools for outreach, communication, and image sharing. Community-based identification of individuals portrayed in the photographs constituted a key methodological component, allowing the enrichment and correction of museum metadata through lived memory and intergenerational knowledge.
Václav Šolc (1919–1995) was an ethnographer working for the Náprstek Museum in Prague, the only museum in the Czech Republic (and formerly Czechoslovakia) dedicated to non-European cultural heritage. Unlike many European ethnographic museums whose collections were formed within explicitly colonial contexts, the Náprstek Museum developed in a country without overseas colonies. Nevertheless, the museum's collections emerged within global asymmetries of power, mobility, and knowledge production, and later became shaped by Global Cold War geopolitics. Until the mid-20th century, the museum was not formally connected to an academic institution, and many objects were acquired through

travellers, collectors, or adventurers rather than systematic research (Woitschová, et al., 2023: 22). It was only in 1949, after becoming part of the National Museum, that the institution began to employ professional researchers. As part of a socialist state, the museum also played a role in representing the so-called colonised world within ideological frameworks influenced by the Global Cold War (Woitschová, et al., 2023: 189). This institutional position, neither colonial nor neutral, has important implications for contemporary engagements with communities of origin and for current debates on digital repatriation.

At the beginning of his ethnographic career, Šolc focused primarily on indigenous peoples of North America, later shifting his attention to Latin America. He conducted research in Bolivia, Peru, Mexico, and most extensively in Chile. Owing to diplomatic and academic cooperation between socialist Czechoslovakia and Chile during the 1960s and early 1970s, Šolc undertook four research trips to Chile, where he worked with Aymara, Mapuche, and Kawésqar communities. These research stays, ranging from several weeks to several months, were financially supported by the Universidad de Chile, for which Šolc also prepared official research reports. In Chile, he was regarded as a respected museologist and ethnographer, contributing particularly to Aymara ethnology and local archaeology. He was among the first researchers to study Aymara communities in northern Chile and prepared a detailed proposal for an archaeological and ethnographic museum in Arica, which ultimately remained unrealized. During his final research trip in 1973, Šolc was forced to leave the country following the military coup and was never able to return. He subsequently served as director of the Náprstek Museum.

While Šolc published scholarly articles and popular travelogues and adventure stories, his most extensive legacy is his photographic collection, which comprises over 4,000 photographs from his travels in Latin America and Asia (Šejbl, 2023). Photography had been integral to anthropology since the discipline's emergence and was long regarded as a privileged scientific method (Perera, 2019: 4). Ethnographic photography, however, is inseparable from colonial epistemologies and from what has been described as the "colonial gaze" (Burke, 2001: 129). Although Šolc operated outside a formal colonial framework, his photographic practice nevertheless reflects structural inequalities between a foreign researcher affiliated with both European and Chilean institutions and members of indigenous communities. These asymmetries are evident, for example, in the frequent absence of personal names in photographic descriptions, where individuals are labelled with generic identifiers such as "Aymara man" or "Kawésqar children." At the same time, Šolc's photographs do not depict their subjects in overtly sensationalist or degrading ways, the relations with him and the people he photographed were in many cases friendly, which complicates their interpretation and situates them within a spectrum of ethically ambivalent ethnographic imagery.

During his first research trip in 1966–1967, Šolc visited the predominantly Aymara village of Chapiquiña in the province of Arica. This fieldwork resulted in a research report for the Universidad de Chile and a series of ethnographic photographs documenting village life, architecture, and local inhabitants. He subsequently travelled to Lake Budi, where he visited a Mapuche community on Budi Island and documented daily activities as well as games of *palín*. These experiences were later described in his travelogue *Pod chilskými sopkami* (Under the Chilean Volcanoes, Šolc, 1969), which incorporated photographic portraits of indigenous inhabitants alongside narrative accounts of his encounters. Šolc returned to Chile in 1968 and spent several months in the Aymara village of Enquelga in the province of Tarapacá. There, he documented everyday life, housing, and construction practices, which he later analysed in his article *Casa Aymara en Enquelga* (Šolc, 2011). His work in Enquelga formed part of a broader engagement with Aymara communities that had begun earlier in Bolivia, where he conducted research near Lake Titicaca and acquired knowledge of the Aymara language, an asset that significantly shaped his subsequent fieldwork. In addition to photographs, Šolc collected approximately 200 objects of everyday use, including textiles, tools, and decorative items, for the Náprstek Museum.

Šolc's third major research trip led him to Puerto Edén, a settlement on Wellington Island in Chilean Patagonia. Supported once again by the Universidad de Chile, he conducted fieldwork among the Kawésqar community at a moment of profound social transformation. Traditionally a maritime nomadic society organised around nuclear families travelling by canoe through the Patagonian channels, the Kawésqar were undergoing a transition toward sedentary life under Chilean state administration. By the late 1960s, Kawésqar families had begun settling in Puerto Edén, which was under the jurisdiction of the *carabineros* (Gleisner & Montt, 2014: 23). Children attended a Spanish-language school, and the community lived for the first time in Western-style housing. Puerto Edén was also inhabited by non-indigenous settlers, primarily from Chiloé Island, as well as members of the Chilean police and military.

These transformations form a crucial backdrop for understanding both Šolc's photographs and their contemporary reception.

In Puerto Edén, Šolc expanded his photographic collection with approximately sixteen portraits of Kawésqar individuals and produced both an ethnographic report and a popular travelogue, *K Alakalufům na konec světa* (To the Alacalufs at the End of the World, Šolc, 1974). These texts reveal a noticeable distance between the researcher and the community he studied, shaped by linguistic barriers, limited duration of fieldwork, and culturally embedded assumptions. These factors are also reflected in the photographic material, where many individuals remain unidentified or are assigned incorrect names. These inaccuracies complicate present-day efforts to re-establish connections between the photographic archive and the descendants of those portrayed.

Today, Šolc's photographic collection is preserved in the Náprstek Museum and is partially digitised within the museum's Ethnographic Photographic Collection, which contains nearly 43,000 photographs. The database[1] allows full-text searching and represents a significant step towards accessibility. However, it remains difficult to navigate for non-specialist users. Linguistic barriers, technical complexity, incomplete digitisation, and the presence of institutional watermarks limit its usability for members of communities of origin. Access to unwatermarked images requires direct contact with the museum, reinforcing institutional mediation. This case illustrates how digitisation alone does not guarantee meaningful access and underscores the continued need for human facilitation in digital heritage infrastructures.

This research builds upon earlier fieldwork conducted by Tereza Melicharová, curator of the Náprstek Museum's Central and South American collections. During her research among the Mapuche community at Lake Budi, she established long-term cooperation with descendants of those photographed, shared physical and digital copies of 234 photographs, provided translated texts, and facilitated access to Šolc's reports (Melicharová, 2024). This project represents an example of participatory museology aimed at sharing authority and knowledge with communities of origin.

Following this approach, the present research focused on Aymara and Kawésqar collections that had not previously been systematically studied. Information had to be reconstructed from inventory cards, archival documentation, and contemporary witnesses. The research encountered significant challenges related to access, communication, and mediation. In the case of the Kawésqar, most community members now live in Puerto Edén or in urban centres such as Puerto Natales and Punta Arenas. Establishing contact proved difficult due to geographic isolation, limited digital visibility, and the absence of local intermediaries. Puerto Edén itself is accessible only via a lengthy boat journey, and accommodation options are scarce.

Contact was eventually established through Fundación Pueblo Kawésqar, an organisation dedicated to cultural revitalisation of the Kawésqar culture which cooperates with European museums mainly in Switzerland. Although this facilitated travel to Puerto Edén, initial attempts to connect with community members through social media were largely unsuccessful due to lack of interest in the photographs and Šolc's travelogue. The only direct identification of individuals in the photographs was achieved through informal encounters with a local host and later through remote interviews and academic networks. These experiences highlight the fragility of communication channels through which digital heritage is expected to circulate and demonstrate that digital repatriation is often a slow, uneven, and relational process rather than a straightforward technological solution. The case of Kawésqar photographs also highlights the uneven relations between museums and communities of origin in terms of knowledge and access to collections. This is illustrated by an interview with a Kawésqar participant whose family appears in numerous photographs held in European museums (including the Náprstek Museum) and who expressed a desire to visit these institutions and establish contact with them. Such engagement, however, often depends on the availability of funding, either from museums allocating resources for collaboration with communities of origin or from external sources.

A similar pattern emerged in the case of Aymara communities, although social media proved more effective in establishing contact. Communication with descendants of individuals portrayed in the photographs led to emotionally significant moments of recognition and to renewed engagement with Šolc's work. Subsequent visits to Chapiquiña enabled the sharing of photographs and archival texts with remaining inhabitants and their families. However, the limitations faced by a researcher without direct authority to distribute digital materials underscore the importance of institutional commitment and

[1] Národní muzeum. Archivní Vademecum: etnografická fotosbírka. https://vademecum.nm.cz/nm/permalink?xid=7b6825d050ade9b82c49b80c5976cc25.

cooperation in digital repatriation efforts. A further issue that was identified pertained to the advanced age of the photographed people, along with other individuals associated with the village. This complicated the possibility of disseminating digital copies of the photographs.
The findings of this research demonstrate that digital repatriation of ethnographic photographs is a complex, uneven process that extends far beyond the act of digitisation itself. While the existence of a publicly accessible digital database represents an important step toward openness, meaningful engagement with communities of origin depends largely on personal mediation, trust-building, and the active involvement of curators and researchers. For the communities involved, the photographs were predominantly perceived not as problematic or offensive heritage, but as powerful vehicles of memory, identity, and continuity, particularly when contextualised and reconnected with personal and family histories. At the same time, the research highlights structural barriers to digital public engagement, including linguistic limitations, non-user-friendly interfaces, watermarking practices, and institutional constraints on image sharing. This case study demonstrates that digitised ethnographic photography acquires meaning through its present-day use and relational activation, and that genuinely participatory digital repatriation requires approaches that prioritise human agency, ethical responsibility, and long-term collaboration with communities of origin.

ACKNOWLEDGEMENTS

The research underlying this proposal is supported by the Charles University Grant Agency (GA UK), project no. 532225.
Generative AI was used for translation and proofreading.

REFERENCES

Burke, P. (2001). *Eyewitnessing: the uses of images as historical evidence*. Cornell University Press.

Edwards, E. (2015). Anthropology and Photography: A long history of knowledge and affect. *Photographies*, 8(3), 235–252. https://doi.org/10.1080/17540763.2015.1103088.

Gleisner, C., & Montt, S. (2014). *Kawésqar serie: Introducción histórica y relatos de los pueblos originarios de Chile*. Ministerio de Agricultura, FUCOA.

Kjellman, U. (2025). Managing Photography of Racial Science in Uppsala University Library. In P. B. Larsen & M. Křížová (Eds.), *European University Legacies: Problematic Heritage and Contemporary Practice*. Edinburgh University Press, 273–295.

Melicharová, T. (2024). Mapuche artifacts: The Václav Šolc and Olga Kandertová collection in the Náprstek Museum. *Annals of the Náprstek Museum*, 45(1), 31–88. https://doi.org/10.37520/anpm.2024.002

Perera, S. (2019). Photography and the Ethnographic Method. *Oxford Research Encyclopedia of Education*. https://oxfordre.com/education/view/10.1093/acrefore/9780190264093.001.0001/acrefore-9780190264093-e-379.

Šejbl, J. (2023). Photographic estate of Václav Šolc in the ethnographic photographic collection of the Náprstek Museum. *Annals of the Náprstek Museum*, 44(2), 3–40. https://doi.org/10.37520/anpm.2023.007.

Šolc, Václav. (2011). Casa aymara en Enquelga. *Chungará (Arica)*, *43*(1), 89-111. https://dx.doi.org/10.4067/S0717-73562011000100006

Šolc, V. (1969). *Pod chilskými sopkami*. Orbis.

Šolc, V. (1974). *K Alakalufům na konec světa*. Orbis.

Woitschová, K., et al. (2023). *Dějiny Náprstkova muzea I*. Národní muzeum.

Una storia da ascoltare. Descrivere, mappare, connettere: un modello per la valorizzazione degli archivi musicali

MARTINA GREMIGNAI[1], LUCA ANDREA LUDOVICO[2]
[1]UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA/CONSERVATORIO VECCHI-TONELLI DI MODENA – MARTINA.GREMIGNAI@UNIMORE.IT
[2]UNIVERSITÀ DEGLI STUDI DI MILANO – LUCA.LUDOVICO@UNIMI.IT

ABSTRACT (ITALIANO)

La digitalizzazione del patrimonio culturale ha conosciuto negli ultimi anni una forte accelerazione, che ne permette un accesso libero e inclusivo. Tuttavia, rendere accessibile non significa rendere comprensibile, e questo problema risulta particolarmente rilevante nel caso del patrimonio musicale, che richiede passaggi di mediazione ulteriori rispetto ad altre tipologie documentarie. Una partitura digitalizzata, infatti, necessita di una sonorizzazione per restituire la dimensione temporale e sonora che le è intrinseca e la caratterizza. Oltre a questa specificità si aggiunge il fatto che l'alfabetizzazione musicale, intesa come capacità di decodifica della notazione, non è una competenza trasversale.
Il presente contributo propone un modello di valorizzazione del patrimonio musicale digitalizzato, attualmente in fase di sviluppo e sperimentazione su un corpus di fonti provenienti da archivi storici modenesi. Il modello si articola in tre dimensioni integrate: descrivere le fonti digitalizzate attraverso trascrizioni testuali e sonore; mappare i luoghi di produzione e circolazione dei materiali tramite georeferenziazione; connettere documenti, agenti e luoghi attraverso una struttura reticolare che superi la consultazione lineare.
Dal punto di vista metodologico, il progetto intercetta la didattica della storia basata sulle fonti materiali, l'approccio microstorico e le strategie di storytelling, i principi dello *Universal Design for Learning* e la *urban musicology*. La piattaforma web adotta standard internazionali per la rappresentazione testuale e musicale (TEI, MEI, IEEE 1599) e prevede una sperimentazione con classi della scuola secondaria di primo grado.
Il fine ultimo è proporre un modello replicabile su qualsiasi *repository* digitale, offrendo una riflessione più ampia su come rendere il patrimonio digitalizzato accessibile e significativo anche per il pubblico non specialistico.
Parole chiave: patrimonio musicale digitale; mediazione educativa; rappresentazione multimodale; modello integrato; IEEE 1599

ABSTRACT (ENGLISH)

A Story Worth Listening. Describing, Mapping, Connecting: A Model for Enhancing Music Archives
The digitisation of cultural heritage has increased significantly in recent years, enabling free and inclusive access. However, access to information does not automatically mean understand them: this problem is particularly relevant in the case of musical heritage, which requires additional mediation compared to other types of documents. A digitised score, in fact, needs sonification to restore its temporal and sonic dimension. Plus, the ability to decode notation is not a universal competence.
This paper proposes a model for enhancing digitised musical heritage, currently under development and experimentation on a corpus of sources from historical archives in the Modena area. The model is structured around three integrated dimensions: describing digitised sources through textual and audio transcriptions; mapping the places of production and circulation of materials through georeferencing; connecting documents, agents, and places through a network structure that overcomes linear consultation.
From a methodological perspective, the project intersects source-based history education, microhistorical approaches and storytelling strategies, Universal Design for Learning principles, and urban musicology. The web platform adopts international standards for textual and music representation (TEI, MEI, IEEE 1599) and includes an experimentation phase with lower secondary school classes.
The goal is to propose a replicable model for any digital repository, offering a broader reflection on how to make digitised heritage accessible and meaningful for non-specialist audiences.

Keywords: digital music heritage; educational mediation; multimodal representation; integrated framework; IEEE 1599

1. INQUADRAMENTO

La presente proposta muove da una specifica domanda di ricerca: come valorizzare il patrimonio musicale digitalizzato in maniera efficace, anche per un pubblico non specialistico?
L'attività di ricerca parte dalla constatazione dell'evidente dicotomia tra accesso e comprensione. Il lavoro di digitalizzazione del patrimonio culturale, incluso quello musicale, che negli ultimi anni ha visto una forte accelerazione, è imprescindibile: senza iniziative come il Piano Nazionale di Digitalizzazione del Patrimonio Culturale (PND)[1], solo per citare un'iniziativa di ampio respiro, non sarebbe neppure possibile formulare questa riflessione.
Tuttavia, rendere il patrimonio culturale liberamente accessibile non comporta una comprensione automatica da parte di chi lo visualizza. Il rischio concreto è che i materiali digitalizzati, pur liberamente accessibili, rimangano appannaggio esclusivo di studiosi e specialisti e non permettano di generare nuova conoscenza condivisa intorno a tali materiali.
Questa criticità assume una rilevanza particolare nel caso del patrimonio musicale: una partitura digitalizzata, a differenza di un dipinto o di un documento testuale, richiede più passaggi di mediazione. Si tratta di una caratteristica intrinseca degli oggetti musicali, che anche nelle forme più analogiche necessitano di una sonorizzazione, una "traduzione" dal segno al suono.
A causa di questa specificità intrinseca al tipo di materiale, e per interesse precipuo degli autori, il progetto esplora le possibilità digitali di mediazione del patrimonio musicale: una volta garantito l'accesso digitale ai documenti, è necessario prevedere strumenti che in primis sonorizzino tali fonti e restituiscano agli oggetti musicali digitalizzati la dimensione temporale e sonora che è loro intrinseca. In secondo luogo, è centrale superare il concetto di musica fine a sé stessa e calarla nel contesto storico-culturale in cui veniva fruita, esplicitando le storie, i luoghi e le reti di relazioni che conferivano senso a tali musiche. In quest'ottica, sono preziosissime tutte quelle fonti indirette in cui la musica è oggetto di interesse, come ad esempio epistolari, ricevute di acquisto e cronache.
A partire da queste premesse, si propone un modello di approccio attualmente in fase di sviluppo e sperimentazione. Il corpus principale di indagine è costituito dal repertorio musicale dei salotti modenesi del primo Ottocento italiano, scelto non solo per la ricchezza delle fonti negli archivi storici, ma anche per la disomogeneità nelle indagini finora condotte. Tale fenomeno è dovuto alla disponibilità di una grande quantità di documenti e alla frammentarietà di alcuni di essi (Indulti & Schiavone, 2009). In particolare, il corpus preso in esame si concentra sull'attività musicale legata alla famiglia Salimbeni ed è costituito da più di duemila esemplari musicali e da un centinaio fra epistole e ricevute d'acquisto.
Il fine più ampio di questo progetto è la determinazione di un modello generalizzabile e replicabile su altre collezioni musicali e, potenzialmente, su altri domini del patrimonio culturale.

2. CORNICE METODOLOGICA

Prima di descrivere la struttura del modello, è opportuno esplicitare quali siano gli ambiti che questo progetto intercetta, essendo per sua natura fortemente interdisciplinare.
In primis, il progetto ha un forte taglio pedagogico, visto l'indirizzo quanto mai attuale di una didattica della storia basata sulle fonti materiali: è ormai condivisa l'opinione che fare storia "con gli oggetti" favorisca un apprendimento significativo, nettamente più profondo della sola trasmissione dei concetti. L'oggetto musicale rappresenta un gancio particolarmente efficace, poiché l'ascolto può attivare un coinvolgimento emotivo meno mediato dal filtro di comprensione linguistica. Allo stesso tempo, come già evidenziato, la natura sonora e performativa dell'oggetto musicale ne complica l'accessibilità: una pagina di musica non "parla" – anzi, non "suona" – di per sé, ma deve essere eseguita. (Maule, 2007; Fabbri, 2015)
Oltre alla presentazione visiva e alla sonorizzazione delle fonti, il progetto si concentra sul concetto di "microstoria" (Ginzburg, 1976; Loeser, 2019; Békéssy, 2023), intesa non come semplice riduzione di narrazione ad avvenimenti di storia locale, ma come metodo indiziario. Essa costituisce un gancio forte con il discente: a partire da una microstoria, circoscritta a una fonte specifica o a un personaggio storico,

[1] *ICDP - Piano nazionale di digitalizzazione del patrimonio culturale*, http://docs.italia.it/italia/icdp/ (cons. 04/02/2026).

è possibile da una parte evidenziare il contesto in cui è calata, e dall'altra ricostruire configurazioni storico-culturali più ampie. Nel caso in esame, è un approccio che si rivela particolarmente efficace per ridurre la distanza percepita fra le fonti e il discente, grazie alla possibilità di ascoltare le fonti dirette. In quest'ottica, le strategie di *storytelling*, già ampiamente utilizzate nella didattica museale, costituiscono uno strumento complementare: la narrazione non è esclusivamente divertente, ma fondata su fonti e fatti storici, e permette di ricostruire le microstorie in maniera coinvolgente e storicamente informata (Barber, 2016; De Luca, 2019).
Per garantire una fruizione multimodale e inclusiva, il modello adotta i principi dello *Universal Design for Learning* (UDL) per l'insegnamento musicale. A partire dai tre principi cardine del framework (*Multiple Means of Engagement, Multiple Means of Representation* e *Multiple Means of Action and Expression*), il modello affronta in primo luogo la rappresentazione dei materiali e prevede molteplici modalità di accesso ai contenuti: modalità visiva (fonte digitalizzata), testuale (trascrizione delle fonti digitalizzate, contestualizzazione storica e narrativa), sonora (esecuzione del brano, lettura ad alta voce dei materiali testuali), esplorativa (mappe, reti). Questo tipo di rappresentazione dei materiali risponde alle variabilità dell'apprendimento e supera anche la possibile assenza di competenze pregresse di lettura musicale: l'alfabetizzazione musicale, intesa in questa sede come la competenza di decodifica della notazione, non è una competenza che può essere considerata come acquisita né in un contesto scolastico, né tantomeno da un possibile pubblico generalista (Armes, Harry, Grimsby, 2022). Il principio dei *Multiple Means of Engagement* trova a sua volta declinazione nelle modalità di navigazione e nei percorsi tematici descritti più avanti, mentre il principio di A*ction and Expression* rimane una linea di sviluppo futuro.
Per la natura del progetto, l'organizzazione di questa massa eterogenea di fonti (spartiti, lettere, ricevute d'acquisto, cataloghi d'epoca digitalizzati) non può seguire una logica esclusivamente gerarchica: ad esempio, una lettera può menzionare brani musicali, momenti conviviali con musica, acquisti di strumenti o di partiture; analogamente, un singolo spartito può essere ricollegato a un momento conviviale o a una specifica ricevuta d'acquisto. Per restituire questa complessa rete di relazioni, il progetto propone una struttura che permette di collegare vari tipi di informazione – documenti, agenti (musicisti, acquirenti, editori) – in modo che siano esplorabili da diversi punti di vista. L'obiettivo è il superamento della consultazione lineare delle fonti per favorire una visione reticolare (Panciroli, 2022).
Questo tipo di visione, unitamente all'applicazione dello *Universal Design for Learning,* non solo trova un'eco dal punto di vista tecnico di organizzazione di fonti e dati, ma si configura come un forte *trait d'union* con una specifica branca della musicologia, la cosiddetta *urban musicology*, che studia i paesaggi sonori (*soundscapes*) (Schafer, 1985; Carter, 2002; Harding, 2022). Questo indirizzo di ricerca invita a considerare fortemente la dimensione spaziale e acustica come ambiente di fruizione attiva e multimodale (Fiore & Belotti, 2020). In questa ottica, la riflessione degli ultimi anni si è concentrata sull'esplorazione della "fonosfera", lo spazio sonoro *indoor* in cui ascolto musicale, conversazione e socialità si intersecano (AA.VV., 2002; Steffan, 2005).

3. STRUTTURA DEL MODELLO

Il modello proposto prevede la realizzazione di una piattaforma interattiva *online* che permette di descrivere, mappare e connettere i materiali digitalizzati, consentendo non solo all'utente di esplorarli singolarmente ma esplicitando le relazioni che intercorrono fra di essi (Baratè, Ludovico, 2020; Faschi, Avanzini, Ludovico, 2026).
In primo luogo, la piattaforma dà accesso alle fonti dirette e indirette digitalizzate: testi musicali (spartiti e partiture), epistolari e ricevute d'acquisto. Nel caso del progetto in corso, le digitalizzazioni sono state realizzate per valorizzare un archivio originariamente in formato analogico. In un'ottica di scalabilità, è però possibile considerare anche materiali specifici digitalizzati e già disponibili su piattaforme *open access*.
A partire dalle fonti visive, vengono prodotte una trascrizione testuale e una sonora: per quanto riguarda i testi musicali, si procede a una trascrizione manuale su MuseScore e alla produzione di una traccia audio di sintesi; per quanto riguarda i testi verbali, si procede a una trascrizione tramite il software Transkribus e alla produzione di una registrazione sonora.
Dopo la codifica dei materiali, si procede alla loro organizzazione in un database relazionale con metadati descrittivi, predisponendo l'applicazione degli standard TEI e MEI (Pierazzo, Rosselli Del Turco, 2023; Viglianti, 2023) secondo un approccio che permetta di esplicitare le relazioni fra i dati stessi.

Per quanto riguarda la visualizzazione ed esecuzione sonora degli oggetti musicali, si prevede l'applicazione dello standard IEEE 1599 (Baratè, Haus, Ludovico, 2016), che consente di visualizzare il testo musicale digitalizzato e la trascrizione in formato moderno e di seguire visivamente l'andamento dell'esecuzione musicale. Tra le caratteristiche dello standard IEEE 1599, ha particolare rilievo la possibilità di accogliere all'interno di un unico *file* la descrizione di testi musicali e performance differenti. La struttura multi-layer di IEEE 1599, che prevede la sincronizzazione fra vari livelli (notazionale, grafico, performativo-sonoro), costituisce la realizzazione tecnica dei *Multiple Means of Representation* enunciati dai principi dello UDL e permette, attraverso l'integrazione coerente dei diversi livelli, un approccio multimodale effettivo.
L'impianto complessivo del modello si regge sul dialogo funzionale fra i tre standard internazionali adottati: TEI per la codifica dei testi verbali (epistolari, ricevute, contratti), MEI per la rappresentazione simbolica della notazione musicale e IEEE 1599 come piattaforma di sincronizzazione multimodale, in grado di integrare la codifica simbolica con le rappresentazioni grafiche, audio e performative.
L'interoperabilità è così garantita su tre livelli complementari: di formato, grazie all'adozione condivisa di XML; di contenuto, attraverso l'aderenza a standard di dominio consolidati; di integrazione, grazie all'architettura multi-layer di IEEE 1599. Questa impostazione permette al modello di non limitarsi a una giustapposizione di codifiche eterogenee, ma di organizzarle in una struttura coerente e navigabile.
In un'ottica di scienza aperta, tutte queste informazioni vengono codificate in formati digitali standard e di interscambio, supportati anche in ambito web: per le immagini si prevede l'uso di .png o .jpeg, per i testi si ipotizza l'adozione di XML, secondo gli standard già citati, per i contenuti audio e video .mp3 e .mp4. La piattaforma è implementata in HTML, CSS e JavaScript, con backend in PHP.
In secondo luogo, il progetto intende procedere nella mappatura dei materiali attraverso la georeferenziazione dei luoghi di produzione ed esecuzione, tracciando inoltre i percorsi di circolazione di tali materiali musicali. Oltre a essere un solido strumento di ricostruzione storica secondo i principi della *urban musicology*, una visualizzazione cartografica permette di situare i materiali in uno spazio e collegarli a coordinate reali, aspetto di particolare interesse per oggetti dematerializzati da un'operazione di digitalizzazione. L'output previsto è una mappa georeferenziata tramite tecnologie *open source* di *web mapping*.
Infine, si passa a esplicitare le connessioni, ricostruendo le reti sociali che legavano musicisti, committenti, luoghi di fruizione e luoghi di produzione. Attraverso collegamenti interni, sarà possibile passare dalla lettura di una lettera al brano musicale e/o alla geolocalizzazione dei luoghi citati nella lettera stessa; allo stesso modo, il testo musicale sarà accompagnato dai collegamenti ai materiali a esso riferibili (Tomasi, 2022).
Per permettere una molteplicità di *Engagement* e rendere così la fruizione più personalizzata e orientata a un apprendimento significativo, si intende affiancare alla forma di navigazione sopra descritta una serie di percorsi didattici che si focalizzano su un tema specifico, come per esempio uno specifico ensemble di strumenti o un personaggio storico centrale in una lettera.

4. STATO DI AVANZAMENTO E SVILUPPI FUTURI

Al fine di testare l'effettiva fattibilità del progetto, è stato realizzato un primo modello su un corpus ristretto di fonti scelte. Nello specifico, per rimanere in ambito modenese, il modello pilota ha avuto come soggetto *Il Pazzariello – scene napolitane in atto unico*, opera lirica in un solo atto del compositore Luigi Gazzotti su libretto di Ferdinando Fontana. Il progetto pilota, già pubblicamente disponibile in rete[2], si è orientato su questo repertorio per motivi sia pratici che metodologici: si tratta infatti di un corpus di piccole dimensioni, costituito da una partitura manoscritta per canto e pianoforte, un libretto dattilografo, cinque lettere e un contratto di scrittura tra compositore e librettista. Nonostante ciò, il soggetto della proposta progettuale offre un corpus coerente e, allo stesso tempo, eterogeneo nelle fonti.
Ciò ha permesso di identificare prematuramente alcune difficoltà operative e di ipotizzare una procedura efficace per il trattamento dei materiali. In particolare, se da una parte è stato testato il software Transkribus per la trascrizione automatica testuale, la questione rimane ancora aperta per la trascrizione automatica del materiale musicale (Simonetta, Mondal, Ludovico, Ntalampiras, 2024).

[2] Si fornisce il link: https://ilpazzariello.lim.di.unimi.it/

In questo primo approccio si è scelto di rimandare a una fase successiva la mappatura dei luoghi, sia per privilegiare l'affinamento della procedura di studio e trattamento dei materiali, sia per realizzare in tempi rapidi un'interfaccia utente prototipale.
Per quanto riguarda il progetto generale legato ai salotti modenesi, il processo di metadatazione e di costruzione del database è in fieri. Allo stesso tempo, è in corso la produzione dei materiali audio, in collaborazione con enti di alta formazione musicale per quanto riguarda il patrimonio musicale, quale il Conservatorio di Musica «Orazio Vecchi- Antonio Tonelli» di Modena e Carpi, in seno a cui si sta costituendo un piccolo gruppo di ricerca dedicato, ed enti teatrali per quanto concerne la lettura dei testi digitalizzati, quale *Lo Spazio Bianco APS*, associazione teatrale senza scopo di lucro.
Per testare l'usabilità della piattaforma e l'efficacia dell'architettura proposta, si prevede una sperimentazione che coinvolgerà classi della scuola secondaria di I grado, in realtà scolastiche sia legate al territorio sia estranee a esso. La procedura prevede una prima fase di utilizzo guidato, seguita da un'esplorazione autonoma; si ipotizza inoltre la somministrazione di un questionario, sia agli studenti sia ai docenti di riferimento.
Oltre a testare l'usabilità e l'efficacia della piattaforma – scopo primario della sperimentazione – si vuole anche operare un confronto fra l'esperienza di classi del territorio modenese, per le quali i materiali costituiscono un patrimonio locale, e classi di altri territori, per valutare quanto la dimensione spaziale incida sull'*engagement* e l'interesse verso i materiali proposti. La fase sperimentale costituisce inoltre l'occasione per raccogliere indicazioni sul principio UDL di *Multiple Means of Action and Expression*, la cui implementazione sistematica è ancora in fase di definizione, ma vede un primo prototipo nelle attività didattiche del progetto su *Il Pazzariello* di Gazzotti-Fontana. Il questionario prevederà specifiche domande sulle modalità con cui studenti e docenti vorrebbero poter interagire con i materiali, così da poter affinare il modello.

5. CONCLUSIONI

Il presente contributo propone un modello che trova una sua implementazione nel rilascio di una piattaforma web pubblicamente accessibile e basata su standard internazionali per la rappresentazione e l'interscambio di informazione musicale, quali TEI, MEI e IEEE 1599. Combinando le tre dimensioni del descrivere, mappare e connettere, è possibile realizzare forme di mediazione efficaci, realmente coinvolgenti per chi fruisce di tali materiali.
Rimangono alcune questioni aperte: sul piano della sostenibilità a lungo termine, la scelta di standard web consolidati garantisce la longevità della piattaforma, di fatto indipendente da soluzioni proprietarie. Restano invece aperti alcuni nodi sulla manutenzione e sulla produzione dei contenuti: in entrambi i casi è necessario prevedere il coinvolgimento di diversi specialisti, e ciò può comportare tempi dilatati. Altro aspetto non trascurabile riguarda la proprietà intellettuale dei materiali proposti dalla piattaforma, eterogenei nella loro natura e spesso appartenenti a istituzioni differenti.
La questione centrale rimane la scalabilità del progetto. Lo standard IEEE 1599 è già stato applicato a corpora di varia natura e di dimensioni molto diverse, come ad esempio la codifica della *Messa da Requiem* di Giuseppe Verdi (Baratè, Ludovico, 2012): questa applicazione dimostra la capacità dello standard di reggere oggetti musicali complessi, presentando all'utente la possibilità di esplorare in forma sincronizzata partitura, tracce audio multiple e video. Un'altra applicazione virtuosa si è concretizzata con *Music Studio*[3], una piattaforma per l'educazione musicale realizzata da Sanoma in collaborazione con il Laboratorio di Informatica Musicale dell'Università degli Studi di Milano, in cui è possibile esplorare oltre 150 brani di varie epoche e generi musicali. Il nodo non è dunque la scalabilità in sé, ma i tempi di preparazione, che variano in base alla natura e alla numerosità dei documenti. Il fine ultimo è proporre un modello replicabile su qualsiasi *repository* digitale, sia realizzato appositamente – come nel caso in esame – sia già esistente, appoggiandosi proprio a tutti quei materiali digitalizzati a cui è possibile avere accesso oggigiorno.
Nonostante il modello descritto nasca da un'esigenza precipua del patrimonio musicale, per il quale è necessaria una competenza specifica anche solo per la decodifica del testo musicale, il progetto vuole essere una proposta metodologica trasferibile anche ad altre discipline e indirizzi di studio, offrendo una riflessione più ampia su come rendere il patrimonio digitalizzato accessibile e significativo anche per il pubblico non specialistico.

[3] Si fornisce il link: https://musicstudio.lim.di.unimi.it/

BIBLIOGRAFIA

AA.VV. (2002). *La romanza italiana da salotto* (F. Sanvitale, A c. Di; EDT srl). EDT.

Armes, J. W., Harry, A. G., & Grimsby, R. (2022). Implementing Universal Design Principles in Music Teaching. *Music Educators Journal*, *109*(1), 44–51. JSTOR.

Baratè, A., & Ludovico, L. A. (2020). An Open and Multi-Layer Web Platform for Higher Music Education. *JOURNAL OF E-LEARNING AND KNOWLEDGE SOCIETY*, *16*(4), 29–37. https://doi.org/10.20368/1971%208829/1135356

Baratè, A., Haus, G., Ludovico, L.A. (2016) A Critical Review of the IEEE 1599 Standard. Computer Standards & Interfaces, vol. 46, 46–51. Elsevier.

Baratè, A., & Ludovico L. A. (2012). New Frontiers in Music Education through the IEEE 1599 Standard. *Proceedings of the 4th International Conference on Computer Supported Education* (CSEDU 2012), 146-151, SciTePress - Science and Technology Publications.

Barber, J. F. (2016). Digital Storytelling: New Opportunities for Humanities Scholarship and Pedagogy. *Cogent Arts & Humanities*, *3*(1). https://doi.org/10.1080/23311983.2016.1181037

Békéssy, L. (2023). The reunion genre: A microhistory of the musical life of Pest-Buda in 1857. *Arti Musices: Hrvatski Muzikološki Zbornik*, *54*(2), 397–413. RILM Abstracts of Music Literature with Full Text (A2411858).

Carter, T. (2002). The sound of silence: Models for an urban musicology. *Urban History*, *29*(I), 8–18. https://doi.org/10.1017/S0963926802001025

De Luca, M. R. (2019). Opera, evento e storytelling. Costruire la storia attraverso la musica d'arte. *Musica Docta*, *9*(1), 151–171. https://doi.org/10.6092/issn.2039-9715/10195

Fabbri, P. (2015). Una sfida didattica: Insegnare storia della musica. *Musica Docta*, *5*(1), 5–10. https://doi.org/10.6092/issn.2039-9715/5866

Faschi, V., Avanzini, F., & Ludovico, L. A. (2026). Fostering Openness and Accessibility of Music Archives through Multilayer Formats. *Umanistica Digitale*, *10*(22), 25–43. https://doi.org/10.60923/issn.2532-8816/21759

Fiore, A., & Belotti, S. (2020). Merging Music and Landscape: Un approccio digitale per lo studio dell'identità culturale della Modena estense. *magazén*, *1*(1), 75–100. https://doi.org/10.30687/mag//2020/01/004

Ginzburg, C. (1999). *Il formaggio e i vermi: Il cosmo di un mugnaio del '500*. Einaudi.

Harding, V. (2022). Introduction: Music and urban history. *Urban History*, *29*(1), 5–7. https://doi.org/10.1017/S0963926802001013

Indulti, G., & Schiavone, R. (2009). Nuclei storici nell'Istituto "Orazio Vecchi" e fonti per una biografia di Giuseppe Venceslao Köhler: Una ricerca d'archivio tra Modena, Praga e Litomerice. *Quaderni Estensi*, *I*, 22.

ICDP - Piano nazionale di digitalizzazione del patrimonio culturale, http://docs.italia.it/italia/icdp/

Il Pazzariello – Alla scoperta dell'opera inedita di Luigi Gazzotti e Ferdinando Fontana, https://ilpazzariello.lim.di.unimi.it/

Loeser, M. (2019). The many faces of everyday musical life: Approaching music history from "below". In *Muzikologija: Časopis Muzikološkog Instituta Srpske Akademije Nauka i Umetnosti* (Vols. 1–27, pp. 31–49). Srpska Akademija Nauka i Umetnosti. RILM Abstracts of Music Literature with Full Text (A2072884). https://research.ebsco.com/linkprocessor/plink?id=f31cfab4-e411-30f2-9cf2-b2ed7b90ff82

Maule, E. (2007). *Storia della musica: Come insegnarla a scuola*. ETS.

Music Studio, https://musicstudio.lim.di.unimi.it/

Panciroli, C. (2022). *Elementi di didattica post-digitale*. Bononia University Press.

Pierazzo, E., & Rosselli Del Turco, R. (2023). Critica testuale e nuovi metodi: L'edizione scientifica digitale. In *Digital Humanities. Metodi, strumenti, saperi*. Carrocci.

Schafer, R. M. (1985). *Il paesaggio sonoro*. Ricordi. (http://id.sbn.it/bid/CFI0063616).

Simonetta, F., Mondal, R., Ludovico, L. A., & Ntalampiras, S. (2024). Optical Music Recognition in Manuscripts from the Ricordi Archive. *Proceedings of the 19th International Audio Mostly Conference: Explorations in Sonic Cultures*. https://doi.org/10.1145/3678299.3678324

Steffan, C. (2005). *Cantar per salotti: La musica vocale da camera (1800-1850): Testi, contesti e consumo*. Istituti editoriali e poligrafici internazionali.

Tomasi, F. (2022). *Organizzare la conoscenza: Digital humanities e web semantico: Un percorso tra archivi, biblioteche e musei*. Editrice bibliografica.
Viglianti, R. (2023). Musicologia digitale e spartiti digitali. In *Digital Humanities. Metodi, strumenti, saperi*. Carrocci.

LETSdigit: Archivio digitale della letteratura a Trieste

ALINA JILL SIMEONE[1], CRISTINA FENU[2], MARINA BUZZONI[3], ROBERTO ROSSELLI DEL TURCO[4]

[1]UNIVERSITÀ CA' FOSCARI VENEZIA, ITALIA – ALINAJILL.SIMEONE@UNIVE.IT
[2]BIBLIOTECA CIVICA "ATTILIO HORTIS" TRIESTE, ITALIA - CRISTINA.FENU@COMUNE.TRIESTE.IT
[3]UNIVERSITÀ CA' FOSCARI VENEZIA, ITALIA - MBUZZONI@UNIVE.IT
[4]UNIVERSITÀ DI TORINO, ITALIA - ROBERTO.ROSSELLIDELTURCO@UNITO.IT

ABSTRACT (ITALIANO)

Il contributo illustra il progetto di ricerca LETSdigit: Archivio digitale della letteratura a Trieste, finanziato nell'ambito di "Dicolab. Cultura al digitale" subinvestimento 1.1.6 "Formazione e Miglioramento delle competenze digitali" del PNRR M1C3. LETSdigit è finalizzato alla progettazione di un archivio digitale dedicato alla letteratura primaria e secondaria relativa alla produzione di scrittrici e scrittori operanti nel contesto triestino tra XIX e XX secolo. A partire dall'edizione genetica digitale del manoscritto del *Canzoniere 1900-1920* di Umberto Saba e dalle collezioni digitali di oggetti di formati eterogenei, inclusi materiali sonori, conservati nella Biblioteca Civica di Trieste, il progetto sperimenta un *workflow* metodologico che combina processi semi-automatici di trascrizione e codifica testuale secondo lo standard XML/TEI, supportati dall'impiego controllato di sistemi LLM. Tali strumenti sono integrati in un quadro metodologico orientato alla supervisione critica dei processi automatizzati di *encoding*, secondo criteri di trasparenza e verificabilità. Il progetto affianca alla modellizzazione testuale l'annotazione semantica, tramite l'identificazione delle *named entities* e il collegamento a risorse esterne secondo i principi dei *Linked Open Data*. Il contributo espone le scelte alla base dell'architettura del progetto, presenta i primi risultati e delinea prospettive di sviluppo inerenti all'integrazione di materiali sonori come componenti strutturate dell'edizione digitale, in linea con le più recenti sperimentazioni su archivi orali e edizioni multimodali.
Parole chiave: edizione scientifica digitale; patrimonio letterario digitalizzato; servizi digitali; accessibilità e riuso

ABSTRACT (ENGLISH)

LETSdigit: A Digital Archive of Literature in Trieste. This contribution illustrates the research project LETSdigit: A Digital Archive of Literature in Trieste, funded under "Dicolab. Cultura al digitale" sub-investment 1.1.6 "Training and Improvement of Digital Skills" of the PNRR M1C3. LETSdigit aims to design a digital archive dedicated to primary and secondary literature relating to the work of writers active in Trieste between the 19th and 20th centuries. Starting from the digital genetic edition of Umberto Saba's *Canzoniere 1900-1920* manuscript and the digital collections of objects in various formats, including sound material, preserved at the Trieste Civic Library, the project experiments with a methodological workflow that combines semi-automatic transcription and text encoding processes according to the XML/TEI standard, supported by the controlled use of LLMs. These tools are integrated into a methodological framework aimed at the critical supervision of automated encoding processes, with a view to transparency and verifiability. The project combines textual modelling with semantic annotation activities, through the identification of named entities and their linking to external resources according to the principles of Linked Open Data. This contribution illustrates the choices underlying the project's architecture, presents the initial results, and outlines future development perspectives related to the integration of sound materials as structured components of the digital edition, in line with the most recent experimental work on oral archives and multimodal editions.
Keywords: digital scholarly editing; digitized literary heritage; digital services; accessibility and reuse

1. INTRODUZIONE

Il progetto LETSdigit: Archivio digitale della letteratura a Trieste (CUP F84D21000010006), nato nell'ambito del Piano Nazionale di Ripresa e Resilienza (PNRR) e frutto di una collaborazione tra la Biblioteca Civica "Attilio Hortis" di Trieste e l'Università Ca' Foscari di Venezia, prevede la sperimentazione di un modello di archivio digitale della letteratura primaria e secondaria relativo alla produzione di scrittrici e scrittori di Trieste. Il programma prende avvio dall'edizione genetica digitale del manoscritto

R.P. 1-18 del *Canzoniere 1900-1920* di Umberto Saba (Buzzoni et al., 2024), conservato presso la Biblioteca Civica di Trieste, per integrarlo e arricchirlo con risorse digitali eterogenee disponibili presso la Biblioteca o esposte presso LETS – Museo della Letteratura Trieste. Il patrimonio di riferimento comprende circa 2.500 oggetti digitali riferiti al contesto letterario triestino e giuliano tra il 1860 e il 1955. Tra questi figurano, oltre al manoscritto sabiano, l'archivio sveviano con la corrispondenza tra Italo Svevo e James Joyce, altri archivi d'autore custoditi presso la Biblioteca, collezioni di periodici storici locali digitalizzati e registrazioni sonore riversate in formato digitale.
Il progetto ha preso avvio a dicembre 2025 e si concluderà a giugno 2026. Dopo una fase di selezione e analisi di tali oggetti, verranno sperimentate procedure semi-automatiche di trascrizione dei documenti tramite *software* OCR/ICR e di codifica secondo lo *standard* XML/TEI con l'ausilio di *chatbot* basati su *Large Language Models*. È previsto, inoltre, l'arricchimento semantico dei testi tramite l'annotazione di *named entities*. La visualizzazione dei documenti codificati avverrà tramite il *software* EVT[1], che nella versione 3 ha compiuto notevoli progressi nella rappresentazione delle entità nominate.

2. STATO DELL'ARTE

LETSdigit si inserisce nell'ambito dei progetti che prevedono la realizzazione di un ambiente digitale integrato. Particolare attenzione è dedicata alle pratiche di integrazione e armonizzazione di oggetti digitali eterogenei — comprendenti DSE, materiali testuali, visivi e sonori — nonché ai processi di annotazione semantica delle *named entities* e alla loro modellizzazione secondo i principi dei *Linked Open Data*, al fine di costruire un ecosistema informativo strutturato capace di contestualizzare, interconnettere e rendere interrogabili le fonti oggetto di analisi.
Negli ultimi anni, nel campo delle *Digital Humanities* si è registrata una crescita significativa di ricerche che impiegano soluzioni basate su LLM per attività comunemente considerate complesse e dispendiose in termini di tempo se svolte manualmente. Nell'ambito della filologia digitale, queste comprendono soprattutto la trascrizione dei documenti, la codifica XML/TEI e l'annotazione semantica delle *named entities*. Diversi studi hanno mostrato come questi sistemi non solo si integrano efficacemente nei *workflow* editoriali, ma in alcuni casi superano le capacità e la *performance* di strumenti OCR e *pipeline* di *Natural Language Processing* (NLP) consolidate.
Nell'ambito della trascrizione di documenti storici, è stato dimostrato (Kim et al., 2025) che *chatbot* basati su modelli come GPT-4o e Claude Sonnet 3.5 possono raggiungere livelli di accuratezza superiori rispetto a *software* come EasyOCR, Tesseract, Keras e TrOCR, specialmente quando l'*input* è opportunamente segmentato.
Per quanto riguarda la codifica secondo lo *standard* XML/TEI, studi recenti (Pollin et al., 2023, 2025) evidenziano come il *prompt engineering* e il *few-shot learning* permettano ai modelli generativi di produrre *output* ben formati e semanticamente coerenti, che richiedono un basso margine di controllo umano in fase di validazione e risultano integrabili nei flussi di lavoro delle edizioni digitali.
Analogamente, nell'ambito della *Named Entity Recognition* (NER), alcuni contributi (Galka & Vogeler, 2025; Hiltmann et al., 2025) rivelano che agenti conversazionali basati su LLM, se opportunamente guidati mediante *prompt* contestualizzati, sono in grado di ottenere prestazioni superiori a *framework* NLP come spaCy e Flair, risultando particolarmente efficaci anche su testi complessi.
In parallelo a queste sperimentazioni orientate all'analisi e strutturazione automatica del testo, la ricerca recente ha iniziato a interrogarsi anche su forme di arricchimento dell'edizione digitale che coinvolgono altri media, in particolare la dimensione sonora.
Negli ultimi anni l'integrazione di componenti audio nelle edizioni scientifiche digitali è stata discussa non soltanto come supporto all'accessibilità, ma anche come possibilità di estendere la dimensione critica dell'edizione, riconoscendo alla *performance* orale uno statuto testuale autonomo. In questa prospettiva, la registrazione della lettura d'autore può essere interpretata come una "versione" del testo, capace di chiarire ambiguità attraverso intonazione e ritmo e, talvolta, di introdurre deviazioni significative rispetto alla redazione fissata a stampa. Levy, Moffatt e Sherren (2020) sottolineano come la lettura dal vivo possa configurarsi come un atto di revisione implicita, mentre la trascrizione dell'audio costituisce essa stessa un processo editoriale non neutro, che richiede scelte interpretative nel passaggio da medium sonoro a medium visivo.

[1] http://evt.labcd.unipi.it/ (cons. 05/02/2026).

Nell'ambito degli archivi sonori e audiovisivi, progetti come *Archivio Vi.Vo.* (Archivio VideoVoce)[2], coordinato dall'Università di Siena, hanno evidenziato la necessità di modellizzare il documento sonoro come unità autonoma, dotata di metadati strutturati e inserita in infrastrutture interoperabili orientate ai principi FAIR (Calamai et al., 2022; Valentini et al., 2023). In continuità con tali approcci, il progetto STOA (*Space and Time in Oral Archives*)[3] amplia la prospettiva verso il riuso e la restituzione degli archivi sonori in contesti esperienziali e territoriali, mettendo in relazione voce, spazio e memoria. Queste esperienze mostrano come il documento sonoro possa essere trattato e valorizzato all'interno di infrastrutture archivistiche.
Altre sperimentazioni si collocano, invece, sul versante editoriale. È il caso di *CantApp: The General Prologue*[4], che integra trascrizione, apparato di supporto alla lettura, immagini del manoscritto e lettura audio sincronizzata, dimostrando come la dimensione sonora possa diventare componente strutturale dell'esperienza editoriale e incidere direttamente sulla leggibilità dell'edizione multimodale (Bordalejo et al., 2021).
Tale cornice teorico-applicativa risulta particolarmente pertinente per progetti che, come LETSdigit, intendono mettere in relazione una versione genetica (manoscritta) e una versione canonica del testo, affiancando a quest'ultima anche la resa sonora d'autore come ulteriore livello di testimonianza. In questa prospettiva, la codifica multimodale (testo, facsimile e audio) non rappresenta un mero arricchimento multimediale, ma un ampliamento metodologico dell'edizione digitale, capace di offrire accesso, interpretazione e stratificazione critica attraverso più canali semiotici.

3. METODOLOGIA

In una prima fase è stata condotta una ricognizione sistematica delle collezioni digitali disponibili presso la Biblioteca Hortis, comprendente fonti manoscritte, a stampa, iconografiche e sonore. La mappatura è stata svolta in funzione della correlazione dei contenuti al manoscritto del *Canzoniere* e considerando le future modalità di accesso e fruizione dei contenuti, sia in ambiente digitale che museale. Mentre le risorse testuali e iconografiche possono rappresentare importanti approfondimenti biografici e risorse di contestualizzazione geografica e storico-culturale, le risorse audio possono essere restituite al pubblico aumentando l'engagement, arricchendo i contenuti testuali e migliorandone l'accessibilità.
Sulla base della mappatura, è stato avviato lo sviluppo di un modello di codifica XML/TEI adatto alla metadatazione e all'annotazione semantica, orientato alla descrizione di oggetti digitali eterogenei, all'identificazione di *named entities* e al collegamento con i *Linked Open Data*.
Nella fase di sperimentazione sono stati avviati dei test di trascrizione e codifica in XML/TEI semi-automatiche, condotti mediante agenti conversazionali basati su LLM, con l'impiego in particolare del modello GPT-5 di OpenAI (vd. sezione 4).
Seguiranno sperimentazioni riguardanti la trascrizione e l'integrazione dei contenuti audio, la creazione di *dataset* testuali arricchiti e la visualizzazione dei dati. I prototipi vengono valutati in termini di accessibilità, usabilità e riusabilità, in linea con i principi FAIR e con l'obiettivo di definire indicazioni operative e un modello di servizio replicabile.
Alla luce della letteratura recente sulle edizioni digitali multimodali e sulle pratiche di sincronizzazione tra testo e registrazioni sonore, il progetto LETSdigit intende sperimentare un'estensione dell'edizione genetica digitale del manoscritto del *Canzoniere* come edizione multimodale. A tal fine, verranno integrate risorse audio d'archivio già presenti nelle collezioni della Biblioteca Hortis, in particolare la registrazione su vinile *La Voce di Umberto Saba* (Fonit, 1974), già riversata in formato digitale, in cui il poeta declama anche tre Sonetti della *Giovinezza* (*Tra la folla*, *Fra chi dice d'amarmi*, *Nella notte di Natale*), tratti dall'edizione del *Canzoniere* del 1921.
Dal punto di vista metodologico, tale sperimentazione si colloca in continuità con quei modelli che, nell'ambito degli archivi orali, distinguono il livello del supporto da quello del contenuto e riconoscono

[2] *Archivio Vi.Vo.* (Archivio VideoVoce) è un'infrastruttura digitale per la conservazione, descrizione e valorizzazione di archivi orali e audiovisivi, coordinata dall'Università di Siena in collaborazione con CLARIN-IT e CNR-ILC. Il progetto, avviato nel 2019 con finanziamento della Regione Toscana, è stato successivamente sviluppato in ulteriori fasi progettuali. La piattaforma è accessibile all'indirizzo: https://archiviovivo.clarin-it.it/ (cons. 19/04/2026).
[3] *Space and Time in Oral Archives* (STOA), coordinato da Silvia Calamai nell'ambito dei progetti PRIN PNRR 2022, è dedicato alla digitalizzazione, metadatazione e valorizzazione di archivi orali, con particolare attenzione alla loro contestualizzazione spaziale e temporale e al riuso dei dati in ambienti di ricerca interoperabili. Il progetto è consultabile *online* https://sites.google.com/unisi.it/stoaproject/home (cons. 19/04/2026).
[4] https://www.sd-editions.com/CantApp/GP/ (cons. 19.04.2026).

nella registrazione di un evento comunicativo una unità documentaria autonoma (Calamai et al., 2022; Valentini et al., 2023). In questa prospettiva, ulteriormente sviluppata in progetti come STOA, l'attenzione si estende al riuso e alla restituzione del dato sonoro in relazione alle sue dimensioni spaziali, temporali ed esperienziali.
Nel caso di LETSdigit, questa impostazione viene trasferita nel dominio dell'edizione scientifica digitale: l'audio è concepito come oggetto digitale strutturato, dotato di autonomia documentaria, ma al tempo stesso collegabile a specifici segmenti testuali e interrogabile in relazione alla stratificazione testuale dell'opera. In questo senso, la registrazione della lettura d'autore è trattata come livello testuale ulteriore, integrato nel sistema di rappresentazione e fruizione dell'edizione.
Poiché il progetto dispone attualmente della trascrizione e codifica genetica del manoscritto (1900–1920), ma non di una trascrizione critica completa della redazione a stampa del 1921, la prototipazione sarà limitata alle tre liriche selezionate. Essa prevede la creazione di un testo di lettura privo di apparato esteso, destinato a fungere da base di riferimento per l'aggancio strutturale dell'audio.
L'integrazione sonora sarà modellata secondo un approccio *standoff*, mantenendo l'audio come risorsa autonoma e collegandolo al testo di lettura tramite un *layer* esterno di allineamento (per versi o segmenti) basato su *timestamp*, intesi come riferimenti temporali relativi alla *timeline* del file audio. In questo modo, la dimensione sonora, pur restando separata sul piano della codifica, viene ricondotta a un livello interpretativo integrato con il testo, rendendo esplicita la corrispondenza tra *performance* orale e articolazione testuale. Tale mappatura consente la sincronizzazione dinamica tra *player* e visualizzazione testuale (*highlight* del segmento corrente e navigazione verso il punto corrispondente), abilitando in interfaccia una fruizione sincrona e bidirezionale.

4. RISULTATI OTTENUTI

I test di trascrizione e di codifica in XML/TEI semi-automatiche mediante *Large Language Models* hanno visto l'impiego del modello GPT-5 in modalità *chat*. La sperimentazione è stata condotta sull'articolo di Umberto Saba intitolato *Il Montenegro*, pubblicato nel 1904 sul quotidiano triestino "Il Lavoratore". Il testo è stato selezionato come approfondimento biografico e come risorsa di contestualizzazione geografica rispetto alla lirica *Salendo il Monte Lovćen*, presente nel *Canzoniere 1900-1920* nella sezione *Versi militari* (1907-1908).
I primi esperimenti, in cui è stata fornita l'intera scansione della pagina di giornale, si sono rivelati insoddisfacenti, con trascrizioni parziali e imprecise, omissioni e allucinazioni. Un miglioramento significativo è stato ottenuto impiegando l'immagine ritagliata dell'articolo e fornendo istruzioni più dettagliate: la trascrizione generata è risultata accurata, con qualche errore riconducibile a scarsa leggibilità dell'immagine a causa del cattivo stato di conservazione del supporto cartaceo. Questi risultati evidenziano il ruolo centrale della qualità dell'immagine nei sistemi di riconoscimento e trascrizione automatica del testo basati su LLM. Un esempio del ritaglio dell'articolo fornito in *input* e del corrispondente *output* codificato in XML/TEI è riportato in Figura 1.



IL MONTENEGRO.

Cettigne, 2 luglio.

E' un paese semiselvaggio; all'aspetto un che di mezzo fra la Stiria e l'Oriente; ha della Stiria la conformazione rocciosa e poco feconda del terreno, dell'Oriente i costumi e la forma delle cose. Il popolo è immensamente simpatico, robusto di una robustezza omerica, ingenuo e primitivo, di sentimenti guerreschi ma inclinato — come tutte le razza slave — alla malinconia. Questa si dichiara specialmente nelle canzoni popolari, tecnicamente una specie di canto fermo; ma che ispirano sentimenti e immagini non dissimili da quelle della musica chopiniana.

Che se i ritmi dell'infelice polacco sono l'espressione di una civiltà raffinata e corrotta fino all'impossibile, i canti montenegrini ottengono per la via opposta, cioè per quella di una semplicità puerile effetti eguali, un senso di melanconia e di nostalgia quasi disperata.

```
<text>
    <body>
        <pb n="1" facs="#surf_1"/>
        <div type="article" xml:id="art_montenegro">
            <head>IL MONTENEGRO.</head>

            <p>
                <lb n="1"/><placeName ref="#cettigne">Cettigne</placeName>, 2 luglio.
                <lb n="2"/>
                <lb n="3"/>E' un paese semiselvaggio; all'aspetto
                <lb n="4"/>un che di mezzo fra la <placeName
                ref="#stiria">Stiria</placeName> e l'<placeName
                ref="#oriente">Oriente</placeName>;
                <lb n="5"/>ha della <placeName ref="#stiria">Stiria</placeName>
                la conformazione rocciosa
                <lb n="6"/>e poco feconda del terreno, dell'<placeName
                ref="#oriente">Oriente</placeName> i
                <lb n="7"/>costumi e la forma delle cose. Il popolo è
                <lb n="8"/>immensamente simpatico, robusto di una
                <lb n="9"/>robustezza omerica, ingenuo e primitivo, di
                <lb n="10"/>sentimenti guerreschi ma inclinato — come
                <lb n="11"/>tutte le razza slave — alla malinconia.
                <lb n="12"/>Questa si dichiara specialmente nelle can-<lb
                break="no" n="13"/>zoni popolari, tecnicamente una specie di
                <lb n="14"/>canto fermo; ma che ispirano sentimenti e
                <lb n="15"/>immagini non dissimili da quelle della mu-<lb
                break="no" n="16"/>sica chopiniana.
            </p>

            <p>
                <lb n="17"/>Che se i ritmi dell'infelice polacco sono
                <lb n="18"/>l'espressione di una civiltà raffinata e
                <lb n="19"/>corrotta fino all'impossibile, i canti mon-<lb
                break="no" n="20"/>tenegrini ottengono per la via opposta, cioè
                <lb n="21"/>per quella di una semplicità puerile ef-<lb
                break="no" n="22"/>fetti eguali, un senso di melanconia e di
                <lb n="23"/>nostalgia quasi disperata.
            </p>
```

Figura 1. Esempio di input e output del workflow semi-automatico basato su GPT-5 per trascrizione, codifica XML/TEI e annotazione semantica. (a) Ritaglio dell'articolo «Il Montenegro» di Umberto Saba, pubblicato sul quotidiano triestino "Il Lavoratore" (1904), utilizzato come input dopo la segmentazione dell'immagine. (b) Estratto dell'output generato dal modello con prompt procedurale strutturato, comprendente trascrizione e marcatura TEI della struttura testuale (<head>, paragrafi, a capo tramite <lb/>) e l'annotazione automatica di entità nominate (<placeName>), con assegnazione di identificativi e riferimenti interni (@ref)

A partire da queste criticità, la metodologia è stata progressivamente raffinata, sia sul piano dei dati forniti in *input*, sia sul piano della definizione del *prompt*. Da una richiesta sintetica si è passati ad un approccio procedurale articolato in precise istruzioni per trascrizione, codifica e individuazione di *named entities*.

Dal punto di vista metodologico la separazione dei passaggi di trascrizione e codifica ha permesso di basare la codifica su un testo già verificato, riducendo il numero di interventi manuali e aumentando il controllo sull'*output* finale. Inoltre, il modello ha riconosciuto correttamente le *named entities* nella maggior parte dei casi; tuttavia, in assenza di una definizione puntuale delle tipologie di entità nominate da annotare, tende a generare errori di classificazione.

L'esperienza ha mostrato come la definizione di un *prompt* strutturato consenta di rendere il processo più efficiente, migliorando la completezza dell'*output* rispetto alle esigenze del progetto.

Un aspetto critico emerso durante il test riguarda la variabilità dell'*output*: utilizzando il medesimo *prompt* in conversazioni diverse, i risultati possono differire sensibilmente. Alcuni errori di trascrizione e codifica risultano ricorrenti, mentre altri variano in modo casuale. Occorre tuttavia considerare che, anche in presenza di un *prompt* dettagliato, non è possibile garantire risultati pienamente stabili, a causa di fattori quali la natura probabilistica dei modelli generativi, la qualità e la risoluzione dell'immagine in *input*, nonché la rapida evoluzione del sistema, che rende difficile ottenere prestazioni costanti nel tempo. Alla luce di queste considerazioni, l'intervento umano rimane centrale nel processo di validazione filologica della trascrizione e della codifica.

Tuttavia, il passaggio da un *prompt* generico a un *prompt* strutturato, in cui viene definita una *pipeline* di lavoro e viene esplicitata la forma attesa dell'*output*, ha permesso di ridurre le ambiguità e di contenere la tendenza alle allucinazioni. Inoltre, la richiesta di emulare un esperto in filologia digitale e codifica TEI nell'elaborazione della risposta, una strategia di *role-based prompting*, ha contribuito ad allineare il comportamento del modello alle richieste.

La breve analisi condotta ha carattere qualitativo e si è basata su un campione limitato. Di conseguenza, i risultati non sono generalizzabili ed è fondamentale affiancare a questo tipo di valutazioni l'adozione di *benchmark* e metriche quantitative, per poter misurare in modo sistematico le prestazioni dei modelli su documenti storici (Levchenko, 2025). Lo schema di codifica XML/TEI, inoltre, è ancora in fase di

definizione: si stanno conducendo campionature per fornire uno schema di codifica completo e per estendere la metodologia a un *corpus* più ampio, inclusi materiali manoscritti.
Inoltre, il progetto prevede l'implementazione di un modello di annotazione semantica interoperabile, applicabile a oggetti culturali di diversa natura e provenienza. Tale modello consente di descrivere e mettere in relazione testi letterari, manoscritti, immagini d'archivio e materiali audiovisivi entro un quadro semantico comune, basato su *standard* aperti e vocabolari controllati.
Un aspetto particolarmente innovativo riguarda l'integrazione dei materiali sonori (letture d'autore, interviste, documenti audio d'epoca) non come semplici allegati, ma come componenti strutturate dell'edizione critica digitale, annotabili, collegabili a specifici passaggi testuali e interrogabili insieme agli altri contenuti.
Il lavoro nei prossimi mesi prevede la costruzione di *dataset* testuali semanticamente arricchiti, in cui i testi saranno corredati da informazioni strutturate su entità nominate (persone, luoghi, organizzazioni), relazioni concettuali, varianti e contesti storici e culturali. Queste risorse potranno essere impiegate in future analisi nell'ambito del NLP, ad esempio per l'addestramento o la valutazione di modelli di NER, per studi di stilometria e per l'analisi delle reti di relazioni tra autori, opere e luoghi. Ao stesso tempo, i dati saranno resi disponibili in forma accessibile, riusabile e interconnessa, favorendo pratiche di *open science* e l'integrazione con altri archivi e progetti affini.
Infine, il progetto ha portato alla definizione e porterà alla sperimentazione di servizi digitali dedicati al museo LETS. Tra questi rientrano percorsi di consultazione tematici che intrecciano testi, documenti sonori e materiali iconografici, strumenti per esplorare le relazioni tra autori e luoghi della città di Trieste e interfacce pensate per diversi profili di pubblico, dal ricercatore allo studente fino al visitatore non specialista. In questa prospettiva, il museo viene concepito non solo come spazio espositivo, ma come piattaforma digitale dinamica per la produzione, la diffusione e la valorizzazione della conoscenza letteraria.

5. CONCLUSIONI

Il progetto LETSdigit: Archivio digitale della letteratura a Trieste dimostra come risorse digitali già disponibili possano costituire il punto di partenza per la progettazione di servizi avanzati, capaci di valorizzare il patrimonio letterario attraverso accessibilità, interoperabilità e riuso.
I primi risultati evidenziano, inoltre, il potenziale dei sistemi LLM nel supportare attività filologiche complesse, quali trascrizione, codifica XML/TEI e annotazione semantica, purché inseriti in *workflow* controllati, basati su *prompt* strutturati e su una validazione umana sistematica.
Nelle fasi successive saranno sviluppate interfacce *user-friendly* per la fruizione digitale e museale, strumenti di analisi del *corpus* mediante tecniche di NLP, visualizzazioni avanzate dei dati e ulteriori integrazioni tra contenuti testuali, iconografici e sonori. In particolare, l'estensione multimodale dell'edizione genetica, attraverso la sincronizzazione tra testo e *performance* orale d'autore, mira a offrire un ulteriore livello interpretativo e critico, oltre che a rendere l'esperienza di consultazione più accessibile e coinvolgente.
Il modello sviluppato si propone infine come riferimento trasferibile ad altri contesti del patrimonio culturale digitalizzato, contribuendo alla definizione di buone pratiche replicabili per la costruzione di archivi e servizi digitali interoperabili.

RINGRAZIAMENTI

Il progetto di ricerca rientra nei cantieri di digitalizzazione dell'investimento PNRR "Strategie e piattaforme digitali per il patrimonio culturale" Missione 1 – Componente 3 – Misura 1 – Investimento 1.1. Sub investimento 1.1.6 "Formazione e miglioramento delle competenze digitali".
Più specificamente, fa parte di "Dicolab. Cultura al digitale", promosso dal Ministero della Cultura – Digital Library nell'ambito del PNRR Cultura 4.0; è realizzato dalla Scuola nazionale del patrimonio e delle attività culturali e finanziato dall'Unione europea – Next Generation EU.

BIBLIOGRAFIA

Bordalejo, B., Gibbings, L., North, R., & Robinson, P. (2021). Making an edition in an app. *Digital Medievalist*, *14*(1), 3, 1–32. https://doi.org/10.16995/dm.8067

Buzzoni, M., Cucurnia, D., Fenu, C., Rosselli Del Turco, R., & Tancredi, G. (2024). Progetto di edizione genetica digitale del Canzoniere manoscritto di U. Saba (1919–20). In A. Di Silvestro & D. Spampinato (Eds.), *Me.Te. Digitali. Mediterraneo in rete tra testi e contesti, Proceedings del XIII Convegno Annuale AIUCD2024* (pp. 215–220). AIUCD. https://doi.org/10.6092/unibo/amsacta/7927

Calamai, S., Piccardi, D., Pretto, N., Candeo, G., Stamuli, M. F., & Monachini, M. (2022). Not just paper: Enhancement of archive cultural heritage. In D. Fišer & A. Witt (Eds.), *CLARIN: The infrastructure for language resources* (pp. 647–666). De Gruyter. https://doi.org/10.1515/9783110767377-025

Galka, S., & Vogeler, G. (2025). Annotating, projecting, and interpreting named entities in digital scholarly editions with LLMs. *International Journal of Digital Humanities*, *7*, 483–509. https://doi.org/10.1007/s42803-025-00114-8

Hiltmann, T., Dröge, M., Dresselhaus, N., Grallert, T., Althage, M., Bayer, P., Eckenstaler, S., Mendi, K., Schmitz, J. M., Schneider, P., Sczeponik, W., & Skibba, A. (2025). *NER4all or context is all you need: Using LLMs for low-effort, high-performance NER on historical texts. A humanities informed approach*. arXiv. https://doi.org/10.48550/arXiv.2502.04351

Kim, S., Baudru, J., Ryckbosch, W., Bersini, H., & Ginis, V. (2025). *Early evidence of how LLMs outperform traditional systems on OCR/HTR tasks for historical record*s. arXiv. https://doi.org/10.48550/arXiv.2501.11623

Levchenko, M. A. (2025). Evaluating LLMs for historical document OCR: A methodological framework for digital humanities. In *Proceedings of the First Workshop on Natural Language Processing and Language Models for Digital Humanities* (pp. 75–85). https://doi.org/10.26615/978-954-452-106-6-007

Moffatt, K., Sharren, K., & Levy, M. (2020). Modeling the audio edition with Mavis Gallant's 1984 reading of "Grippes and Poche". *ESC: English Studies in Canada, 46*(2–4), 141–159. https://doi.org/10.1353/esc.2020.a903555

Petridou, E., & Tiktopoulou, K. (2022). Reaching out to the reader: The audio guide as a tool in digital genetic editions. *Textual Cultures, 15*(1), 126–134. https://doi.org/10.14434/tc.v15i1.34505

Pollin, C., Fischer, F., Sahle, P., Scholger, M., & Vogeler, G. (2025). When it was 2024 – Generative AI in the field of digital scholarly editions. *Zeitschrift für digitale Geisteswissenschaften*, *10*. https://doi.org/10.17175/2025_008

Pollin, C., Steiner, C., & Zach, C. (2023). New ways of creating research data: Conversion of unstructured text to TEI XML using GPT on the correspondence of Hugo Schuchardt with a web prototype for prompt engineering. *FORGE 2023*. https://doi.org/10.5281/zenodo.8425163

Valentini, C., Piccardi, D., Calamai, S., & Stamuli, M. F. (2023). Da cornice a soggetto. Il documento sonoro nell'infrastruttura Archivio Vi.Vo. *Archivi*, *18*(1), 88–125. https://hdl.handle.net/11365/1223925

La mostra *Coerenza In Coerenza* (1984) oltre la fotogrammetria: ricostruzione digitale tramite reti neurali

FILIPPO YAHIA MASRI[1], MARIA CARMELITI[2]

[1] UNIVERSITÀ DI GENOVA, ITALIA, FILIPPOYAHIA.MASRI@EDU.UNIGE.IT
[2] ACCADEMIA DI BELLE ARTI STATALE DI VERONA, ITALIA, MARIA.CARMELITI@ACCADEMIABELLEARTIVERONA.IT

ABSTRACT (ITALIANO)

Il contributo presenta un approccio sperimentale alla ricostruzione digitale della mostra *Coerenza In Coerenza*, allestita presso la Mole Antonelliana di Torino nel 1984, affrontando criticamente i limiti della fotogrammetria applicata a dataset archivistici ridotti, non strutturati e privi di finalità documentarie. La collettiva, curata da Germano Celant, rappresentò un episodio di assoluto rilievo nel panorama artistico nazionale e internazionale, configurandosi come un momento di storicizzazione e rilancio dell'Arte Povera. Inserita nel complesso contesto culturale torinese degli anni Ottanta, la mostra riunì dodici artisti sotto uno spazio simbolo del capoluogo piemontese ma allora privo di una vocazione museale stabile. Nonostante la rilevanza storica e critica, dell'evento non esiste una documentazione fotografica professionale: la memoria visiva dell'allestimento sopravvive unicamente attraverso un corpus frammentario di fotografie private, prive di metadati, con scarsa sovrapposizione e condizioni di ripresa eterogenee. A partire da questo materiale d'archivio, lo studio confronta diversi flussi di lavoro basati su immagini, tra cui fotogrammetria, NeRF e 3DGS. I risultati evidenziano come il 3DGS rappresenti la soluzione più efficace nel restituire continuità volumetrica, qualità luminosa e percezione spaziale dell'allestimento, configurandosi non come replica metrica, ma come *reenactment* digitale dell'evento. Il modello risultante è stato integrato in una ricostruzione digitale della Mole Antonelliana e costituisce la base per futuri sviluppi in ambiente VR, orientati alla fruizione immersiva e alla valorizzazione del patrimonio culturale immateriale legato a mostre effimere.
Parole chiave: museologia; *reenactment*; *Structure from Motion*; rappresentazione 3D; *Virtual Reality*

ABSTRACT (ENGLISH)

The exhibition *Coerenza In Coerenza* (1984) beyond photogrammetry: digital reconstruction using neural networks

This contribution presents an experimental approach to the digital reconstruction of the exhibition *Coerenza In Coerenza*, held at the Mole Antonelliana in Turin in 1984, critically addressing the limitations of photogrammetry applied to small, unstructured archival datasets without documentary purposes. The collective exhibition, curated by Germano Celant, represented an episode of great importance in the national and international art scene, constituting a moment of historicisation and relaunch of Arte Povera. Set in the complex cultural context of Turin in the 1980s, the exhibition brought together twelve artists in a space that was symbolic of the Piedmontese capital but at the time lacked a stable museum vocation. Despite its historical and critical importance, there is no professional photographic documentation of the event: the visual memory of the exhibition survives only through a fragmentary corpus of private photographs, lacking metadata, with little overlap and heterogeneous shooting conditions. Starting from this archival material, the study compares different image-based workflows, including photogrammetry, NeRF and 3DGS. The results show that 3DGS is the most effective solution for restoring volumetric continuity, light quality and spatial perception of the exhibition, configuring itself not as a metric replica, but as a digital reenactment of the event. The resulting model has been integrated into a digital reconstruction of the Mole Antonelliana and forms the basis for future developments in the VR environment, aimed at immersive enjoyment and the enhancement of the intangible cultural heritage linked to ephemeral exhibitions.
Keywords: museology; *reenactment*; *Structure from Motion*; rappresentazione 3D; *Virtual Reality*

1. LA MOSTRA COERENZA IN COERENZA: CONTESTO STORICO-ARTISTICO

La ricostruzione digitale delle mostre passate è un tema attuale per una necessità storico-critica: poter visualizzare e partecipare a eventi ormai conclusi costituisce una risorsa imprescindibile per lo studio della storia dell'arte. Nonostante un continuo aggiornamento delle tecnologie, gli strumenti più idonei a questa operazione non sono ancora stati individuati. Tale condizione è motivata dal fatto che la fotogrammetria, mezzo principe per le ricostruzioni digitali, sia ancorata a dati e materiali difficilmente reperibili per le esposizioni precedenti agli anni Novanta. I documenti sono spesso fotografie amatoriali o scatti parziali, difficilmente utilizzabili. I sistemi neurali permettono però di lavorare attraverso materiali sino ad ora reputati come collaterali, basando la propria correttezza sullo studio filologico dell'evento di interesse. L'analisi di cataloghi e opere d'arte risulta dunque imprescindibile per un lavoro di ibridazione totale tra discipline umanistiche e tecniche. Due problemi sorgono però rispetto all'utilizzo delle fotografie amatoriali sopra citate: la questione della reperibilità e della conservazione. I materiali di uso personale, di carattere effimero, per quanto preziosi, poco si prestano a lavori di riordino a decenni di distanza. Lo studio illustra tuttavia un episodio felice di ricostruzione digitale, riguardo una mostra cardine degli anni Ottanta, di respiro internazionale, *Coerenza In Coerenza*, del 1984 (Celant, 1984). La collettiva costituì un momento di assoluta rilevanza, di cui non vi era testimonianza fotografica. Durante alcune ricerche presso vari archivi privati, è stato però ritrovato del materiale inedito, testato attraverso l'uso di sistemi neurali. Base di partenza per un lavoro dalle molteplici implementazioni, il progetto presenta una nuova modalità di ricostruzione e rappresentazione digitale. *Coerenza In Coerenza* rappresenta un'opportunità per la perlustrazione di tutte le mostre organizzate presso la Mole Antonelliana negli anni Ottanta, prima della conversione dello spazio in Museo del Cinema. Quest'ultimo è stato coinvolto nella realizzazione del lavoro con lo scopo di rendere la mostra digitale visibile all'interno del palinsesto dello spazio CineVR del Museo del Cinema di Torino. Le fotografie utilizzate per il lavoro sono opera di Carlo Viano, curatore di alcuni degli eventi culturali più importanti della Torino degli anni Ottanta e oggi preziosa memoria storica.
«Il Giornale dell'Arte», sul numero 12 del maggio 1984, aprì la recensione alla mostra *Coerenza In Coerenza* con il titolo *Celant e i suoi 12 apostoli* (Torriani, 1984). Il riferimento al critico come a una figura divina, attorno alla quale ruotano gli artisti esposti, era calzante. Germano Celant aveva assemblato un gruppo senza precedenti nell'arte italiana, per coesione ed eterogeneità. Nel 1984 sembrava però esaurita la parabola più intensa della recente arte contemporanea nazionale e, per diversi aspetti, internazionale. La volontà alla base della mostra era quella di rilanciare personalità ormai cristallizzate nel decennio precedente, senza però imbrigliarle alla definizione a cui furono legati i loro successi tra la fine degli anni Sessanta e i primissimi anni Settanta: «perché ognuno prese strade individuali all'inizio degli anni '71 e '72, quando l'arte povera iniziò ad essere storicizzata» (Torriani, 1984). Il fine era altrettanto esplicito: «un momento di storicizzazione dell'Arte Povera composta di 12 ritratti autonomi». Celant, con l'aiuto di Ida Gianelli e Paolo Fossati, organizzò una mostra diffusa per la città di Torino, con opere – recenti e non – degli artisti coinvolti: Giovanni Anselmo, Alighiero Boetti, Pier Paolo Calzolari, Luciano Fabro, Jannis Kounellis, Mario Merz, Marisa Merz, Giulio Paolini, Pino Pascali, Giuseppe Penone, Michelangelo Pistoletto, Gilberto Zorio.
Coerenza In Coerenza definisce una generazione di artisti legandoli a Torino come città epicentro del movimento poverista, scindendo l'identità geografica da quella artistica e riunendo il gruppo sotto la struttura simbolo della città. La scelta della Mole Antonelliana venne certamente condizionata dall'assenza di uno spazio museale attivo, ma questa risultò vincente in termini di ricezione. Slegare gli artisti da un'istituzione pubblica o privata permise di indirizzare l'attenzione verso Torino tutta, riunendo le opere in uno spazio simbolo della città ed evitando associazioni tra l'evento e una specifica programmazione – o personalità – culturale. Un momento che chiariva l'identità degli artisti esposti e il loro legame, ratificandone il luogo di nascita: Torino. Lo spazio si sviluppava su due livelli, di cui uno interrato. Per il piano superiore, l'altezza della struttura consentiva agli artisti di sperimentare con verticalità difficilmente ripetibili.
La mostra, organizzata talvolta con strappi importanti, talvolta con grande linearità, non aveva interesse nel chiarire la carriera degli artisti o metterne in luce un determinato periodo. Una collettiva aperta al pubblico ma dalla dimensione privata, utile a chi era già coinvolto negli eventi dell'Arte Povera. Un episodio che includeva tutte le personalità imprescindibili per Celant, limitando all'Arte Povera l'unico risultato per la combinazione di arte contemporanea e Torino. Questa condizione, solo superficialmente corretta, irritò diversi artisti. Una lettera aperta all'Assessore alla Cultura Balmas da parte di artisti tra i quali Marco Gastini, Piero Gilardi e Giorgio Griffa lamentava l'esclusione dalla mostra e una

programmazione miope da parte della politica culturale del Comune. La mostra era effettivamente percepita come compendio dell'esperienza torinese per i due decenni '60 e '70, ma la volontà era quella di rivitalizzare la corrente partendo da Torino, città seminale per l'esperienza. Celant era stato chiaro nella scelta di cosa rappresentare: «una situazione che è nata qui o che, quando non è nata qui, vi è passata al 90%» (Torriani, 1984).
L'intento della mostra venne raggiunto. *Coerenza In Coerenza* ricevette recensioni positive sulle maggiori riviste e rotocalchi nazionali. Ma fu il riscontro internazionale a coronare l'esperienza. L'esposizione venne riproposta, in misura rivista rispetto allo spazio, ai Palacios de Cristal y de Velázquez a Madrid e al MoMA PS1 di New York.
Rispetto alle esperienze estere, la mostra presentava un problema di difficile risoluzione: l'assenza di materiale fotografico utile a definirne una memoria. La chiusura della GAM negli anni Ottanta aveva infatti spinto il Comune a cercare nuove sedi per le attività culturali, senza però avere personale specializzato per le varie mansioni che ruotavano attorno a tali eventi (Foot & Lumley, 2007). L'assenza di un fotografo sembrava aver relegato l'episodio a un passato irrecuperabile, ma un'attenta ricerca in diversi archivi privati torinesi ha portato alla luce documenti inediti imprescindibili per un lavoro di digitalizzazione.

2. LA RICOSTRUZIONE VOLUMETRICA DELL'ALLESTIMENTO ESPOSITIVO

La sfida metodologica risiede nell'utilizzo di immagini d'archivio non concepite per scopi fotogrammetrici, caratterizzate da una bassa risoluzione, parametri di camera ignoti e, soprattutto, una scarsa sovrapposizione e un numero assai limitato di scatti. In uno scenario informativo così lacunoso, le tecniche tradizionali di fotogrammetria mostrano limiti evidenti nella completezza della ricostruzione e nella gestione delle superfici poco strutturate.
Per affrontare tali problematiche, la ricerca ha sperimentato e confrontato diversi flussi di lavoro, basati sulle tecnologie della *Structure from Motion* (SfM) come fotogrammetria, *Neural Radiance Fields* (NeRF) e *3D Gaussian Splatting* (3DGS) (Mildenhall et al., 2021; Kerbl et al., 2023).
I risultati ottenuti dimostrano come la tecnologia *3D Gaussian Splatting* abbia rappresentato la soluzione più efficace per la ricostruzione da un dataset archivistico limitato, eccellendo non solo nella coerenza visiva e nella resa realistica dei materiali, ma soprattutto nella capacità di generare e restituire l'aspetto dell'allestimento originario, anche a livello di illuminazione e atmosfera.

3. STATO DELL'ARTE

La documentazione e la ricostruzione tridimensionale di spazi effimeri (quali siti archeologici compromessi nel corso dei secoli o ad oggi non più esistenti) si inseriscono in un filone di ricerca consolidato che vede l'uso di immagini d'archivio come fonte primaria di dati. In questi contesti le fotografie storiche sono considerate testimonianze essenziali per la storiografia e la tutela del patrimonio culturale materiale e immateriale (Al Khalil & Grussenmeyer, 2019; Maiwald et al., 2017; Bevilacqua et al., 2019).
Per decenni, le tecniche di SfM hanno rappresentato lo standard per la creazione di repliche digitali metricamente accurate. Questi flussi di lavoro estraggono informazioni geometriche e cromatiche da serie di fotografie parzialmente sovrapposte, generando nuvole di punti dense e mesh poligonali; per ottenere risultati affidabili, i dataset richiedono generalmente una sovrapposizione tra le immagini compresa tra il 60% e il 90% nelle sequenze di scatto (Croce et al., 2024).
Tuttavia, l'applicazione della fotogrammetria tradizion bassa ale a dataset archivistici presenta criticità strutturali: la scarsa risoluzione delle immagini, l'assenza di metadati della camera e, soprattutto, l'insufficiente sovrapposizione tra gli scatti (*overlapping ratio*) compromettono una ricostruzione completa o accurata (Al Khalil & Grussenmeyer, 2019; Bevilacqua et al., 2019).
Nonostante tali limitazioni, la letteratura scientifica ha ampiamente esplorato l'uso di immagini storiche d'archivio e contenuti generati dagli utenti per la ricostruzione di siti perduti o danneggiati, affrontando le medesime criticità di geometria della camera ignota, risoluzione e scarsa sovrapposizione.
Uno degli ambiti di applicazione più rilevanti riguarda la documentazione di siti archeologici o edifici storici compromessi durante eventi bellici o modificati nei secoli per adattarli ad altri fini. Nel caso di Bosra al-Sham, in Siria, flussi di lavoro fotogrammetrici sono stati utilizzati per documentare la Porta Occidentale e la Moschea Al Omari, duramente colpiti dalla guerra. Sebbene le immagini d'archivio non rispettassero i requisiti di sovrapposizione ideali per le tecnologie di SfM, è stato possibile generare

nuvole di punti dense e fotomosaici bidimensionali utili alla documentazione e all'analisi (Al Khalil & Grussenmeyer, 2019).
Analogamente, nel caso del Castello di Beaufort (Libano) le immagini d'archivio dell'esercito francese (1935-1937) sono state integrate con rilievi moderni e immagini aeree per una restituzione 3D delle strutture sepolte o distrutte, supportando la pianificazione degli scavi archeologici (Grussenmeyer & Yasmine, 2004).
Altri approcci propongono l'orientamento di immagini d'archivio su modelli 3D esistenti (ottenuti tramite laser scanning o fotogrammetria) per recuperare volumi o texture perdute. Un esempio significativo è presentato dal Palazzo di Cosimo de' Medici a Livorno: gravemente danneggiato a seguito di eventi bellici, l'edificio è stato oggetto di una ricostruzione volumetrica integrando immagini terrestri degli anni '30 con modelli 3D dello stato attuale della Fortezza Vecchia. La principale criticità è stata superata tramite l'identificazione manuale di c*ontrol points* su dettagli architettonici invariati (Bevilacqua et al., 2019).
In altri casi, invece, le fotografie d'archivio (anche scansioni da libri o immagini web) sono state orientate su modelli 3D attuali, permettendo ai restauratori di visualizzare porzioni di affreschi oggi degradate o mancanti, come illustrato negli studi di Bevilacqua et al. (2018).
Parallelamente, questo settore ha tratto grande beneficio anche dalle enormi collezioni di immagini disponibili online. Sistemi avanzati di SfM sono stati testati con successo su migliaia di fotografie scattate da turisti e fotocamere diverse, dimostrando la possibilità di ricostruire noti siti culturali (Snavely et al., 2006).
Negli ultimi anni, i progressi introdotti da Mildenhall hanno portato allo sviluppo dei sistemi NeRF, che si sono affermati come una promettente alternativa alla fotogrammetria per la creazione di modelli 3D a partire da immagini. I NeRF rappresentano la scena come una funzione volumetrica continua, permettendo una resa fotorealistica dipendente dal punto di vista dell'osservatore (*view-dependent*).
Studi comparativi nel settore dei beni culturali dimostrano che tali approcci superano la fotogrammetria tradizionale nel preservare la completezza del modello e la descrizione dei materiali anche in presenza di dataset ridotti o a bassa risoluzione (Croce et al., 2024). Tuttavia, i lunghi tempi di addestramento e la generazione di nuvole di punti intrinsecamente rumorose limitano ancora la loro efficacia in contesti puramente metrici.
Il 3DGS, invece, è un'alternativa in tempo reale che modella le scene utilizzando insiemi sparsi di primitive gaussiane. Questo metodo consente una visualizzazione efficiente, offrendo prestazioni migliorate in termini di aspetto dei materiali, percezione della profondità e gestione della luminosità (Kerbl et al., 2023). È stato osservato che, in condizioni di dati ridotti o bassa risoluzione e artefatti riflettenti, questi nuovi sistemi offrono una maggiore completezza del modello e fedeltà dei materiali rispetto alla fotogrammetria, sebbene quest'ultima rimanga superiore in termini di pura accuratezza metrica su superfici opache e ben strutturate (Clini et al., 2025).
il presente studio si propone di ricreare virtualmente l'ambiente della mostra *Coerenza In Coerenza*, svoltasi negli spazi della Mole Antonelliana nel 1984. La ricerca affronta le criticità tipiche dei dataset non strutturati, disponendo di soli 49 scatti d'archivio utilizzabili. Per superare i limiti della fotogrammetria, che spesso fatica a garantire risultati completi in presenza di scarse sovrapposizioni e geometrie di camera ignote, la metodologia qui adottata prevede un confronto sperimentale con i sistemi di rendering neurale citati, per valutare quale di queste tecnologie sia in grado di restituire con maggiore fedeltà non solo la volumetria dello spazio, ma anche l'atmosfera e l'aspetto luminoso dell'allestimento originario.
L'obiettivo finale è trasformare il frammento fotografico in un'esperienza immersiva, un *reenactment* digitale di un momento perduto della storia espositiva torinese

4. METODOLOGIA

Il metodo adottato per la ricostruzione volumetrica dell'allestimento espositivo si basa su un approccio sperimentale e comparativo, volto a valutare l'efficacia di diverse tecnologie di ricostruzione tridimensionale applicate a un dataset archivistico limitato e non strutturato.
Il punto di partenza della ricerca è costituito da un corpus di immagini provenienti da un archivio fotografico privato. Dal materiale disponibile sono state recuperate complessivamente 112 fotografie – amatoriali – relative alla mostra in oggetto. Le immagini documentano sia le opere esposte sia lo spazio architettonico che ha ospitato l'evento, in particolare durante le fasi di allestimento.
A seguito di una fase preliminare di analisi qualitativa e selezione, volta a valutare la leggibilità spaziale e la presenza di informazioni utili alla ricostruzione tridimensionale, solo 49 fotografie sono risultate

effettivamente utilizzabili per la sperimentazione. Il dataset finale si caratterizza per una marcata eterogeneità in termini di risoluzione, inquadrature e condizioni di illuminazione, nonché per l'assenza di metadati relativi alla geometria di camera e per una sovrapposizione visiva estremamente ridotta tra gli scatti.
Come descritto in precedenza, alla luce delle problematiche individuate, la metodologia adottata ha previsto l'implementazione e il confronto di tre diversi flussi di lavoro basati su immagini.
La fotogrammetria è stata utilizzata come termine di riferimento, al fine di valutare i limiti delle tecniche SfM in un contesto archivistico caratterizzato da dataset ridotti e scarsamente sovrapposti. Tale approccio ha mostrato difficoltà significative sia nell'orientamento delle immagini sia nella generazione di una nuvola di punti sufficientemente densa e coerente, con lacune evidenti o creazione di geometrie inesistenti nelle superfici poco strutturate e nelle zone a illuminazione complessa.
Successivamente, il dataset è stato sottoposto a tecniche di rendering neurale. I sistemi NeRF non sono riusciti a ricostruire la scena a causa della scarsità di immagini di partenza. Non è un risultato legato al numero di input ma quanto al valore di sovrapposizione, quasi nullo.
Il 3DGS, invece, è risultato l'approccio più efficace nel contesto specifico della ricerca. La tecnologia dietro la rappresentazione della scena ha permesso di gestire in modo più robusto la scarsità di dati, offrendo una ricostruzione volumetrica continua e una resa visiva particolarmente efficace degli aspetti luminosi e materici dell'allestimento (vd. Fig.1).
Rispetto agli altri flussi di lavoro testati, il 3DGS ha dimostrato una maggiore capacità di restituire la percezione spaziale complessiva dell'ambiente espositivo, includendo riflessi e gradienti di luce. Questo aspetto si è rivelato centrale per l'obiettivo del progetto, che non mira esclusivamente alla produzione di un modello metricamente accurato, ma alla ricostruzione dell'esperienza percettiva e atmosferica della mostra.

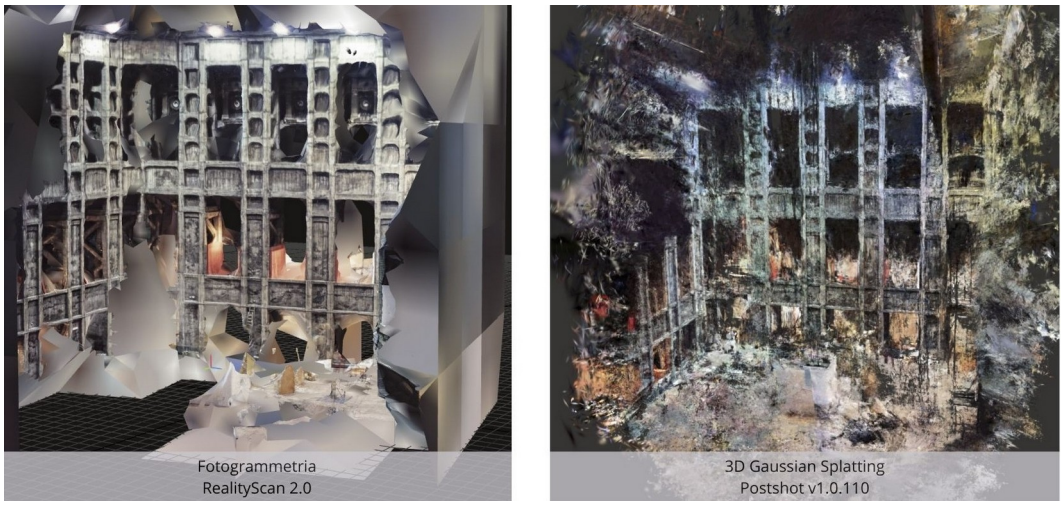


Figura 1. Confronto tra i risultati ottenuti utilizzando la fotogrammetria, a sinistra, e il 3D Gaussian splatting, a destra. L'immagine evidenzia i limiti della fotogrammetria in condizioni di scarsa sovrapposizione e la maggiore capacità del 3DGS di restituire coerenza volumetrica.

5. RISULTATI E FUTURE WORKS

Il metodo adottato si colloca in una prospettiva che privilegia la dimensione esperienziale e interpretativa della restituzione digitale. La ricostruzione volumetrica proposta non ambisce a costituire una replica oggettiva e definitiva dello spazio espositivo, ma si configura come un *reenactment*, capace di trasformare il frammento fotografico in un ambiente immersivo e navigabile.
In tale prospettiva, il 3DGS si è rivelato particolarmente efficace nel restituire non solo la volumetria dell'allestimento, ma anche le condizioni luminose, percettive e spaziali che hanno caratterizzato l'evento originario. La ricerca ha tuttavia superato la fase di sola ricostruzione, estendendo il risultato verso una sperimentazione orientata alla fruizione e alla valorizzazione del patrimonio culturale.
Il modello gaussiano dell'allestimento è stato infatti integrato all'interno di una ricostruzione digitale della Mole Antonelliana, realizzata in ambiente *Blender v. 4.4.3* mediante l'utilizzo di un'estensione dedicata alla gestione dei modelli gaussiani (*KIRI Engine v4.1.0*). Opportunamente scalato e collocato nello spazio architettonico.

Le prospettive future vedono una successiva implementazione in *Unity* per consentire la creazione di *hotspot* interattivi in corrispondenza delle opere esposte così da garantire l'accesso a contenuti informativi e iconografici.
La progettazione dell'ambiente si svolgerà in funzione della realtà virtuale in modo da trasformare il modello in un'esperienza immersiva, in cui spazio, memoria e informazione dialogano tra loro: il lavoro apre dunque nuove prospettive per la valorizzazione di archivi fotografici e per la conservazione e trasmissione del patrimonio culturale legato a eventi espositivi effimeri.

BIBLIOGRAFIA

Al Khalil, O., & Grussenmeyer, P., 2019. 2d & 3d reconstruction workflows from archive images, case study of damaged monuments in bosra al-sham city (syria). The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLII-2/W15, 55–62. https://doi.org/10.5194/isprs-archives-XLII-2-W15-55-2019

Bevilacqua, M. G., Caroti, G., Martínez-Espejo Zaragoza, I., & Piemonte, A., 2018. Orientation of archive images on 3D digital models of painted vaults: An interesting tool for restorers. https://doi.org/10.1007/s12518-018-0212-8

Bevilacqua, M. G., Caroti, G., Piemonte, A., & Ulivieri, D., 2019. Reconstruction of lost architectural volumes by integration of photogrammetry from archive imagery with 3-d models of the status quo. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLII-2/W9, 119–125. https://doi.org/10.5194/isprs-archives-XLII-2-W9-119-2019

Celant, G., I., 1984. Coerenza in coerenza. Mondadori, Milano.

Clini, P., Perdicca, R., Nespeca, R., Angeloni, R., Coppetta, L., 2025. Comparing MVS and Gaussian Splatting for the 3D Reconstruction of Reflective and Texture-less Cultural Heritage Artifacts. DIGITAL HERITAGE (2025). Eurographics-The European Association for Computer Graphics. https://doi.org/10.2312/dh.20253144

Croce, V., Billi, D., Caroti, G., Piemonte, A., Luca, L. D., Véron, P., Croce, V., Billi, D., Caroti, G., Piemonte, A., Luca, L. D., & Véron, P., 2024. Comparative Assessment of Neural Radiance Fields and Photogrammetry in Digital Heritage: Impact of Varying Image Conditions on 3D Reconstruction. Remote Sensing, 16(2). https://doi.org/10.3390/rs16020301

Foot, J., Lumley, R., 2007, *Le città visibili. Spazi urbani in Italia, culture e trasformazioni dal dopoguerra a oggi*, Il Saggiatore, Milano.

Grussenmeyer, P., & Yasmine, J., 2004. Photogrammetry for the preparation of archaeological excavation. A 3d restitution according to modern and archive images of Beaufort Castle landscape (Lebanon). XXth Congress of the International Society for Photogrammetry and Remote Sensing (ISPRS), Jul 2004, Istanbul, Turkey. pp.809-814. ⟨halshs-00264014⟩

Kerbl, B., Kopanas, G., Leimkuehler, T., & Drettakis, G., 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. ACM Trans. Graph., 42(4), 139:1-139:14. https://doi.org/10.1145/3592433

Maiwald, F., Vietze, T., Schneider, D., Henze, F., Münster, S., & Niebling, F., 2017. Photogrammetric analysis of historical image repositories for virtual reconstruction in the field of digital humanities. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLII-2/W3, 447–452. https://doi.org/10.5194/isprs-archives-XLII-2-W3-447-2017

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., & Ng, R., 2021. NeRF: Representing scenes as neural radiance fields for view synthesis. Commun. ACM, 65(1), 99–106. https://doi.org/10.1145/3503250

Snavely, N., Seitz, S. M., & Szeliski, R., 2006. Photo tourism: exploring photo collections in 3D. ACM Trans. Graph. 25, 3 (July 2006), 835–846. https://doi.org/10.1145/1141911.1141964

Torriani, F., 1984. Celant e i suoi dodici apostoli. Il Giornale dell'Arte, n. 12, maggio, Torino, 34

[3/A] AI e NLP per testi e recupero dell'informazione

Creating a new gender-fair Italian morphological classifier based on UmBERTo fine-tuning

IRENE CAIAZZO[1], GIOVANNA MARIA DIMITRI[2]

[1] UNIVERSITÀ PER STRANIERI DI SIENA, ITALIA – I.CAIAZZO@DOTTORANDI.UNISTRASI.IT
[2] UNIVERSITÀ DEGLI STUDI DI MILANO, ITALIA - GIOVANNA.DIMITRI@UNIMI.IT

ABSTRACT (ENGLISH)

The paper describes the design of a fine-tuning procedure for a BERT model and the subsequent construction of an annotated dataset composed of texts from a gender-fair corpus and synthetic generated text for training. It presents the stages of automatic annotation and manual revision, as well as the definition of classes and labels designed to represent morphologically innovative gender-inclusive language in Italian strategies, such as - ə (schwa), -* (asterisk), -x, and other variants.
The goal is to train a deep learning–based morphological classifier capable of automatically annotating uses of gender-inclusive language realized through these strategies.
Keywords: POS annotation; morphological annotation; LLM fine-tuning; neo-morphemes; gender-fair Italian

ABSTRACT (ITALIANO)

Per la creazione di un classificatore morfologico di italiano "gender-fair" basato sul fine-tuning di UmBERTo. L'articolo descrive la progettazione del fine-tuning di un modello BERT e la successiva costruzione di un dataset annotato composto da testi prelevati da un corpus di linguaggio inclusivo e da testi sintetici generati per il training. Vengono illustrate le fasi di annotazione automatica e di revisione manuale, nonché la definizione di classi ed etichette progettate per rappresentare strategie morfologicamente innovative di linguaggio inclusivo, quali -ə (schwa), -* (asterisco), -x e altre varianti. L'obiettivo è addestrare un classificatore morfologico basato su deep learning in grado di annotare automaticamente gli usi del linguaggio inclusivo realizzati attraverso tali strategie.
Parole chiave: annotazione per parti del discorso; annotazione morfologica; LLM fine-tuning; neo-morfemi; italiano inclusivo

1. INTRODUCTION

The morphological classification task has been a consolidated research area in Natural Language Processing (NLP) for several decades. As far as contemporary standard Italian is concerned, this task can be considered largely solved, thanks to the availability of annotated linguistic resources and mature computational models, some tools are LinguA[1] (Dell'Orletta 2009, Attardi, Dell'Orletta 2009, Attardi et al. 2009), AnIta (Tamburini, Melandri 2013) Tint[2] (Palmero Aprosio, Moretti 2016 and 2018). Scholars have also started to get interested in different varieties of Italian, both from a diachronic perspective and in relation to the different components of the Italian diasystem (e.g. Pennacchioni, Zanzotto 2008, Tonelli et al. 2010, De Felice et al. 2018, Bolognesi et al. 2025), but at the moment gender-fair Italian - which present in some of its uses modifications in the morphological system - has yet to be taken into consideration even if corpora of these uses exist (Comandini 2021, Safina 2023, De Cesare 2024, Formato 2024).
The present study aims to illustrate the creation of a labelled training dataset to fine-tune a Large Language Model (LLM), specifically of UmBERTo[3] (Parisi et al. 2020), aimed at the automatic annotation of gender-fair linguistic strategies based on morphological processes. The objective is to prepare a dataset for the successive training to then determine the model's ability to recognize and classify innovative morphological forms that do not fully fit into the paradigms of standard Italian morphology.

[1] http://www.italianlp.it/demo/linguistic-annotation-tool/ (cons: 07/02/2026)
[2] https://dh.fbk.eu/research/tint/ (cons: 07/02/2026)
[3] https://github.com/musixmatchresearch/umberto (cons: 07/02/2026)

2. GENDER-FAIR ITALIAN

With the expression "gender-fair Italian" we refer to an emerging linguistic variety that is configured through contemporary uses aimed at overcoming the gender binary, in order to allow a linguistic representation that breaks free from the grammaticalized gender system of the Italian language. This term is used instead of the also widespread "inclusive language", as the latter, over time, has come to overlap with a plurality of instances of inclusivity and exclusion (for example, linked to racism, ableism, etc.), generating a conceptually heterogeneous category. Furthermore, the label "inclusive" has been the subject of criticism in the linguistic and sociolinguistic fields, both for its vague definition and for the ideological implications associated with it (Gheno 2023, Formato 2025).
The gender-fair variety of Italian can manifest itself through multiple linguistic strategies, some internal to the standard norm and others external to it (Comandini 2021). This work focuses specifically on innovative morphological strategies—here referred to as neomorphemes—which introduce new endings to mark or neutralize grammatical gender (Carlo 2024: 85–86). These strategies do not account for all forms of gender-inclusive language in Italian: syntactic and lexical alternatives, such as the use of specific nouns (e.g. collective nouns) or syntactic constructions (e.g. impersonal clauses) to avoid gender-marked forms, are also employed (MIUR 2018, Giovine 2024). However, such strategies fall outside the scope of this study, as their identification would require a different computational approach.

3. MODEL ARCHITECTURE AND WORKFLOW

The workflow adopted for fine-tuning the model is divided into the following phases:

- **Data collection**: selection of raw texts from IncluInstIT (Caiazzo et al. 2025) a corpus containing examples of gender-fair Italian. The corpus is composed of posts crawled from the social media platform Instagram through the "research for hashtag" method. The obtained dataset was pre-processed to exclude repetitions of documents and posts in foreign languages and also to include the texts located within images, which were retrieved with OCR procedures and later manually checked.
- **Preliminary automatic annotation**: application of an automatic annotation system for the categories of part-of-speech and for morphological traits, based on STANZA[4] (Qi et al. 2020), to a sample of the dataset (1209 files, that is 25% of the subcorpus *captions*[5]).
- **Manual review and enrichment of labels**: manual checking of annotations and introduction of new specific labels for gender-fair forms and for neomorphemes not covered by traditional schemes.
- **Construction of the annotation training dataset**: definition of an annotated reference dataset, used as a basis for supervised training.
- **Supervised fine-tuning of the UmBERTo**: training of the LLM on the training dataset in order to adapt its internal representations to the recognition of gender-fair morphological strategies.

The end result will be a fine-tuned linguistic model capable of performing morphological annotation that is more sensitive to the forms and strategies specific to gender-fair Italian than models trained exclusively on standard Italian data. This will enable the implementation of automatic annotations for corpora which, due to their focus on gender-fair language, are currently annotated through manual procedures.

3.1 DATASET DESCRIPTION AND SAMPLE COLLECTION

The corpus used for the construction of the labelled training dataset and for the model's training and testing phases was derived from IncluInstIT (Caiazzo et al. 2025) containing texts from social media, specifically designed to collect and analyze discourses related to gender-fair language in Italian (Table 1 for dimensions details, more specifically number of text files and tokens extracted).[6] This corpus is therefore characterized by a distinct presence of the linguistic strategies being analyzed, which allows for a significant amount of data relevant for both training and evaluating the model. Such coverage would be difficult to achieve using currently available generalist corpora, in which the presence of gender-fair forms

[4] https://stanfordnlp.github.io/stanza/ (cons: 07/02/2026)
[5] The corpus is composed of two different subcorpora, one for the captions, and the other of the texts retrieved in the posts' images. We chose to use in this initial phase only the subcorpus captions to have a more cohesive dataset.
[6] Data obtained using LancsBox 6.0 (Brezina et al. 2020).

is marginal or sporadic, both due to the relatively short lifespan of these expressions and their often limited use in certain specific contexts and genres.

Dataset	Number of documents	Tokens[6]
Corpus	4838 txt files	512,461
Training dataset (25% of the corpus)	1209 txt files	136,298

Table 1- Dimension of the corpus and of the training dataset

3.2 DEFINING LABELS FOR MANUAL ANNOTATION

The collected sample was automatically annotated using Stanza (Qi et al. 2020), with the output produced in CoNLL format, the standard representation for annotated linguistic data, in which each line corresponds to a token and its associated labels, separated by tabs. Each part-of-speech category is associated with a specific set of labels; for categories that include additional features beyond Gender and Number (e.g. PronType), these annotations were manually corrected where necessary in accordance with the Universal Dependencies (UD) framework.
The structure of the new labels was designed to optimize the chances of training success. For this reason, a two-step classification was developed: first, the "Non-binary" attribute was added[7] within the "Gender" class, assigned to all forms that do not conform to the binary distinction between male and female. Second, the "SNG" (*Strategie di neutralizzazione di genere,* gender neutralization strategies, Comandini 2021: 51) class has been added, where these forms are further subdivided into specific categories depending on the gender-fair strategy used. Furthermore, the "Ambiguo" attribute has also been added to the "Number" class in the manual annotation, to be able to flag cases in which the use of a neomorpheme leads to ambiguity in the number of the form. A token containing a gender-fair strategy will be annotated as:
Ragazz* → NOUN _ Gender=Non-binary|Number=Ambiguo|SNG=Ast

In Table 2, we show the labels created in the class "Incl" and the respective examples of use.

Tag	Strategies	Example
SNG=Ast	-*	tutt*
SNG =Ch	-@	tutt@
SNG =Spl	use of splitting	tutti/e, tutti/tutte,[8] tutti/e/ə, tutti/e/ə/*
SNG =Sw	-ə	tuttə/tutt3/tutt3[9]
SNG =X	-x	tuttx
SNG =U	-u	tuttu
SNG =altro	various strategies	tutty tutt. tutt\ tutt Tutt_ tutts

Table 2 - Labels for manual annotation and examples

[7] The labeling of these forms as "non-binary" raises considerations regarding the accuracy of this description in capturing the phenomenon. At present, however, it is regarded as the most appropriate option, as emphasizing the transcendence of the binary structure encompasses both their function as a generic form and their role as a specific form.
[8] These cases (extended splitting) have been annotated for the time being; however, we will evaluate whether they affect the accuracy of the annotation, for example, cases where the splitting is not used in an inclusive sense, and, if necessary, we will remove them from the training dataset.
[9] The "long schwa," a proposed plural form of schwa, appears in the corpus represented by both the symbol 3 (the IPA open-mid central unrounded vowel) and the symbol 3. The occurrence of the symbol 3 is plausibly attributable to its greater accessibility compared to the IPA character, which, like the standard schwa, was not originally available on most keyboards.

During the manual annotation, some particular results were noticed, such as examples of the presence of schwa interfering with the annotation of nearby words, e.g. “Ti sei appena iscrittə” (Table 3):

Annotation	Token	POS	Morphological feature
Automatic annotation	Ti	PRON	Clitic=Yes|Number=Sing|Person=2|PronType=Prs
	is	X	Foreign=Yes
	i	X	Foreign=Yes
	appena	ADV	
	iscrittə	VERB	Mood=Ind|Number=Sing|Tense=Pres|VerbForm=Fin
Manual annotation	Ti	PRON	Clitic=Yes|Number=Sing|Person=2|PronType=Prs|SNG=NONE
	sei	AUX	Mood=Ind|Number=Sing|Person=2|Tense=Pres|VerbForm=Fin|SNG=NONE
	appena	ADV	SNG=NONE
	iscrittə	VERB	Mood=Ind|Gender=Non-binary|Number=Sing|Tense=Past|VerbForm=Part|SNG=Sw

Table 3 - Example of tagging with automatic and manual annotation

Furthermore, some documents had to be removed from the sample, as they were considered examples of uses which could not be categorized as inclusive, for instance ironic uses, presence of the strategies on random words (e.g. inanimate nouns) and in words from other languages (as Neapolitan). These cases were removed as not representative of the phenomena that we wanted to analyze and, also, since they were hapax, they were not significant.
Some borderline cases were also observed. In certain instances, gender-inclusive strategies were applied to words that would not conventionally take them (e.g. individuə). These forms were nonetheless included in the sample, as we hypothesize that they carry semantic value that can being described as a deliberate attempt to foreground the non-binary identity of the referent, and that similar cases are likely to occur elsewhere in the corpus.
An additional issue concerns the use of ə and * as articles. We have found occurrences of these uses in the sample, but ə and * appear more often in isolation to refer to explanations of the strategy itself rather than to actual linguistic usage.

3.3 ANNOTATED TRAINING DATASET

At the end of the manual annotation and the review, morphological inclusive strategies were found in 199 documents (16,45% of the dataset). The distribution of the different strategies is not equal, some of them are more present than others, as shown in Table 4.

SNG value	Occurrences
Sw	311
Ast	90
Spl	49
Altro	8
U	3
X	2

Ch	1
Total	464

Table 4 - Occurrences for each strategy

The unbalanced distribution of the classes posed a potential issue for training; therefore, we adopted a twofold approach. First, we searched for additional instances of these strategies within a randomly selected 25% of the corpus. This procedure yielded only a small number of new examples: 3 for **-**u, 2 for **-**@, and 1 for **-**x. This outcome is unsurprising, given the limited diffusion of these strategies compared to **-**ə, **-***, and splitting forms, this last strategy benefits from being often still binary and are therefore more is perceived as standard and used by writers that want to be inclusive but within the current norms. Secondly, due to this data sparsity, we decided to generate an additional synthetic dataset for each strategy using ChatGPT 5.2.[10] It was crucial to keep the **-**x, **-**u, and **-**@ strategies separate from the "*altro*" category, as they are widely discussed in the literature on inclusive language strategies and in empirical research on gender-fair language use (Comandini 2021).
The example generation was carried out using the following prompt: "Scrivi 10 testi simili a quelli che ti sottopongo e che contengano la strategia di linguaggio inclusivo -x". The same procedure was applied to -u and -@.
After generating the initial texts, we observed that the model frequently combined the requested strategy with -ə and, in some cases, failed to apply it altogether. We therefore explicitly instructed the model to use it exclusively; however, this constraint did not result in a substantial improvement. The chatbot continued to exhibit this behaviour and also began to place the strategy in incorrect positions, for example: "Cucina e cultura: ricett@ tradizional@ possono essere reinterpretat@ in chiave inclusiv@ per tutt@ l@ appassionat@." It is evident that some applications of the strategy (tutt@, l@, appassionat@) are appropriate, whereas others (ricett@, tradizional@, reinterpretat@, inclusiv@) are not. Furthermore, when we required to stick to the way the strategy was used in the example, the output did not significantly change.
Overall, text generation required multiple prompt iterations to ensure that the LLM produced usable outputs and, consequently, we retained only the generated texts that consistently employed the required strategy and manually correct some of them when they contained a small number of localized errors within an otherwise correct text.
In total, we collected 53 new texts (24 with -x, 14 with -@, and 15 with -u), which were processed using the same automatic and manual annotation as the rest of the dataset.
At the end, the new numbers of occurrences for each strategy are shown in Table 5:

SNG value	Occurrences
Sw	323
Ast	90
Spl	55
Altro	8
X	36
U	37
Ch	37
Total	573

Table 5 - Occurrences for each strategy after generation

3.4 CONCLUSIONS

In this paper, we present the architecture designed for the creation and fine-tuning of a large language model for classification tasks aimed at identifying morphologically inclusive linguistic innovation, as well as the process adopted to construct the annotated training dataset and generated synthetic data. Since the training phase is still ongoing and the evaluation has not yet begun, our discussion of the results is

[10] The current quality of the synthetic dataset is considered to be improvable. In future work, we plan to incorporate texts generated by additional LLMs and to refine our prompt engineering strategies in order to increase the size of the dataset, enhance generation quality, and systematically compare the models' ability to produce texts employing these strategies.

necessarily speculative. Nevertheless, we already outline several possible limitations and interesting points of discussion during the system design:

- A limited ability to generalize to gender-fair strategies that are not represented in the training data.
- Sensitivity to non-normative or inconsistent uses of gender-fair forms (e.g., "individuə").
- An inability to distinguish between genuine and ironic uses of gender-fair language.
- The continued need for manual annotation to determine whether a given usage is generic or specifically nonbinary.

At the same time, we emphasize the importance of developing this tool, which will facilitate the handling of and provide deeper insight into the Italian gender-fair variety, whose use has been emerging for several years now.

REFERENCES

Attardi G. & Dell'Orletta F. "Reverse Revision and Linear Tree Combination for Dependency Parsing". In NAACL-HLT 2009 – North American Chapter of the Association for Computational Linguistics – Human Language Technologies (Boulder, Colorado, June 2009). Proceedings, pp. 261 – 264. Association for Computational Linguistics, 2009.

Attardi G., Dell'Orletta F., Simi M. & Turian J. "Accurate Dependency Parsing with a Stacked Multilayer Perceptron". In Proceedings of EVALITA 2009 – Evaluation of NLP and Speech Tools for Italian 2009 (Reggio Emilia, Italy, December 2009).

Brezina, V., Weill-Tessier, P., & McEnery, A. (2020). #LancsBox 5.x and 6.x. [software package]

Bolognesi, C., Cinini, A., Cutugno, P., Ferretti, M., & Chiarella, D. (2025). Towards a preschooler corpus of Italian: An experimental journey. Research Methods in Applied Linguistics, 4(3), 100252. https://doi.org/10.1016/j.rmal.2025.100252.

Caiazzo I., Dimitri G. M. & Tronci L. (2025). IncluInstIT: Un nuovo corpus per lo studio di linguaggio inclusivo nei social media. In Rebora S., Rospocher, M., Bazzaco, S. (Eds.) Diversità, Equità e Inclusione: Sfide e Opportunità per l'Informatica Umanistica nell'Era dell'Intelligenza Artificiale, Proceedings del XIV Convegno Annuale AIUCD2025.

Carlo, G. S. di. (2024). Is Italy ready for gender-inclusive language?: An attitude and usage study among Italian speakers. In Inclusiveness Beyond the (Non)binary in Romance Languages. Routledge.

Comandini, G. (2021). Salve a tuttə, tutt*, tuttu, tuttx e tutt@: L'uso delle strategie di neutralizzazione di genere nella comunità queer online: Indagine su un corpus di italiano scritto informale sul web. Testo e Senso, (23).

De Cesare, A. M. (2024). Sulla diacronia breve dello schwa nei testi di 'mediattivisti di movimento': forme, frequenza e funzioni socio-identitarie. Lingue e Culture dei Media v.8, n.1, 2022 (pp. 37-54).

De Felice, I., Dell'Orletta, F., Venturi, G., Lenci, A., & Montemagni, S. (2018). Italian in the Trenches: Linguistic Annotation and Analysis of Texts of the Great War. In Cabrio E., Mazzei A., & Tamburini F. (Eds.), Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018) (pp. 161–165). CEUR Workshop Proceedings.

Dell'Orletta F. "Ensemble system for Part-of-Speech tagging". In: Proceedings of EVALITA 2009 – Evaluation of NLP and Speech Tools for Italian 2009 (Reggio Emilia, Italy, December 2009).

Formato, F. (2024). Feminism, corpus-assisted research, and language inclusivity. Cambridge University Press. https://doi.org/10.1017/9781009236379.

Formato, F. (2025). Lingua, Linguaggio, Genere e Inclusività: Conceptualising Language and Gender in Italian. In Pzfalgraf F. (Ed), Gender-neutral, Gender-fair, Gender-inclusive, Palgrave Studies in language, Gender and Sexuality (pp. 145-166).

Gheno, V. (2023). Linguaggio ampio: una possibile strada verso la convivenza delle differenze. In Biemmi I. (Ed.) Quanti generi di diversità? Promuovere nuovi linguaggi, rappresentazioni e saperi per educare alle differenze e prevenire l'omofobia e la transfobia (pp. 57-68), Firenze University Press and USiena PRESS.

Giovine, M. (2024). Gender Visibility: Linguistic Strategies to Challenge Stereotypes in Italian. Rivista Italiana Di Filosofia Del Linguaggio, 161–171.

MIUR. (2018). Linee guida per l'uso del genere nel linguaggio amministrativo del MIUR.

Palmero Aprosio, A. & Moretti, G. (2018). Tint 2.0: An All-inclusive Suite for NLP in Italian. In Cabrio E., Mazzei A. & Tamburini F. (Eds.), Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018) (pp. 312–318). CEUR Workshop Proceedings.

Palmero Aprosio, A. & Moretti, G. (2018). Tint 2.0: An All-inclusive Suite for NLP in Italian. In Cabrio E., Mazzei A. & Tamburini F. (Eds.), Proceedings of the Fifth Italian Conference on Computational Linguistics CLiC-it 2018 (pp. 311–317). Accademia University Press.

Parisi, L., Francia, S. & Magnani, P. (2020). UmBERTo: an Italian Language Model trained with Whole Word Masking. https://github.com/musixmatchresearch/umberto.

Pennacchiotti, M. & Zanzotto, F. M. (2008). Natural Language Processing Across Time: An Empirical Investigation on Italian. *GoTAL,* S371–382. https://doi.org/10.1007/978-3-540-85287-2_36

Qi, P., Zhang. Y., Zhang, Y., Bolton, J. & Manning, C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In Association for Computational Linguistics (ACL) System Demonstrations. 2020.

Safina E. S. (2023) Siamo di fronte a una pericolosa deriva? Le strategie morfologiche di neutralizzazione del genere nell'italiano digitale tra opinione e uso, In Pietrini D. (Ed.) Lingua e Discriminazione. La lingua contro la discriminazione. (pp. 335-356) Martin Luther University Halle-Wittenberg.

Tonelli, S., Pianta, E., Delmonte, R. & Brunelli, M. (2010). VenPro: A Morphological Analyzer for Venetan. In Calzolari N., Choukri K., Maegaard B., Mariani J., Odijk J., Piperidis S., Rosner M. & Tapias D. (Eds.), Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). European Language Resources Association (ELRA).

Tamburini, F. & Melandri, M. (2012). AnIta: A powerful morphological analyser for Italian. In Calzolari N., Choukri K., Declerck T., Doğan M. U., Maegaard B., Mariani J., Moreno A., Odijk J. & Piperidis S. (Eds.), Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12) (pp. 941–947). European Language Resources Association (ELRA).

Assessing LLM Capabilities in Cultural Heritage: Recent Trends and Gaps in Benchmarks

REMO GRILLO [1,2], GIANMARCO SPINACI[1,2], LUKAS KLIC[2], GIOVANNI COLAVIZZA[1,3]

[1]DIPARTIMENTO DI FILOLOGIA CLASSICA E ITALIANISTICA, UNIVERSITÀ DI BOLOGNA, BOLOGNA, ITALY
[2]I TATTI - THE HARVARD UNIVERSITY CENTER FOR ITALIAN RENAISSANCE STUDIES, FIRENZE, ITALY
[3] DEPARTMENT OF COMMUNICATION AND THE CENTRE FOR DIGITAL AND COMPUTATIONAL HUMANITIES, UNIVERSITY OF COPENAGHEN, COPENAGHEN, DENMARK

ABSTRACT (ENGLISH)

Over the last two years (2024–2026), Large Language Models (LLMs) and multimodal models have shifted from experimental use to systematic integration in Cultural Heritage (CH) and Digital Humanities (DH) workflows across GLAM institutions and research projects. Benchmarks and leaderboards claim to measure model capabilities on domain-specific tasks (e.g., VQA, document classification, information extraction, knowledge synthesis, question answering). This paper critically examines these benchmarks as both technical and epistemic tools. While they enable comparability and can accelerate progress, these benchmarks can also yield misleading indicators in a domain where "correctness" is inherently contextual, interpretive, and inseparable from sources. We find that headline scores often overstate practical capability because they under-specify provenance, evidential support and ambiguity, and they seldom reflect end-to-end performance in real CH pipelines. By examining three task families—structured data, unstructured texts, and images—we observe a recurring gap between quantitative performance and operational quality and usefulness. Building on evidence from recent CH/DH benchmark initiatives, we outline actionable gaps and mitigation-oriented practices—such as safeguards against memorization and contamination—toward evaluations that better accommodate provenance, uncertainty, interpretive plurality, and end-to-end pipeline behaviour.
Keywords: benchmarks; LLM evaluation; question answering; KGQA; retrieval-augmented generation; cultural heritage

ABSTRACT (ITALIANO)

Negli ultimi due anni (2024–2026), i Large Language Models (LLM) e i modelli multimodali sono passati da un uso sperimentale a un'integrazione sistematica nei flussi di lavoro del Patrimonio Culturale (CH) e delle Digital Humanities (DH) presso istituzioni GLAM e progetti di ricerca. Benchmark e classifiche dichiarano di misurare le capacità dei modelli in compiti specifici del dominio (ad es. VQA, classificazione di documenti, estrazione di informazioni, sintesi della conoscenza, question answering). Questo articolo esamina criticamente tali benchmark come strumenti sia tecnici sia epistemici. Se da un lato consentono confrontabilità e possono accelerare il progresso, dall'altro possono produrre indicatori fuorvianti in un ambito in cui la "correttezza" è intrinsecamente contestuale, interpretativa e inseparabile dalle fonti. Osserviamo che i punteggi di sintesi spesso sovrastimano la capacità pratica perché specificano in modo insufficiente provenienza, supporto evidenziale e ambiguità, e raramente riflettono prestazioni end-to-end in pipeline CH reali. Analizzando tre famiglie di compiti—dati strutturati, testi non strutturati e immagini—rileviamo uno scarto ricorrente tra prestazioni quantitative e qualità. Sulla base di evidenze tratte da recenti iniziative di benchmark in CH/DH, delineiamo lacune operative e pratiche orientate alla mitigazione—incluse misure contro memorizzazione e contaminazione—per valutazioni più sensibili a provenienza, incertezza, pluralità interpretativa e comportamento end-to-end delle pipeline.
Parole chiave: benchmark; valutazione LLM; question answering; KGQA; retrieval-augmented generation; patrimonio culturale

1. INTRODUCTION

Mapping recent trends in LLM benchmarking for CH is not an exercise in taxonomy, but an attempt to clarify what today's evaluation practices are actually rewarding—and what they systematically miss. In many humanities contexts, LLM-based technologies are entering everyday activities: access to collections, description and indexing, research assistance, and knowledge organization. In this scenario, a benchmark is not merely a neutral objective description: it does not just check fit, because it rather

defines fit criteria, therefore implicitly stating what counts as "understanding" for a model, which practices are valued, and which remain invisible. The question, therefore, cannot be only "which model is better?", but rather "what are we really measuring, and at what epistemic cost?". This question has become more pressing for two converging reasons. First, the rise of retrieval-augmented and multimodal systems shifts attention from the model "in isolation" to the behavior of the system as a whole (retrieval, grounding, citations, tool use, structured outputs). Second, the saturation of general-purpose benchmarks has spurred 'frontier' initiatives that increase challenge by emphasizing closed-ended, verifiable questions designed to resist answer-by-retrieval. This is not a CH-specific observation: in mainstream NLP, models have reached near-ceiling performance on established suites faster than anticipated, eroding the discriminative value of static leaderboards and motivating calls for dynamic, adversarially constructed evaluation (Kiela et al., 2021; Bowman & Dahl, 2021). This trajectory—exemplified by *Humanity's Last Exam* (Phan & others, 2025) —is illuminating but not automatically transferable to CH, where provenance and the management of ambiguity must be integral to the task. The aim of this paper is twofold. On the one hand, it offers a reasoned synthesis of the benchmark lines most relevant to CH/DH, without claiming exhaustiveness given the rapid pace of change. On the other hand, it brings into focus the gap—both technical and epistemic—that helps explain why many current metrics are only weakly predictive of real research needs in cultural institutions. We seek to identify these key strands of research, the reasons why certain directions have emerged, and the open problems they are attempting to address.

2. SCOPE AND WORKING DEFINITIONS

A fundamental distinction informs our definition of benchmarks for systems utilizing LLMs in CH. We do not restrict ourselves to benchmarks that evaluate the model in isolation—in a closed-book fashion—but also include suites that assess end-to-end tasks or essential components of the CH pipeline, provided that the LLM is a core part of the process and the evaluation is specifically designed for humanities materials and tasks. By CH, here we refer to a pragmatic set of computational scenarios: museum collections, archives, library holdings, and art-historical contexts, alongside the digital infrastructures that represent them (knowledge graphs, metadata, thesauri, and search systems). For the purpose of this review, we distinguish three recurring ways LLMs enter CH workflows: (i) generation of narrative or descriptive text; (ii) extraction and normalization into structured fields; and (iii) tool- and retrieval-mediated use, where the model coordinates external resources (e.g., search, KG queries, citation/grounding). We adopt a tripartite division of data types—structured data, unstructured text, and images—viewed as the three primary material types typical in CH. Literature was identified via targeted explorations on arXiv and Google Scholar indexed papers via title and keyword searches[1]. From an initial seed of identified articles we proceeded with forward/backward references navigation to detect more. Three criteria guided the inclusion. A temporal criterion (works published between 2020 and 2026), a domain criterion (works explicitly targeting CH/DH materials and tasks), and a task criterion (evaluation frameworks and benchmarks rather than architectures). Together, these benchmarks are a representative snapshot of the landscape as they are not meant to be a comprehensive scenario of the field. This is also reflecting the rapid publication pace inherent in the area of LLMs benchmarks.

3. STRANDS AND TRENDS

3.1 KNOWLEDGE GRAPHS QUESTION ANSWERING

In CH/DH, structured representations—often knowledge graphs and linked-data models—are a key way to encode collections, authority data, events, and relationships for reuse and validation across systems (Tomasi, 2022). One prominent evaluation track focuses on Knowledge Graph Question Answering (KGQA): whether a system can translate domain questions into executable graph queries (often SPARQL) in a way that respects the target schema/ontology and modelling choices.
The rise of ontology-oriented text-to-SPARQL (in particular using the CIDOC-CRM) is a notable trend, evidenced by new benchmarks. First *TCRMQ* (Gounakis et al., 2023) and finally *CIDOC-QA-BENCHMARK* (Mountantonakis & Tzitzikas, 2025) addressed the structural difficulty of long, multi-hop paths in event-based data models by proposing ontology path-based prompting strategies, up to a scale of 10,000

[1] e.g., KGQA cultural heritage, VQA benchmark art, historical document LLM evaluation

templated questions across 20 query types explicitly using "depth" as an evaluation axis. These benchmarks increasingly operationalize "difficulty" as graph-traversal depth, but risking bias toward easily template-able metrics (execution accuracy) over scholarly relevance; moreover the questions in the dataset were, for scalability reasons, automatically generated and lack real world grounding and language.

Another CH trend re-centers evaluation on Competency Questions (CQs), originally intended as natural-language requirements used to scope and validate an ontology/KG (Grüninger & Fox, 1995). CQs can form a link between knowledge engineering and user practice. These question sets function as requirements, stemming from real research needs and implicitly defining "usefulness". Earlier examples like *ArCo* (Carriero et al., 2019), *Zeri* (Daquino et al., 2017), *HiCO (Daquino & Tomasi, 2015),* and *Polifonia CQs* (De Berardinis et al., 2023) prioritized authentic relevance reflecting specialist inquiry over sheer size. *PolifoniaCQ* explicitly uses extracted CQs as data for classification and validation, with the *IDEA* extraction pipeline (Polifonia Project, 2023) aims to provide a reproducible methodology. This shifts CH evaluation to requirements fidelity, ensuring benchmarks remain anchored to documented user needs and query patterns.

A third trend involves a "meta" step, namely benchmarking the benchmark-building itself for structured heritage KGs. When CQs are formalized (often with example queries), they effectively define what a KGQA system should be able to answer. Bench4KE (Lippolis et al., 2025) makes CQ generation itself comparable by providing curated CQ resources and evaluation metrics for automated CQ elicitation methods. AskCQ (Alharbi et al., 2025) similarly supports controlled comparisons by producing CQs from the same user-story requirements via manual, template-based, and LLM-based approaches, enabling analysis of how elicitation choices shape downstream evaluation. Crucially, evaluation is increasingly focusing on upstream steps—how CQs are elicited, normalized, disambiguated, and judged "good enough"—because these steps significantly determine the meaning of subsequent KGQA scores.

3.2 VISUAL DATA - VISUAL QUESTION ANSWERING

Early Visual Question Answering (VQA) benchmarks for art and cultural heritage (Antol et al., 2015) were developed prior to the advent of large language models. LLM based approaches, such as AQUA (Art Question Answering) (Garcia et al., 2020), utilize synthetically generated question–answer pairs evaluated with exact-match metrics with models fine-tuned on domain-specific training sets. Reported results typically range between 35% and 60% accuracy, reflecting the ability of models to exploit regularities within questions rather than robust multimodal semantic understanding. This design choice introduces pattern exploitation rather than genuine visual reasoning making accuracy scores poor proxies for interpretative competencies required in CH tasks. More recently, with the introduction of vision–language models the field marked a shift toward evaluation protocols that reduce sensitivity to linguistic biases and dataset-specific correlations. In this context, VQArt-Bench (Alfarano et al., 2025) represents a new generation of art-focused VQA benchmarks, explicitly targeting symbolic interpretation, narrative inference, and complex visual reasoning in zero-shot settings. Results showed a substantial performance gap with respect to general-domain VQA, with accuracy degrading systematically as reasoning complexity increases. This pattern suggests that even more recent models are anchored to surface-level feature recognition rather than deeper reasoning.

Beyond art-specific evaluations, other developments have introduced large-scale, cross-disciplinary benchmarks, such as MMMU (Yue et al., 2024) and HSSBench (Kang et al., 2025) with explicit coverage of cultural heritage domains. Their evaluation results indicate that, despite strong performance on general-domain tasks, LLMs often achieve lower accuracy, frequently around or below 60%, on humanities-focused questions. Unlike specifically-designed datasets these sets have broader topical scope and less domain specialization. This is a structural limitation often observed in broad benchmarks as they are optimized for wide coverage at the expense of cultural specificity. Therefore, domain-specific benchmarks remain essential for systematically contextual grounding and interpretive reasoning challenges in cultural heritage.

3.3 UNSTRUCTURED DOCUMENTS

In CH/DH, "unstructured documents" denotes materials whose informational content is not readily machine-actionable as clean fields or records: narrative catalogue descriptions, finding aids and archival inventories, scholarly prose, correspondence, card catalogues, and other primary sources—often available as digitizations where OCR and layout introduce additional noise. Benchmarks in this strand therefore

probe not only answer generation, but how systems cope with the conditions that make these sources hard to use computationally: historical spelling and orthographic variation, multilinguality, and the requirement that outputs remain traceable to specific documents (and, where applicable, pages or passages) rather than floating free of evidence.

One track evaluates question answering over large historical collections, where the task implicitly requires retrieval and grounding. *ArchivalQA* (Wang et al., 2021) operationalizes this by organizing questions along difficulty and time sensitivity (e.g., questions that hinge on dates or periods), making it explicit that what counts as a correct answer can depend on temporal context and shifting reference points. *ChroniclingAmericaQA* (Piryani et al., 2024) pushes the setting closer to institutional reality by evaluating across different representations of the same underlying sources (raw OCR, corrected OCR, and page images), exposing how performance depends on digitization quality and on where breakdowns occur across the retrieval–reading–answering chain.

A second, complementary track focuses upstream: extracting and normalizing structured representations from document-like sources, which then shapes what later retrieval and linking can even access. The *Humanities Data Benchmark* (Hindermann, 2025) exemplifies this approach by packaging heterogeneous humanities tasks with prompts, gold data, and evaluation scripts, treating "making text usable" as part of what should be benchmarked rather than assumed. Within it, *Zettelkatalog* (Humanities Data Benchmark sub-benchmark) evaluates field-level bibliographic extraction from index cards, illustrating a broader institutional pattern: before one can meaningfully run QA or synthesis, collections often require controllable, scoreable structuring steps that translate messy documentary traces into actionable records.

A third line connects these two by benchmarking intermediate, evidential claims that sit between extraction and downstream synthesis—especially relational statements that later support linking, knowledge graph construction, and provenance-aware querying. *HIPE-2026* (Opitz et al., 2025) moves in this direction by targeting relational, temporally qualified information extraction from historical texts ("who was where, when?"), aligning evaluation with the kinds of statements that subsequent systems must reconcile across documents and over time. Finally, DH-facing RAG evaluations increasingly combine automatic scoring with expert judgement because usefulness depends not only on factuality but also on how well systems manage ambiguity, consolidate dispersed evidence, and avoid overconfident synthesis when sources are thin or conflicting. *Topic-RAG* (Murugaraj et al., 2025) and *FolkRAG* (Kelly et al., 2025) reflect this shift, pointing to an open gap: benchmarks that test not just whether systems can "answer," but whether they can support exploratory inquiry by making their reasoning auditable, their failure cases legible, and their outputs stable under realistic variations in retrieval and context.

3.4 METHODOLOGIES TRENDS

Contamination and Memorization: A precondition for interpreting any benchmark score is confidence that the model has not already encountered the test data during training. Data contamination—the overlap between training corpora and test sets—is now recognized as a systemic validity threat. Empirical investigations have shown that some commercial models can reconstruct missing portions of well-known test sets in ways consistent with memorization rather than reasoning (Deng et al., 2024). Anti-leakage strategies emerging in general NLP—such as constructing reference documents that have never appeared on the web (Monteiro et al., 2024)—are directly relevant to CH, where many heritage sources are publicly indexed and therefore plausibly present in pre-training corpora.

Primary Source Noise: evaluation is increasingly treating digitization artefacts—OCR errors, layout effects, and uneven scan quality—as an explicit robustness axis, rather than assuming clean, already-normalized text. As discussed above with *ArchivalQA*, *ChroniclingAmericaQA*, and *MultiOCR-QA*, keeping the task fixed while varying representations or noise levels makes performance shifts more diagnostic at the system level and helps localize where breakdowns concentrate (Wang et al., 2021; Piryani et al., 2024; Piryani et al., 2025). Complementarily, Levchenko (2025) argues for OCR evaluation protocols that emphasize historically consequential error patterns and include leakage/instability safeguards, reflecting that degraded digitizations are routine in DH inputs.

Historical Interpretation: an emerging multimodal line evaluates models on collection-grounded questions tied to museum records, moving beyond generic VQA toward exhibit-centric competencies such as attribute identification and context-dependent reasoning. *Museum-65* (Balauca et al., 2024) exemplifies the scale-first approach—large, multilingual, and aligned with catalogue-style inquiry—making it valuable for stress-testing and training; at the same time, its format highlights a construct-

validity trade-off, since catalogue questions capture only part of the interpretive work that matters in heritage research.

Cultural Bias: A closely related trend is the transition from "culture as content" to culture as a measurable bias. Benchmarks are now being designed to expose geo-linguistic asymmetries and the "translation effect". *CulturalVQA* (Nayak et al., 2024) demonstrates performance disparities between regions using questions tied to specific cultural practices. *CVQA* (Romero et al., 2024) radicalizes this approach by including images and questions culturally situated across 31 languages and 30 countries, developed in collaboration with native speakers and experts.

Expert-Level Knowledge: In the realm of "humanities knowledge", we observe a shift from accuracy on factoid questions (questions focused on isolated, verifiable facts like dates or names) to benchmarks that test historical reasoning, reliability, and utility. *HiST-LLM* (Benam et al., 2024) converts structured global history data (Seshat) into multiple-choice questions with broad coverage, targeting "expert-level knowledge" while adopting a format that tends to minimize ambiguity. Conversely, *HistBench* (Qiu et al., 2025) attempts to include multimodal and interpretive questions on primary sources (manuscripts, inscriptions, objects), explicitly addressing the difficulty of evaluating open-ended responses and maintaining consistent scoring rubrics. These projects aim to measure the gap between apparent performance and usable performance in scholarly contexts.

Scalability and Synthetic Data Generation Finally, an important aspect is synthetic data generation of evaluation datasets. Reliance on manually curated benchmarking datasets from existing collections has become increasingly impractical at scale. In the years, data generation approaches have progressed from rule-based question generation and metadata-driven synthesis for GLAM-oriented benchmarks (Balauca et al., 2025; Becattini et al., 2023) to LLM-driven question generation, resulting in informative accuracy scores that better reveal models' limitations in symbolic and contextual reasoning. However, the use of synthetic data generation introduces challenges related to reliability and validity. For instance, reliance on LLM-generated questions risks confining evaluation within the epistemic boundaries of the generating models, potentially favoring architectures with similar training distributions and limiting the assessment of truly out-of-distribution reasoning (Liu et al., 2025). A bifurcation is consolidating that explains many current scalability tensions. On one side are massive benchmarks useful for training and stress-testing (e.g., *Museum-65*); on the other are smaller, high-validity benchmarks (e.g., *CulturalVQA*, *HistBench*) that prioritize tailored expert curation and interpretive tasks. This split indicates that the community is implicitly dividing between benchmarks that optimize for base capabilities and those that optimize for fidelity to humanities practice. The open challenge remains how to bridge these two levels through end-to-end pipelines, specialized rubrics, and anti-contamination measures.

4. CONCLUSION - GAPS AND POINTS OF INTERVENTIONS

What is at stake is not whether models are "good" in the abstract, but whether full workflows produce outputs that remain inspectable and defensible in humanities settings. Across the strands surveyed, a recurring gap emerges between isolated scores and whole performance under the messy conditions of heritage data: evaluations that vary input conditions (e.g., noisy OCR vs. cleaned text vs. page images) show that apparent accuracy can be fragile and pipeline-dependent, complicating leaderboard-style comparisons. For structured knowledge tasks, shifting from recall to verifiable, queryable representations is promising, but metrics should better reflect modelling assumptions and user-facing requirements, not only whether a generated query happens to execute. In multimodal and interpretive settings, newer culture and art-oriented VQA benchmarks highlight the need for reasoning-focused questions and scoring schemes that can accommodate warranted alternatives rather than forcing single-gold answers. Finally, keeping evaluation trustworthy and usable calls for stronger benchmark governance contamination-resistant test design and leakage diagnostics, alongside reporting that includes operational constraints (cost, latency, maintainability) as part of "what counts" in CH/DH practice. Without these shifts, benchmarks risk rewarding test-shaped behavior instead of systems that can be adopted responsibly in cultural institutions and scholarship.

ACKNOWLEDGEMENTS

During the preparation of this submission, we used generative AI tools for drafting support beyond basic spelling and grammar checks: summarization and translation in particular. All passages were reviewed

and edited by the authors, who take full responsibility for the final content, including the accuracy of claims, interpretations, and references.

REFERENCES

Alfarano, A., Venturoli, L., Castillo, D. N. del, & others. (2025). VQArt-Bench: A semantically rich VQA Benchmark for Art and Cultural Heritage. https://doi.org/10.48550/arXiv.2510.12750

Alharbi, R., Tamma, V., Payne, T. R., & de Berardinis, J. (2025). A Comparative Study of Competency Question Elicitation Methods from Ontology Requirements (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2507.02989

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2015). VQA: Visual Question Answering. Proceedings of the IEEE International Conference on Computer Vision (ICCV). https://doi.org/10.48550/arXiv.1505.00468

Balauca, Ada-Astrid, Garai, S., Balauca, S., Shetty, R. U., Agrawal, N., Shah, D. S., Fu, Y., Wang, X., Toutanova, K., Paudel, D. P., & Van Gool, L. (2024). Understanding Museum Exhibits using Vision-Language Reasoning (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2412.01370

Balauca, A.-A., Paudel, D. P., Toutanova, K., & Van Gool, L. (2024). Taming CLIP for Fine-Grained and Structured Visual Understanding of Museum Exhibits. In Computer Vision – ECCV 2024. https://doi.org/10.1007/978-3-031-73116-7_22

Becattini, F., Bongini, P., Bulla, L., Di Bimbo, A., Marinucci, L., Mongiovì, M., & Presutti, V. (2023). VISCOUNTH: A Large-scale Multilingual Visual Question Answering Dataset for Cultural Heritage. ACM Transactions on Multimedia Computing, Communications, and Applications. https://doi.org/10.1145/3590773

Benam, M., Bennett, J., Cioni, E., Del Rio-Chanona, R., Francois, P., Hauser, J., Hoyer, D., Kondor, D., Reddish, J., Turchin, P., & Villa, F. (2024). Large Language Models' Expert-level Global History Knowledge Benchmark (HiST-LLM). Advances in Neural Information Processing Systems 37, 32336–32369. https://doi.org/10.52202/079017-1016

Bowman, S. R., & Dahl, G. (2021). What Will it Take to Fix Benchmarking in Natural Language Understanding? In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 4843–4855). https://doi.org/10.18653/v1/2021.naacl-main.385

Carriero, V. A., Gangemi, A., Mancinelli, M. L., Marinucci, L., Nuzzolese, A. G., Presutti, V., & Veninata, C. (2019). ArCo: The Italian Cultural Heritage Knowledge Graph. In The Semantic Web – ISWC 2019 (Vol. 11779, pp. 36–52). Springer. https://doi.org/10.1007/978-3-030-30796-7_3

Daquino, M., Mambelli, F., Peroni, S., Tomasi, F., & Vitali, F. (2017). Enhancing Semantic Expressivity in the Cultural Heritage Domain: Exposing the Zeri Photo Archive as Linked Open Data. Journal on Computing and Cultural Heritage, 10(4), 1–21. https://doi.org/10.1145/3051487

Daquino, M., & Tomasi, F. (2015). Historical Context Ontology (HiCO): A Conceptual Model for Describing Context Information of Cultural Heritage Objects. In E. Garoufallou, R. J. Hartley, & P. Gaitanou (Eds), Metadata and Semantics Research (Vol. 544, pp. 424–436). Springer International Publishing. https://doi.org/10.1007/978-3-319-24129-6_37

De Berardinis, J., Carriero, V. A., Jain, N., Lazzari, N., Meroño-Peñuela, A., Poltronieri, A., & Presutti, V. (2023). The Polifonia Ontology Network: Building a Semantic Backbone for Musical Heritage. In T. R. Payne, V. Presutti, G. Qi, M. Poveda-Villalón, G. Stoilos, L. Hollink, Z. Kaoudi, G. Cheng, & J. Li (Eds), The Semantic Web – ISWC 2023 (Vol. 14266, pp. 302–322). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-47243-5_17

Deng, C., Zhao, Y., Tang, X., et al. (2024). Investigating Data Contamination in Modern Benchmarks for Large Language Models. NAACL 2024.

Garcia, N., Ye, C., Liu, Z., Hu, Q., Otani, M., Chu, C., Nakashima, Y., & Mitamura, T. (2020). A Dataset and Baselines for Visual Question Answering on Art. In Computer Vision – ECCV 2020 Workshops. https://doi.org/10.1007/978-3-030-66096-3_8

Gounakis, N., Mountantonakis, M., & Tzitzikas, Y. (2023). Evaluating a Radius-based Pipeline for Question Answering over Cultural (CIDOC-CRM based) Knowledge Graphs. Proceedings of the 34th ACM Conference on Hypertext and Social Media, 1–10. https://doi.org/10.1145/3603163.3609067

Grüninger, M., & Fox, M. S. (1995). The Role of Competency Questions in Enterprise Engineering. In A. Rolstadås (Ed.), Benchmarking—Theory and Practice (pp. 22–31). Springer US. https://doi.org/10.1007/978-0-387-34847-6_3

Hauser, J. E. (2025). HiST-LLM [Data set]. Zenodo. https://doi.org/10.5281/ZENODO.14671247

Hindermann, M. (2025). LLM-Benchmarking für die Digital Humanities: Ein praxisorientierter Ansatz aus der Forschungsberatung. https://doi.org/10.5281/ZENODO.17582144

HIPE-2026 Shared Task: Evaluating Accurate and Efficient Person–Place Relation Extraction. (2026). https://hipe-eval.github.io/HIPE-2026/about

Humanities Data Benchmark. (2025). https://rise-unibas.github.io/humanities_data_benchmark/

IDEA: A framework for LM-assisted ontology engineering (CQs extraction/analysis). (2023). https://github.com/polifonia-project/idea | https://doi.org/10.5281/zenodo.7660284

Kang, Z., Gong, J., Yan, J., Xia, W., Wang, Y., Wang, Z., Ding, H., Cheng, Z., Cao, W., Feng, Z., He, S., Yan, S., Chen, J., He, X., Jiang, C., Ye, W., Yu, K., & Li, X. (2025). HSSBench: Benchmarking Humanities and Social Sciences Ability for Multimodal Large Language Models (arXiv:2506.03922). arXiv. https://doi.org/10.48550/arXiv.2506.03922

Kelly, P., Schild, J., & Jafari, A. (2025). FolkRAG: A retrieval-augmented generation system for cultural heritage materials. Neural Computing and Applications, 37(24), 20281–20297. https://doi.org/10.1007/s00521-025-11455-4

Khadangi, A., Sartipi, A., Tchappi, I., & Fridgen, G. (2025). CognArtive: Large Language Models for Automating Art Analysis and Decoding Aesthetic Elements. https://doi.org/10.48550/arXiv.2502.04353

Kiela, D., Bartolo, M., Nie, Y., et al. (2021). Dynabench: Rethinking Benchmarking in NLP. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 4110–4124). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.324

Levchenko, M. (2025). Evaluating LLMs for Historical Document OCR: A Methodological Framework for Digital Humanities (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2510.06743

Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., Chen, K., & Lin, D. (2025). MMBench: Is Your Multi-modal Model an All-Around Player? In A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, & G. Varol (Eds.), Computer Vision – ECCV 2024 (Vol. 15064, pp. 216–233). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-72658-3_13

Lippolis, A. S., Ragagni, M. D., Ciancarini, P., Nuzzolese, A. G., & Presutti, V. (2025). Bench4KE: Benchmarking Automated Competency Question Generation (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2505.24554

Monteiro, J., Noel, P.-A., Marcotte, E., et al. (2024). RepLiQA: A Question-Answering Dataset for Benchmarking LLMs on Unseen Reference Content. arXiv:2406.11811.

Mountantonakis, M., & Tzitzikas, Y. (2025). Generating SPARQL Queries over CIDOC-CRM Using a Two-Stage Ontology Path Patterns Method in LLM Prompts. Journal on Computing and Cultural Heritage, 18(1), 1–20. https://doi.org/10.1145/3708326

Murugaraj, K., Lamsiyah, S., During, M., & Theobald, M. (2025). Topic-RAG for historical newspapers: Enhancing information retrieval in humanities research through topic-based retrieval-augmented generation. Computational Humanities Research, 1, e15. https://doi.org/10.1017/chr.2025.10018

Nayak, S., Jain, K., Awal, R., Reddy, S., van Steenkiste, S., Hendricks, L. A., Stanczak, K., & Agrawal, A. (2024). Benchmarking Vision Language Models for Cultural Understanding. https://mair-lab.github.io/papers/emnlp24_culturalvqa.pdf

Opitz, J., Ehrmann, M., Clematide, S., Corina, R., Boros, E., Michail, A., & Romanello, M. (2025). CLEF HIPE-2026—Shared Task Participation Guidelines. https://doi.org/10.5281/ZENODO.17800136

Phan, L. & others. (2025). Humanity's Last Exam. https://arxiv.org/abs/2501.14249

Piryani, B., Mozafari, J., Abdallah, A., Doucet, A., & Jatowt, A. (2025). Evaluating Robustness of LLMs in Question Answering on Multilingual Noisy OCR Data (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2502.16781

Piryani, B., Mozafari, J., & Jatowt, A. (2024). ChroniclingAmericaQA: A Large-scale Question Answering Dataset based on Historical American Newspaper Pages. Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2038–2048. https://doi.org/10.1145/3626772.3657891

PolifoniaCQ dataset. (2023). https://github.com/polifonia-project/polifoniacq-dataset | https://doi.org/10.5281/zenodo.7924354

Qiu, J., Xiao, F., Wang, Yimin, Mao, Y., Chen, Yijia, Juan, X., Zhang, S., Wang, S., Qi, X., Zhang, T., Yao, Z., Guo, J., Lu, Y., Argon, C., Cui, J., Chen, D., Zhou, J., Zhou, S., Zhou, Z., … Wang, M. (2025). On Path to Multimodal Historical Reasoning: HistBench and HistAgent (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2505.20246

Romero, D., Lyu, C., Wibowo, H. A., & others. (2024). CVQA: Culturally-diverse Multilingual Visual Question Answering Benchmark. https://arxiv.org/abs/2406.05967

Tomasi, F. (2022). Organizzare la conoscenza: Digital Humanities e Web semantico. Editrice Bibliografica. https://doi.org/10.53134/9788893573573

Wang, J., Jatowt, A., & Yoshikawa, M. (2021). ArchivalQA: A Large-scale Benchmark Dataset for Open Domain Question Answering over Historical News Collections. https://arxiv.org/abs/2109.03438

Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., Wei, C., Yu, B., Yuan, R., Sun, R., Yin, M., Zheng, B., Yang, Z., Liu, Y., Huang, W., … Chen, W. (2024). MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI (arXiv:2311.16502). arXiv. https://doi.org/10.48550/arXiv.2311.16502

Zettelkatalog (Humanities Data Benchmark sub-benchmark). (2025). https://rise-unibas.github.io/humanities_data_benchmark/benchmarks/zettelkatalog/ https://doi.org/10.5281/zenodo.18293269

L’intelligenza artificiale per il potenziamento dell’information retrieval nei discovery tool delle biblioteche

ILARIA BELVEDERE[1], SIMONA TURBANTI[2]

[1]SCUOLA IMT ALTI STUDI LUCCA, ILARIA.BELVEDERE@IMTLUCCA.IT
[2]UNIVERSITÀ DEGLI STUDI DI MILANO, DIPARTIMENTO DI STUDI STORICI F. CHABOD, SIMONA.TURBANTI@UNIMI.IT*

ABSTRACT (ITALIANO)

I discovery tool si sono affermati come strumenti di accesso unificato alle risorse analogiche e digitali delle biblioteche, sostituendo o inglobando progressivamente gli OPAC tradizionali. Studi recenti ne hanno tuttavia messo in evidenza alcune criticità, in particolare in relazione alla quantità dei risultati restituiti, ai criteri di ordinamento e alla difficoltà di supportare processi di ricerca esplorativa. Il contributo analizza tali limiti alla luce dei bias strutturali, tecnologici e culturali dei sistemi di discovery. Vengono, quindi, esaminate le potenzialità dell’intelligenza artificiale per il miglioramento dell’information retrieval, in particolare dei Large Language Models e dell’approccio Retrieval Augmented Generation, presentando alcuni casi di applicazione, tra i quali Alphy, il chatbot del portale Alphabetica dell’Istituto centrale per il catalogo unico delle biblioteche italiane. Si ricordano, infine, i limiti dei modelli di IA riguardanti l’attendibilità delle fonti, la tutela della privacy dei fruitori, la presenza di bias algoritmici e la scarsa trasparenza.
Parole chiave: information retrieval; discovery tool; cataloghi; intelligenza artificiale; retrieval augmented generation

ABSTRACT (ENGLISH)

Discovery tools have become key systems for unified access to libraries’ analogue and digital resources, progressively replacing or integrating traditional OPACs. Recent studies, however, have identified several critical issues, particularly regarding result overload, ranking mechanisms, and limited support for exploratory search. This paper examines these challenges in relation to the structural, technological, and cultural biases embedded in discovery systems. It then explores the potential of artificial intelligence to enhance information retrieval, focusing on Large Language Models and the Retrieval-Augmented Generation approach. Selected applications are discussed, including Alphy, the chatbot developed for the Alphabetica portal of the Central Institute for the Union Catalogue of Italian Libraries. The paper concludes by outlining key limitations of current AI models, notably concerning source reliability, user privacy, algorithmic bias, and transparency.
Keywords: information retrieval; discovery tool; catalogs; artificial intelligence; retrieval augmented generation

1. I DISCOVERY TOOL OVVERO UNA RICERCA IN APPARENZA EFFICACE E NEUTRALE

Da tempo le istituzioni della memoria mettono in contatto chiunque voglia con il proprio materiale attraverso i cataloghi, prima cartacei e poi online; in particolare, le biblioteche sono state i primi enti a dotarsi di OPAC. Negli ultimi quindici/venti anni gli OPAC, soprattutto quelli a disposizione delle università, sono stati “inglobati” all’interno di contenitori che consentono di accedere a tutte le risorse bibliografiche di un ateneo a partire da un’unica interfaccia di ricerca, i cosiddetti *discovery tool* (da qui in avanti discovery); la coesistenza, infatti, tra materiale analogico e in formato digitale nelle collezioni delle biblioteche comporta l’inevitabile mescolamento tra descrizioni di oggetti cui l’utente può accedere recandosi in un luogo fisico in cui ne è conservata una o più copie e descrizioni di oggetti esistenti nel web, di accesso immediato direttamente dalla registrazione catalografica. Ormai quasi tutti i sistemi bibliotecari di atenei e centri di ricerca in tutto il mondo e le grandi istituzioni bibliotecarie presentano agli utenti maschere di ricerca semplificate basate su questi strumenti di ricerca unificata i cui vantaggi sono

***** Il contributo è stato impostato e scritto in piena collaborazione tra le due autrici; tuttavia, le responsabilità sono così ripartite: Ilaria Belvedere ha steso il paragrafo 2 e parte del 3, Simona Turbanti i paragrafi 1, parte del 3 e 4.

facilmente intuibili in termini di rapidità e (falsa) facilità di uso, ma i cui limiti e rischi non sono stati sottolineati sufficientemente. Confrontando vari contributi sull'usabilità dei discovery sia in ambito italiano (Bardelli & Verga, 2019; Francese, 2013; Machetti, 2016; Quaquarelli, 2022; Raieli, 2020) sia internazionale (Chickering & Yang, 2014; Fagan et al., 2012; Hanrath & Kottman, 2015; Spencer & Millson-Martula, 2014), emerge la sensazione che non sia affatto chiaro se e in quale misura i cataloghi attuali stiano rispondendo alle funzioni che sono chiamati a svolgere - ossia trovare, identificare, selezionare, ottenere - salvando il tempo del lettore, come definito efficacemente già alla fine del XIX secolo (Cutter, 1876: 10). Le criticità dei discovery messe in luce dalle analisi condotte si concentrano sul rumore prodotto dai numerosi risultati ottenuti e sulla loro presentazione in un certo ordine (ranking), al pari di quanto avviene in un motore di ricerca, ma forse ci si può spingere oltre chiedendosi se questi ambienti *Google-like* non stiano rendendo più tortuoso il complesso processo cognitivo sotteso alla ricerca bibliografica. Secondo l'interpretazione di Kuhlthau (1991), infatti, l'Information Search Process (ISP) non è semplicemente un'attività tecnica di recupero dell'informazione, ma un cammino articolato in varie fasi durante il quale un utente non si limita a ottenere dati, ma realizza gradualmente una comprensione personale di un argomento; in altre parole, l'oggetto della ricerca non è sempre chiaro in partenza e molto spesso si definisce pian piano nella mente di chi consulta uno strumento grazie all'integrazione di nuove informazioni con conoscenze pregresse, valutando e interpretando criticamente le fonti e, non da ultimo, sviluppando emozioni quali incertezza, ansia, fiducia. Non è opportuno soffermarci, in questa sede, in questioni legate alla neutralità e oggettività, ma è importante ricordare come secondo Reidsma (2019) e Gualtieri et al. (2023) i sistemi di discovery bibliotecari presentino bias di tipo strutturale, tecnologico e culturale che la fiducia riposta generalmente nelle biblioteche tende a "mascherare". Dall'analisi, basata sui discovery tool Summon e Primo di Ex libris Clarivate, Ebsco discovery service (EDS) e OCLC's WorldCat Discovery, emerge come questi pregiudizi non derivino solo dagli algoritmi, ma anche dalla qualità e incoerenza dei metadati, dalle scelte dei fornitori commerciali e dalle logiche di indicizzazione su larga scala.

> The problems with algorithms are not only in the code or the steps of each algorithm; they are not in engineering teams, or the engineering education that leaves out ethical reasoning; they are not in the problematic metadata or the content; they are not even in the business models of the companies that make the algorithms. Rather, the problems can be found only by examining the embedded, historical context in which business practices, engineering education and the lack of diversity, systemic racism and sexism embedded in information classification systems, designs based on prejudiced stereotypes, and a culture that glorifies efficiency above all else (Reidsma, 2019: 147-148).

Oltre alla necessità di un approccio etico e critico ai sistemi di discovery in cui i bibliotecari si facciano carico della mediazione algoritmica dell'informazione, in questi strumenti di ricerca che attingono da numerose fonti bibliografiche eterogenee esistono altri fattori di cui tenere conto. Come evidenziato in Turbanti (2020: 398), nei cataloghi attuali sono sacrificate due componenti essenziali che dovrebbero contraddistinguerli, vale a dire la funzione esplorativa e la presentazione ordinata e logica dei risultati di una ricerca. Abbiamo già avuto modo di accennare alla complessa azione cognitiva della ricerca che non viene soddisfatta semplicemente con il reperimento di uno o più risultati; la serendipità che può rappresentare una "scintilla" importante nella ricerca degli studiosi (Foster & Ford, 2003; Watson, 2008) rischia di venir meno in ambienti troppo vasti e disomogenei. A questo si aggiunge la presentazione delle entità bibliografiche nei cataloghi che obbliga l'utente a faticose analisi dei risultati delle ricerche, in quanto non basata sui livelli stabiliti dai modelli concettuali come FRBR (IFLA, 1998) e IFLA LRM (Riva, P., Le Boeuf, P., & Žumer, M., 2017).
Da rilevazioni condotte sull'uso dei discovery tool in alcuni atenei italiani emerge che nella quasi totalità delle ricerche effettuate vengono usate parole chiave all'interno di un solo campo, al pari di quanto accade nei motori di ricerca, e che si tende a cercare oggetti bibliografici della cui esistenza si è già informati (Quaquarelli, 2022: 353-354; Gruppo di lavoro test di usabilità Minerva (2025). Nonostante studi condotti in realtà bibliotecarie estere, soprattutto nordamericane, che adottano discovery abbiano consentito di analizzare le abitudini di ricerca, le aspettative e le difficoltà degli utilizzatori nell'uso di questi strumenti permettendo la personalizzazione dell'interfaccia ai fini di una risposta più centrata (Nichols et al., 2017), gli attuali sistemi di information retrieval non sembrano, dunque, rispondere in modo efficace alle funzioni di identificazione, localizzazione e raggruppamento per le quali il catalogo è nato. Cerchiamo, quindi, di capire se e come le tecnologie di intelligenza artificiale in rapida evoluzione

aprano realmente "new possibilities for libraries to build conversational interfaces and smarter search engines that understand context and intent beyond simple keywords" (Szabó, 2025: 5).

2. L'AI COME STRUMENTO PER MIGLIORARE L'INFORMATION RETRIEVAL NEI DISCOVERY

Negli ultimi anni al centro del dibattito in svariati campi del sapere, gli strumenti di intelligenza artificiale sono in corso di applicazione anche nelle discipline del libro e del documento: per quanto riguarda le biblioteche, l'intelligenza artificiale può essere impiegata in contesti differenti, dalla catalogazione al reference e al supporto alla ricerca bibliografica (Georgieva, Georgiev, 2025: 705-706; La Gorga et al., 2025: 64-66). L'intelligenza artificiale può essere utilizzata dalle biblioteche in modo da migliorare l'efficienza dei servizi e il coinvolgimento dell'utenza (Georgieva, Georgiev, 2025: 703), in particolare per l'information retrieval (Szabó, 2025: 7-9). Proprio nel contesto della ricerca bibliografica e del recupero dell'informazione, il discovery tool in quanto "ponte" tra la biblioteca e l'utenza, può ampliare le sue potenzialità attraverso l'adozione dell'IA.
Si parla di *Intelligent Library Management Systems* (ILMS) quando i *Library Management Systems*, ovvero i software usati per gestire i servizi di una biblioteca, utilizzano soluzioni innovative, tra cui l'IA (Georgieva, Georgiev, 2025: 706). Per quel che concerne l'implementazione, l'IA può essere direttamente integrata al software o può essere introdotta dall'esterno attraverso l'uso di API o *plug in*, può apparire come un'interfaccia conversazionale (*chatbot*) o come un assistente di ricerca (Georgieva, Georgiev, 2025: 707-709). In ogni caso, vengono utilizzate tecniche di elaborazione del linguaggio naturale (NLP) per migliorare l'esperienza d'uso della ricerca, che passa dalla staticità delle parole chiave e degli operatori booleani, a una "discovery" basata sulla cognizione della semantica e del contesto nel quale viene svolta la *query* (Szabó, 2025: 5). L'IA può, dunque, servire al discovery tool per offrire un tipo di ricerca diversa, esplorativa, andando a recuperare una delle funzioni fondamentali del catalogo.
Come ricordato da Gino Roncaglia nel suo più recente contributo relativo a catalogazione e IA (Roncaglia, 2026: 112-113), nel quale riprende la sua metafora dell'architetto e dell'oracolo, fino agli '80 e '90 nel campo dell'IA era presente un tipo di visione architetturale in cui si cercava di sviluppare le tecnologie artificiali attraverso regole deterministiche; successivamente, si è passati alle rete neurali e a un'intelligenza artificiale basata sulla probabilità e la statistica, "oracolare" appunto. I nuovi sistemi si basano su Large Language Models (LLM), modelli di IA pre allenati su grandi moli di dati, e possono, oltre a comprendere il testo a partire da una *query*, rispondere, interrogare database e generare contenuto (Di Marcantonio, 2024: 156): si parla, infatti, di IA generativa. Nonostante l'alto livello raggiunto da queste tecnologie, esse portano con sé anche allucinazioni e bias, oltre a considerazioni di carattere etico e relative alla sicurezza dei dati e alla privacy (Szabó, 2025: 9-12). I LLM sono addestrati su una mole di dati che è in gran parte non strutturata (Roncaglia, 2026: 113) e questo apporta come conseguenza che non vi è certezza dell'affidabilità dei dati stessi e delle fonti da cui essi derivano. Per limitare i fenomeni di allucinazioni, uno degli approcci che sta prendendo piede è la tecnica RAG (Retrieval Augmented Generation). Ritornando alla metafora precedente, la RAG rappresenta un modalità di collaborazione tra le due visioni contrapposte - l'una deterministica, l'altra oracolare -, servendosi da una parte di strumenti strutturati di recupero dell'informazione, dall'altra degli LLM (Roncaglia, 2026: 118). In particolare, la RAG si basa su una fase di ricerca ("retrieval") e una di generazione di contenuto ("generation"):

> Nella fase di retrieval, il sistema utilizza algoritmi di ricerca per identificare i documenti o i contenuti più pertinenti alla query dell'utente. Questi algoritmi possono basarsi su tecniche di embedding semantico, che permettono di catturare il significato delle parole e delle frasi, andando oltre la semplice corrispondenza lessicale. La fase di generation, invece, sfrutta la capacità del LLM di sintetizzare informazioni e generare risposte coerenti, utilizzando come input sia la query originale che le informazioni recuperate (Izacard & Grave, 2021). Un'ulteriore caratteristica della RAG è la sua possibilità di incorporare diverse fonti di conoscenza. Queste possono includere database strutturati, archivi di documenti, pagine web, o persino flussi di dati in tempo reale. (La Gorga, 2025: 61).

Il reperimento dell'informazione e la successiva risposta non si basano unicamente sulla conoscenza pregressa del modello; nella RAG si fornisce all'IA un insieme di fonti definito, esterno e verificabile, dal quale l'IA va ad attingere al momento in cui deve fornire una risposta a una *query*, integrandolo con quanto già "conosce". I sistemi basati su un'architettura RAG permettono quindi:

- maggiore "pertinenza e accuratezza delle risposte generate" (Di Marcantonio, 2024: 159);

- maggiore trasparenza rispetto all'informazione fornita, in quanto sono indicate le fonti stesse della risposta generata dall'IA - e quindi, maggiore fiducia da parte degli utenti (La Gorga et al., 2025: 62);
- un tipo di risposta facilmente adattabile al contesto comunicativo dell'utente grazie all'uso degli LLM nella parte di *generation* (Roncaglia, 2026: 118).

La RAG può essere utilizzata per migliorare e rivoluzionari i servizi bibliotecari, compreso l'information retrieval (Mukhopadhyay, 2025: 109-110). Tuttavia, è fondamentale interrogarsi sulle sue criticità, per esempio il bilanciamento tra la parte di *retrieval* e di *generation* e l'attinenza tra le risposte fornite dall'IA e la richiesta dell'utente (La Gorga et al., 2025: 62-63). Rimane inoltre aperta l'inevitabile questione del controllo sui dati forniti al sistema AI come fonte di conoscenza; infine, per quanto riguarda la parte di AI generativa, bisogna considerare come spesso non sia garantita la trasparenza necessaria in riferimento ai modelli effettivamente impiegati. Nel prossimo paragrafo analizzeremo alcuni esempi pratici di istituzioni che hanno scelto di adottare strumenti IA nell'ambito della ricerca bibliografica, in modo da comprendere più da vicino le loro funzionalità.

3. ALPHABETICA E ALPHY DELL'ISTITUTO CENTRALE PER IL CATALOGO UNICO DELLE BIBLIOTECHE ITALIANE E PER LE INFORMAZIONI BIBLIOGRAFICHE (ICCU) E I RESEARCH ASSISTANT NEI DISCOVERY

A dicembre 2021 è stato presentato al pubblico Alphabetica,[1] il portale di ricerca integrata dell'Istituto centrale per il catalogo unico delle biblioteche italiane e per le informazioni bibliografiche (ICCU) che consente di ricercare nelle banche dati e risorse gestite dall'ICCU a partire da un unico punto di accesso (Buttò, 2020; Battaggia, 2022; Castro, 2022). Tramite la piattaforma è possibile interrogare l'OPAC SBN, Manus Online, EDIT16, SBN Teca, Cataloghi Storici Digitalizzati e banche dati esterne, ad es. il Catalogo collettivo digitale IMAGO. A differenza di quanto accade con il catalogo tradizionale SBN, i risultati ottenuti sono disposti in "contenitori" (box) tipologici o tematici - Libri, Grafica, Audio e Video, Musica, Cartografia, Manoscritti, Periodici, Biblioteche, Protagonisti - offrendo quindi l'opportunità di scegliere ai visitatori di costruire un percorso *ad hoc*; da ciascun box, espandibile, inizia la navigazione all'interno del portale.
Un'ulteriore funzione offerta è la mappa concettuale con la quale estendere l'ambito della ricerca a termini/concetti correlati indicizzati nel thesaurus del Nuovo soggettario di Firenze. Alphabetica si pone, quindi, come un "servizio self-service" che agisce in base a principi diversi da quelli di tipo commerciale su cui sono costruiti i motori di ricerca, amplia i confini tradizionali dei cataloghi bibliografici e offre un sostegno ai docenti nei percorsi scolastici (Castro 2022a: 42; 45-47; Castro, 2022b: 563-564; 568).
Per le finalità di questo contributo, l'aspetto più interessante è costituito dal progetto Alphy, un servizio sperimentale di chatbot basato sull'intelligenza artificiale generativa, progettato tra la fine del 2023 e gli inizi del 2024 per assistere gli utenti nella navigazione e nell'accesso alle informazioni bibliografiche di Alphabetica; si basa sull'infrastruttura cloud I.PaC per la gestione, l'arricchimento semantico e l'interoperabilità delle risorse culturali digitali all'interno dell'ecosistema digitale per la cultura Ecomic, un ambiente collaborativo distribuito per sostenere la trasformazione digitale del settore culturale, promosso da Digital Library per la Direzione generale per la Digitalizzazione e la comunicazione del Ministero della cultura.[2] Nella versione 1.9.0, Alphy è in grado di rispondere alle domande poste in linguaggio naturale e, grazie all'approccio RAG, effettuare ricerche più accurate nei testi digitalizzati; rispetto ai modelli precedenti il chatbot restituisce, inoltre, le citazioni estratte dal materiale digitalizzato in un formato più chiaro, presenta dati maggiormente accurati sulle localizzazioni delle biblioteche che possiedono una copia del documento cercato e consente di geolocalizzare le biblioteche vicine alla posizione dell'utente.
Prove effettuate con ricerche di vario tipo hanno mostrato un livello piuttosto accurato nelle risposte; di seguito un esempio.
Prompt: Concetto di serendipità in biblioteca: quali opere ne parlano?

[1] https://alphabetica.it/web/alphabetica (cons. 31/01/2026).
[2] https://ecomic.cultura.gov.it/ipac/ecomic/ (cons. 31/01/2026).

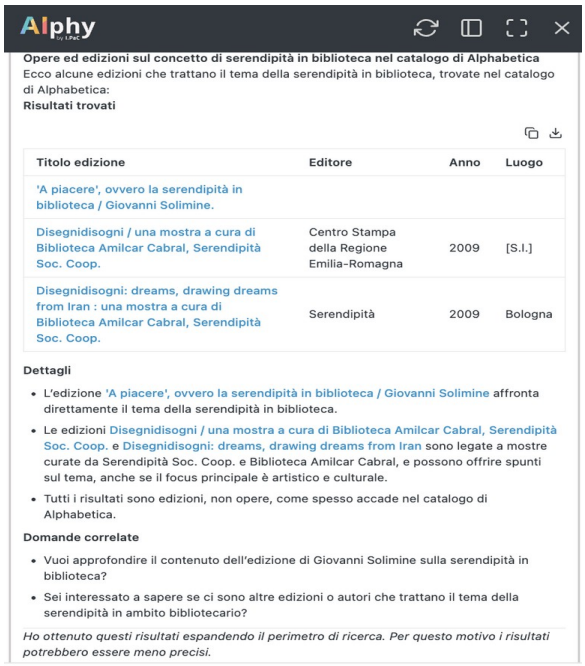


Figura 1. Risultati forniti da Alphy

Un altro esempio di chatbot basato sull'intelligenza artificiale generativa è Cat-IA,[3] sviluppato dall'Istituto centrale per il catalogo e la documentazione - ICCD insieme all'Istituto centrale per la digitalizzazione del patrimonio culturale - Digital Library per migliorare la ricerca all'interno del portale del Catalogo generale dei beni culturali. Al pari di Alphy, Cat-IA si inserisce nell'ecosistema Ecomic e intende facilitare la consultazione delle oltre 3 milioni di schede di beni culturali eterogenei (artistici, archeologici, fotografici, architettonici, naturalistici, numismatici, ecc.) presenti nel Catalogo generale dei beni culturali, nato nel 2021 e basato su dati aperti e strutturati correlati tra loro semanticamente (linked open data, LOD). Le risposte fornite dal chatbot ai quesiti posti, diversi come ambito e grado di precisione, si sono rivelate corrette e centrate, in alcuni casi dopo aver affinato meglio il prompt. Gli esempi esistenti di chatbot in biblioteche estere si dividono essenzialmente in strumenti progettati per soddisfare richieste di base che riguardano i servizi, le collezioni e gli spazi della biblioteca, ma non a ricerche bibliografiche vere e proprie per le quali rimandano ai cataloghi e discovery - ad esempio, T-Rex dell'University of Calgary in Canada e KingbotGPT della University Library's Artificial Intelligence Team della San Jose State University - e bot ideati più simili ad Alphy e Cat-IA, come AISHA, sviluppato dalla Zayed University Library.[4]
Per quanto riguarda, invece, lo sviluppo di assistenti di ricerca IA nei discovery, l'esempio più diffuso è quello di Primo Research Assistant, reso disponibile da fine 2024 su Primo VE (Georgieva, Georgiev, 2025: 707). Primo Research Assistant[5], accessibile tramite autenticazione, utilizza la RAG avvalendosi, per quanto riguarda il modello di AI generativa alla base, del mini GPT-4.1 di Open AI; inoltre, basa la sua ricerca sul *Central Discovery Index* (CDI) di Primo, con l'eccezione di alcuni contenuti. In seguito alla richiesta dell'utente, posta in linguaggio naturale, l'assistente genera come *output* i primi cinque risultati più rilevanti basati sugli abstract o le descrizioni, una loro *overview* comprendente citazioni dagli stessi, e una sezione "Related research questions". Le risposte generate sono direttamente collegate alle fonti e l'utente può porre domande in un linguaggio quotidiano, anche in una lingua che non sia l'inglese[6]. Primo Research Assistant è già stato adottato o testato in varie biblioteche a livello internazionale, per esempio, presso la biblioteca del KTH Royal Institute of Technology di Stoccolma e la biblioteca dell'Università di Westminster[7]. In Italia, l'assistente è stato incorporato nel discovery delle biblioteche dell'Università di

[3] https://catalogo.cultura.gov.it/ (cons. 31/01/2026).
[4] https://library.ucalgary.ca/contact, https://library.sjsu.edu/kingbot, https://aisha.zu.ac.ae/ (cons. 31/01/2026).
[5] https://knowledge.exlibrisgroup.com/Primo/Product_Documentation/020Primo_VE/Primo_VE_(English)/015_Getting_Started_with_Primo_Research_Assistant (cons. 31/01/2026).
[6] https://clarivate.com/news/clarivate-launches-generative-ai-powered-primo-research-assistant/ (cons. 31/01/2026).

Padova, dell'Università Roma Tre, IULM e della Libera Università di Bolzano[8]. Inoltre, la Direzione SBA dell'Università di Milano, in fase di *testing* dello strumento, ha svolto una prima analisi della sua efficacia raccogliendo i feedback di parte del personale bibliotecario e degli utenti in relazione a pertinenza ed esaustività del *tool* e presenza di allucinazioni, rilevando che l'assistente sembra uno strumento adeguato per orientarsi nella ricerca, ma riscontrando anche criticità in relazione alla esaustività e alle fonti utilizzate, talvolta datate e sempre incentrate sulla lingua inglese (Sozzi et al., 2026).

4. SCENARI FUTURI

Se è indubbio che l'intelligenza artificiale abbia ormai un ruolo importante nelle biblioteche e in tutte le istituzioni della memoria e offra opportunità significative per la personalizzazione dei servizi e la scalabilità dei processi di gestione dell'informazione i cui benefici siano già rilevabili in vari ambiti - non solo quello dell'information retrieval, ma anche il reference, la catalogazione e soprattutto indicizzazione -, è necessario tenere conto delle criticità non trascurabili legate alla tutela della privacy degli utenti, alla presenza di bias algoritmici e alla limitata trasparenza dei modelli di IA.
In riferimento a quest'ultimo punto, si è visto come, anche negli esempi di tecnologie RAG analizzate, solo in un caso, ovvero Primo Research Assistant, si menzioni il modello di AI generativa alla base, senza tuttavia entrare nei dettagli della sua applicazione tecnica. Tale "opacità" costituisce un punto problematico per le istituzioni pubbliche e i dati dei relativi utenti che finiscono in mano di terzi senza alcuna garanzia sui termini di uso (non riutilizzo e non conservazione); da quando la rete ha iniziato ad essere utilizzata in larga parte per scopi commerciali e grandi piattaforme economiche sono entrate nel mercato digitale si è affacciata, infatti, la questione della sovranità digitale, ossia del controllo indipendente - da parte di uno Stato, di un ente o di una persona - di risorse, dati e tecnologie. Esercitare il controllo su dati, infrastrutture e modelli di addestramento dei sistemi di intelligenza artificiale rappresenta una condizione essenziale per assicurare la sicurezza nazionale, soprattutto in momenti di incertezza geopolitica. Sempre in riferimento alla mancanza di trasparenza, nella sfera algoritmica, si parla di "scatola nera", vale a dire un modello in cui neanche gli sviluppatori stessi riescono a comprendere come il sistema abbia elaborato i dati per fornire le informazioni richieste. Tale inesplicabilità, di cui si occupa la branca dell'eXplainable Artificial Intelligence (XAI), è connaturata alla struttura stessa dei sistemi, spesso estremamente complessi, in particolar modo le reti neurali composte da centinaia di strati nascosti che trasmettono milioni di segnali; peraltro, nello sviluppo di un modello di apprendimento automatico, considerare il fattore interpretabilità può permettere di individuare e correggere i bias presenti nel set di dati di addestramento rafforzando il funzionamento del sistema (Barredo Arrieta et al., 2020: 83). Un altro fronte aperto riguarda il rapporto tra l'uso di modelli di IA generativa e l'integrità della ricerca scientifica; la capacità di elaborare testi, immagini, audio, video e codici comporta, inevitabilmente, ricadute a livello di responsabilità autoriale e di plagio, i cui confini divengono incerti. Secondo molti studiosi, questa situazione non va affrontata intervenendo soltanto in ambito normativo, ma stimolando soprattutto la conoscenza degli strumenti, il loro utilizzo critico e i loro limiti, in equilibrio tra adesione alle innovazioni e rispetto delle responsabilità e dell'etica. Le biblioteche e i docenti di ogni grado sono i soggetti deputati a farsi carico di questo processo di *AI literacy*; in particolare, i bibliotecari, professionisti dell'informazione, dovrebbero verificare che i sistemi di IA adottati dalle loro istituzioni non siano solo convenienti, ma anche in linea con la tutela dei dati degli utenti, l'integrità accademica e i valori sociali. L'IA vista, quindi, non semplicemente come un nuovo strumento, "but a catalyst forcing a re-examination of professional identity and skill sets in academic libraries" (Sousa, 2026: 12).
Le biblioteche rappresentano non soltanto i luoghi fisici nei quali si conserva e rende disponibile il materiale bibliografico, e non di rado documentale, in forma analogica e digitale, ma anche i luoghi più idonei in cui accertare l'autorevolezza delle fonti per mezzo di strumenti di ricerca; nonostante le modifiche di questi ultimi - da cataloghi cartacei a OPAC, da OPAC a *next generation catalogue* e a discovery - le loro funzioni sono rimaste immutate e rispondono al desiderio umano, prima che

[7] https://kth-ch.primo.exlibrisgroup.com/discovery/search?vid=46KTH_INST:46KTH_VU1_L&lang=en, https://libguides.westminster.ac.uk/librarysearchguide/aiassistant (cons. 31/01/2026).
[8] Phttps://biblio.unipd.it/biblioteche/scienze-farmaco/notizie-eventi/lassistente-di-ricerca-di-galileodiscovery, https://discovery.sba.uniroma3.it/discovery/search?vid=39RM_CABTRE:39CAB_V1, https://digger.iulm.it/discovery/search?vid=39IUL_INST:39IULM_VIEW , https://unibz.primo.exlibrisgroup.com/discovery/search?vid=39UBZ_INST:UNIBZ (cons. 31/01/2026).

bibliotecario, di attuare l'Universal Bibliographic Control (UBC). La sfida attuale è quella di reinterpretare l'utopia del controllo bibliografico universale all'interno di un vasto contesto informativo che interagisce con sistemi "intelligenti".

RINGRAZIAMENTI

Si ringraziano la dott.ssa Tiziana Morocutti, Dirigente del Servizio bibliotecario d'ateneo dell'Università degli studi di Milano, e i componenti dei Gruppi di lavoro dei test di usabilità del discovery tool e del Research assistant per il confronto e la condivisione dei risultati delle analisi svolte e di quelle in corso; un ringraziamento, inoltre, a uno dei revisori per gli utili spunti forniti.

BIBLIOGRAFIA

Bardelli, A., & Verga, F. (2019). The Discovery Tool is a Growing Organism. JLIS.It, 10(3), 97–124. https://doi.org/10.4403/jlis.it-12575.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina. D., Benjamins, R., Chatila, R. & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and Challenges Toward Responsible AI. Information Fusion, 58, 82-115. https://doi.org/10.1016/j.inffus.2019.12.012.

Battaggia, M. (2022). Alphabetica: Aspettative, Opinioni e Ricadute attraverso una Ricerca Esplorativa. DigItalia, 17(1), 96-105. https://digitalia.cultura.gov.it/article/view/2961.

Buttò, S. (2020). Alphabetica, il nuovo Portale per la Ricerca Integrata: Un Salto di Qualità per le Biblioteche Italiane. DigItalia, 15(2), 9-15. https://digitalia.cultura.gov.it/article/view/2624.

Castro, E. (2022a). Alphabetica: uno strumento poliedrico di accesso ai servizi bibliografici nazionali. DigItalia, 17(1), 39-47. https://digitalia.cultura.gov.it/article/view/2955.

Castro, E. (2022b). Alphabetica e il reference digitale: suggerimenti e mappe concettuali per gli utenti del portale. AIB Studi, 62(3), 559–570. https://doi.org/10.2426/aibstudi-12055.

Chickering, F. W., & Yang, S. Q. (2014). Evaluation and Comparison of Discovery Tools: An Update. Information Technology and Libraries, 33(2), 5–30. https://doi.org/10.6017/ital.v33i2.3471.

Cutter, C. (1876). Rules for a Printed Dictionary Catalogue, vol 2. of Public Libraries in the United States of America: Their History, Condition and Management: Special Report. Washington: Government Printing Office.

Di Marcantonio, G. (2024). Intelligenza artificiale, Large Language Models (LLMs) e Retrieval-Augmented Generation (RAG). Nuovi strumenti per l'accesso alle risorse archivistiche e bibliografiche. Bibliothecae.it, 13(1), 146-173. https://dx.doi.org/10.6092/issn.2283-9364/19982.

Fagan, J. C., Mandernach, M. A., Nelson, C. S., Paulo, J. R., & Saunders, G. (2008). Usability Test Results for a Discovery Tool in an Academic Library. Information Technology and Libraries, 31(1), 83–112. https://doi.org/10.6017/ital.v31i1.1855.

Foster, A, Ford N. (2003). Serendipity and Information Seeking: An Empirical Study. Journal of Documentation, 59(3), 321–340, doi: https://doi.org/10.1108/00220410310472518.

Francese, E. (2013). Test di Usabilità sul Discovery Tool "Primo" all'Università di Torino. Biblioteche Oggi, 31(10), 10-17. https://doi.org/10.3302/0392-8586-201310-010-1.

Georgieva, E., & Georgiev, T. (2025). How Artificial Intelligence Can Transform Services in Research Libraries. In: Kahraman, C., et al. Intelligent and Fuzzy Systems. INFUS 2025. Lecture Notes in Networks and Systems, vol 1530. Springer, Cham. https://doi.org/10.1007/978-3-031-98565-2_76.

Gruppo di lavoro test di usabilità Minerva (2025). Minerva allo Studio: Test di usabilità, a cura di Laura Gioppo. Milano, Direzione SBA, Università degli studi di Milano.

Gualtieri, F., Maoret R., Viazzi, F., Molinari, S., Truccolo, I., Di Lauro, E., Sabbatini, V., Formigoni, C. & Rossi, W. (2023). Same Search, Different Results: Algorithm Bias in Various Discovery Tools in Library Search: GIDIF-RBM Monza (Italy) Working Group in collaboration with TDNet. Journal of EAHIL, 19(3), 21-23. https://ojs.eahil.eu/JEAHIL/article/view/578.

Hanrath, S., & Kottman, M. (2015). Use and Usability of a Discovery Tool in an Academic Library. Journal of Web Librarianship, 9(1), 1–21. https://doi.org/10.1080/19322909.2014.983259.

International Federation of Library Associations and Institutions, Study Group on the Functional Requirements for Bibliographic Records (1998). Functional Requirements for Bibliographic Records: Final Report, approved by the Standing Committee of the IFLA Section on Cataloguing. München, Saur.

Kuhlthau, C.C. (1991). Inside the search process: Information Seeking from the User's Perspective. Journal of the American Society for Information Science, 42(5), 361-371. https://doi.org/10.1002/(SICI)1097-4571(199106)42:5<361::AID-ASI6>3.0.CO;2-%23.

La Gorga, A., Testa, R., Verna L. (2025). LLM e Retrieval Augmented Generation (RAG) per le biblioteche. Sperimentazioni, prospettive e valorizzazione del patrimonio. @digitcult, 10(1), 59-73. https://dx.doi.org/10.36158/97912566920714.

Machetti, C. (2016). Biblioteche e Discovery tool: Il Caso OneSearch e l'Ateneo di Siena. AIB Studi, 56(3), 391-408. https://doi.org/10.2426/aibstudi-11501.

Mukhopadhyay, P. (2025). Designing Conversational Search for Libraries: Retrieval Augmented Generation through Open Source Large Language Models. DESIDOC Journal of Library & Information Technology, 45(2), 109–115. https://doi.org/10.14429/djlit.20206.

Nichols, A. F., Crist, E., Sherriff, G., & Allison, M. (2017). What Does it Take to Make Discovery a Success?: A Survey of Discovery Tool Adoption, Instruction, and Evaluation Among Academic Libraries. Journal of Web Librarianship, 11(2), 85–104. https://doi.org/10.1080/19322909.2017.1284632.

Quaquarelli, M. (2022). Search Behavior on University of Bologna Discovery Tool: A Quantitative and Qualitative Analysis. Bibliothecae.It, 11(1), 315–360. https://doi.org/10.6092/issn.2283-9364/15079.

Raieli, R. (2020). Web-scale Discovery Services: Principi, Applicazioni e Ipotesi di Sviluppo. Roma: Associazione Italiana Biblioteche. ISBN 978-88-7812-295-6.

Reidsma, M. (2019). Masked by Trust: Bias in Library Discovery. Library Juice Press. https://scholarworks.gvsu.edu/library_books/30.

Riva, P., Le Boeuf, P., & Žumer, M. (2017). IFLA Library Reference Model: A Conceptual Model for Bibliographic Information: Definition of a Conceptual Reference Model to Provide a Framework for the Analysis of Non-administrative Metadata Relating to Library Resources. August 2017 Revised after Worldwide Review, Endorsed by the IFLA Professional Committee, as Amended and Corrected through December 2017. https://www.ifla.org/publications/node/11412.

Roncaglia, G. (2026). Cataloguing, Metadata, and Generative AI. Early Experiences and Future Perspectives. JLIS.It, 17(1), 106–127. https://doi.org/10.36253/jlis.it-693.

Sousa, N. (2025). Ethical and Practical Implications of AI in Academic Library Research. IFLA Journal. https://doi.org/10.1177/03400352251391753.

Sozzi, I., Rampichini, F., & Turchetti, M. F. (Eds.). (2026). Research assistant [slide]. Direzione SBA, Università degli Studi di Milano.

Spencer, J. S., & Millson-Martula, C. (Eds.). (2014). Discovery Tools: The Next Generation of Library Research. London and New York: Routledge. ISBN 9781032923314.

Szabó, N. (2025). The Role of AI in Enhancing Library Information Retrieval: Opportunities and Challenges. Zenodo. https://doi.org/10.5281/zenodo.15581788.

Turbanti, S. (2020). Della "giungla del digitale", della Tendenza a Semplificare e della Incompiutezza, ovvero I Cataloghi Oggi. In Vittorio Ponzani (a c. di). L'orgoglio di Essere Bibliotecari. Saggi in Ricordo di Maria A. Abenante (393-402). Roma: Associazione italiana biblioteche.

Watson, E. (2008). Going Fishing: Serendipity in Library and Information Science. A Master's paper for the M.S. in L.S. degree. https://doi.org/10.17615/tsra-0e20.

Sinergie ludonarrative: l'evoluzione del racconto videoludico tra divulgazione, tecnica ed intelligenza artificiale

GIULIA ANGELINI[1], TOMMASO MAZZOLI[2]
[1]UNIVERSITÀ DEGLI STUDI DI TRIESTE - ANGELINIGIULIA08@GMAIL.COM
[2]UNIVERSITÀ DEGLI STUDI DI TRIESTE - MAZZOLI@UNITS.IT

ABSTRACT (ITALIANO)

Il presente elaborato si propone di analizzare il rapporto tra meccaniche di gioco (gameplay) e narrazione all'interno del medium videoludico, esplorando il concetto di "piano ludonarrativo". Attraverso l'esame di casi studio specifici (*God of War*, *Dark Souls*, *The Stanley Parable*, *The Quarry* e altri), viene indagato come il videogioco rielabori i canoni della letteratura e del cinema per offrire nuove forme di fruizione culturale e apprendimento (game-based learning), contribuendo così anche alla conservazione delle nozioni. Infine, si osservano le prospettive future legate all'intelligenza artificiale e all'impatto sociale del medium.
Parole chiave: Game Studies; Game Culture; game-based learning; ludonarratività; Game AI

ABSTRACT (ENGLISH)

Ludonarrative Synergies: The Evolution of Video Game Storytelling between Dissemination, Technique, and Artificial Intelligence.
This paper aims to analyze the relationship between gameplay mechanics and narrative within the video game medium, exploring the concept of the “ludonarrative plane.” Through the examination of specific case studies (God of War, Dark Souls, The Stanley Parable, The Quarry, and others), the study investigates how video games reinterpret the conventions of literature and cinema to offer new forms of cultural engagement and learning (game-based learning), contributing to the conservation of knowledge. Finally, the paper considers future perspectives related to artificial intelligence and the social impact of the medium.
Keywords: Game Studies; Game Culture; game-based learning; ludonarrativity; Game AI

1. INTRODUZIONE: IL PIANO LUDONARRATIVO

L'analisi del videogioco contemporaneo richiede un approccio che integri il linguaggio audiovisivo di matrice cinematografica con l'attività ludica interattiva. Al centro di questa sinergia si trova il concetto di "piano ludonarrativo", definito da Ivan Fulco come il punto di incontro tra la meccanica di gioco e il racconto, ovvero il luogo in cui si realizza l'evento-gioco nella sua totalità e complessità (Fulco, 2004: 65).
A differenza della narrazione letteraria, che l'utente possiede nella sua interezza fin dal primo istante e a cui ha accesso in ogni sua parte in ogni momento, la narrazione videoludica possiede una natura temporale unica: essa avviene e si scrive nell'esatto momento in cui il giocatore agisce. Sebbene il codice costituisca un'impalcatura predeterminata, l'aggiornamento della trama è subordinato all'agency dell'utente, configurando l'esperienza come una “scrittura procedurale in divenire”.
La struttura fondamentale del racconto videoludico non si discosta dai modelli classici della teoria della narrazione (Propp, 1928): uno stato di equilibrio iniziale viene interrotto da un conflitto, generando una tensione che spinge verso la risoluzione e un nuovo equilibrio. Tuttavia, nel videogioco, questa tensione è amplificata dall'azione diretta dell’utente in ogni passaggio della sua esperienza: ogni incertezza o sfida (combattimenti, enigmi) richiama le abilità del giocatore, creando un livello di coinvolgimento emotivo e sensoriale difficilmente immaginabile precedentemente (Fulco, 2004: 76). Si potrebbe definire come un livello della struttura interna di un videogioco che permette al pubblico di interagire attivamente, in diversi modi, con la narrazione.

2. ARCHITETTURE NARRATIVE E STRUMENTI DI DIVULGAZIONE

Esistono diverse strategie per veicolare la trama e i significati nel medium videoludico. Henry Jenkins parla di environmental storytelling (Jenkins, 2004: 118-130) per descrivere come le informazioni narrative siano distribuite nello spazio esplorabile del gioco, che diventa esso stesso parte della

narrazione. Parallelamente, Astrid Ensslin introduce il literary gaming (Ensslin, 2014), indicando quei titoli che fanno ampio affidamento su materiali testuali, come diari, didascalie o brani introduttivi, riducendo talvolta il coinvolgimento attivo a favore di una fruizione simile alla lettura. Si tratta quindi di meccanismi complementari e non escludenti, che danno spesso completezza ad un prodotto se opportunamente integrati. Sono paragonabili alle tecniche di "show and tell" introdotte nell'ambito del storytelling, dove l'ambiente mostra mentre i testi raccontano. Nonostante la possibile somiglianza con il contesto cinematografico, in questo caso un comparto eccessivamente testuale renderebbe l'esperienza del pubblico troppo passiva, compromettendone l'efficienza, mentre la natura ergodica del medium (Aarseth, 1997) mitiga il rischio di passività: l'agency richiesta al giocatore controbilancia la densità testuale, mantenendo attivo il coinvolgimento.
Un elemento chiave per la divulgazione e l'approfondimento narrativo è il "codex" (o codice), un manuale virtuale interno al gioco che raccoglie regole e descrizioni di oggetti e personaggi incontrati lungo il proprio percorso ludico. L'analisi comparata di due titoli, *God of War* (2018) e *Hades* (2020), evidenzia come il codex possa trasformare il gioco in un'esperienza paragonabile ad un "museo mitologico" digitale (Vandewalle, 2023: 90-112):

- ***God of War***: Il codex è diegetico, scritto dal punto di vista di Atreus, figlio del protagonista. Esso funge da diario di viaggio, integrandosi perfettamente nell'architettura narrativa senza interrompere la sospensione dell'incredulità. Le descrizioni e i disegni riflettono ciò che è l'esperienza di Atreus, impiegando un apparato estetico e di design elaborato. Il titolo è ambientato in un universo controllato dalle divinità della mitologia e della cultura norrena, permettendo così al giocatore di conoscerne i personaggi protagonisti come ne viene a conoscenza lo stesso Atreus. L'utente può muoversi ed interagire con un museo vivo.
- ***Hades***: Il codex è attribuito ad Achille, mentore del protagonista Zagreus. Qui la funzione è più meccanica: le informazioni si sbloccano progressivamente sconfiggendo i nemici, interrompendo talvolta il flusso narrativo. Zagreus (quindi il giocatore) non sta scrivendo ciò che vede e incontra, ma prova ad orientarsi grazie a questo manuale scritto da qualcun altro. Le informazioni sono già a sua disposizione, deve però sbloccarle giocando in maniera continuativa ed interagendo con gli altri personaggi, quasi la totalità di questi tratti dalla mitologia greca classica.

In entrambi i casi, questi strumenti supportano il game-based learning, permettendo al pubblico di avvicinarsi alla mitografia e alla storia attraverso l'interazione ludica, incoraggiando un apprendimento che possa concentrarsi maggiormente sui modi in cui avviene e non solo sul che cosa esso voglia trasmettere (Baldassarre et al., 2023). Un esempio emblematico di questa commistione tra cultura alta e pop è *Dante's Inferno* (2010), dove, pur discostandosi dalla trama originale, il personaggio di Virgilio recita i versi originali della *Divina Commedia*, onorando la fonte letteraria originale. Ciò non vuole assolutamente sostituire l'opera dantesca, ma qui si vuole porre l'attenzione sul fatto che, in questo modo, il giocatore non sta solo vivendo un'esperienza ludica, ma sta personalmente (anche se virtualmente) interagendo con i diretti personaggi "culturali". L'interattività si fa quindi capace di enfatizzare e rafforzare il concetto secondo cui l'azione e l'esperienza diretta sono strumenti estremamente efficaci per l'apprendimento (Baldassarre et al., 2023). Risulta chiaro che maggiore sarà l'investimento intellettuale da parte degli sviluppatori alle spalle di questi titoli, maggiore sarà la qualità con cui il pubblico verrà in contatto con queste nozioni. In ogni caso, viene incentivato l'avvicinamento a questi argomenti in maniera diversa rispetto ad una classica lezione frontale o ad un supporto unicamente testuale.

3. ANALISI COMPARATA: QUATTRO APPROCCI ALLA NARRAZIONE

Per comprendere come le teorie narrative si applichino concretamente, si prendono in esame quattro titoli paradigmatici che mostrano diverse sfumature del rapporto tra giocatore e storia.

- **Riscrivere la Letteratura: *Alice Madness Returns***

Il videogioco si inserisce in un continuum evolutivo della narrazione che comprende letteratura e cinema. *Alice: Madness Returns* (2011) opera una riscrittura dei classici di Lewis Carroll, mantenendo i personaggi originali ma inserendoli in un contesto horror e psicologico. Questa operazione dimostra come una storia amata possa essere manipolata per esplorare temi più crudi come la follia, offrendo un design distorto ma

riconoscibile, come un romanzo riscritto da un altro autore. La narrazione è sia ambientale che testuale, anche se spesso vengono preferiti i dialoghi doppiati tra i personaggi iconici.

- **Il cinema interattivo: *The Quarry***

Definito un titolo story-driven, *The Quarry* (2022) punta su una qualità grafica iperrealistica e un linguaggio cinematografico, motivo per cui è stata di gran lunga preferita la narrazione ambientale. Il gameplay è semplificato (quick-time event, scelte multiple) per favorire la gestione della trama, che si ramifica in base alle decisioni del giocatore. Un elemento distintivo è la rottura della quarta parete tramite un personaggio esterno alla vicenda (simile al "Curatore" della *Dark Pictures Anthology*) che si rivolge direttamente al giocatore con il "tu", mettendolo di fronte alla responsabilità delle sue scelte. In questi momenti, non esiste una mediazione dell'avatar digitale all'interno del mondo di gioco: il pubblico viene personalmente chiamato in causa a rispondere alle conseguenze delle proprie azioni.

- **Narrazione frammentata ed ambientale: *Dark Souls***

Dark Souls (2011) rappresenta un approccio opposto, dove la trama è velata e deve essere ricostruita attivamente dal giocatore tramite l'esplorazione e la raccolta di indizi. Il codex con le sue descrizioni diventa l'unica fonte onnisciente e affidabile in un mondo criptico. La narrazione, che fa grosso affidamento quindi sulle spiegazioni testuali, è opzionale: il giocatore può scegliere di ignorare la storia o persino uccidere gli NPC (personaggi non giocanti), eliminando con loro porzioni di racconto, esaltando così il libero arbitrio e la componente puramente ludica.

- **Meta-narrazione e illusione della scelta: *The Stanley Parable***

The Stanley Parable (2013) è un'opera che decostruisce il concetto stesso di scelta nei videogiochi. Il protagonista, Stanley, è guidato da un narratore onnisciente; il gameplay si basa sulla decisione di obbedire o ribellarsi a questa voce. Il titolo esplora il conflitto tra l'agire del giocatore e la predeterminazione del codice: anche la ribellione è prevista dagli sviluppatori e a tratti incoraggiata. Come notano Backe e Thon (2019: 1-25), sia la libertà del giocatore che il controllo del narratore sono illusioni, in quanto nessuna delle azioni apparentemente fuori dagli schemi del giocatore potrà mai essere del tutto originale, svelando la natura costrittiva dell'artefatto videoludico che in realtà mantiene sempre il controllo sull'esperienza complessiva.

4. NUOVE FRONTIERE: INTELLIGENZA ARTIFICIALE E IMPEGNO SOCIALE

Il videogioco non è un sistema chiuso, ma dialoga costantemente con altri media e con la realtà sociale, come dimostrato dagli studi sulla Game Culture (Salvador, 2013: 67). Essendo composto per definizione da una matrice informatica, non si allontana mai del tutto dal mondo della tecnologia e dal suo incessante sviluppo, motivo per cui continua ad essere influenzato dalle innovazioni più recenti. Con l'evoluzione dei componenti hardware e la progressiva espansione del pubblico, sono aumentate anche le aspettative sui prodotti finali, obbligando le case produttrici a trovare soluzioni che garantiscano allo stesso tempo costi contenuti, efficienza e qualità. In questo contesto, l'intelligenza artificiale ha trovato un ottimo impiego, trasformandosi in una valida alleata.

L'intelligenza artificiale ha fatto la sua comparsa nel mondo dei videogiochi in tempi relativamente precoci. La sua integrazione, tramite Machine Learning, sta trasformando il gameplay rendendo i nemici adattivi (come in *Alien: Isolation*) o con comportamenti estremamente realistici (*Horizon Zero Dawn*), oppure permettendo interazioni vocali con gli NPC (*Suck Up!*), riuscendo a dare l'impressione di poter parlare realmente con i personaggi di un titolo, con un'accuratezza e contestualità che eguagliano le capacità di elaborazione del linguaggio naturale dei moderni LLM (Large Language Models). Le categorie applicative sono riassumibili in tre ruoli principali, in cui l'intelligenza artificiale diventa cruciale: quello attoriale all'interno del titolo, nel caso gestisca alleati o nemici; un ruolo da designer, quando si ritrova ad adattare i contenuti a seconda delle azioni dei giocatori e dei loro obiettivi; oppure come produttore, nel momento in cui gli sviluppatori videoludici se ne avvalgono per raccogliere dati legati alle esperienze dei giocatori all'interno dei titoli per poterli integrare ad esperienze di intrattenimento più vaste (Aravind et al., 2024). In generale, si può distinguere una forma di AI sistemica, strutturale, quando interviene direttamente sull'architettura tecnica del titolo, sul suo stile simulativo e di riadattamento della realtà, mentre è individuabile una forma più ludonarrativa quando viene applicata ai contesti di interazione e dialogo.

Un utilizzo ormai consolidato dell'intelligenza artificiale all'interno del mondo videoludico è proprio quello di supporto (Wu et al., 2023): buona parte dei processi produttivi richiedono operazioni meccaniche e strettamente tecniche, anche se necessarie. In questo frangente, un "tool" come l'AI permette, in parte, di svolgere queste operazioni fondamentali riducendo le tempistiche ed i margini di errore, consentendo agli sviluppatori di poter ottimizzare il lavoro e potersi concentrare su aspetti più originali e creativi. Il contesto applicativo va dalla stesura del codice informatico alla strutturazione dell'interfaccia grafica più efficace per l'utente. Non si tratta di lasciare in mano all'intelligenza artificiale l'intero sviluppo del gioco, ma di sfruttare le sue potenzialità informatiche lasciando all'uomo la possibilità di concentrarsi su quello che lo rende unico: creare materiale originale e potenzialmente innovativo.
Ciò dimostra che questa tecnologia non solo offre delle opportunità per costruire al meglio mondi di gioco, per esempio sviluppando mappe casuali che rispettino determinate caratteristiche ambientali o migliorando il matchmaking dei titoli multiplayer, ma propone delle interessanti interazioni tra i personaggi interni ed esterni. Uno degli elementi che attira di più i giocatori è proprio l'interattività e, molto spesso, la verosimiglianza delle azioni: l'intelligenza artificiale si configura come strumento chiave per potenziare l'immediatezza delle interazioni uomo-macchina, riducendo il divario tra le aspettative del giocatore e le risposte del sistema. Anche la grafica si sta avvalendo di questo sostegno per cercare di rappresentare ancora più fedelmente la realtà (Centro Sperimentale delle Arti Interattive, 2022). Per questi motivi, si è cominciato a parlare di Game AI quando si fa riferimento a tutti quegli algoritmi e a quelle tecniche di rappresentazione che vengono studiate specificatamente per facilitare l'intrattenimento interattivo digitale in tempo reale (Aravind et al., 2024). Si tratta di tecniche che hanno avuto origine in questo ambito ma che hanno trovato terreno fertile ed una propria strada anche in altri contesti, come il cinema e la televisione, mentre molti altri potrebbero ugualmente prestarsi. Pratiche come quella della motion capture hanno dimostrato la loro efficacia nel portare sullo schermo immagini e movimenti a volte difficilmente scindibili dalla vita reale, mentre l'avvento della realtà virtuale (VR) ha dimostrato la possibile efficacia della sua capacità di creare situazioni virtuali controllate: dalla simulazione di interi metaversi alle applicazioni in campo medico-scientifico (Wu et al., 2023).
Con questi strumenti, le già presenti potenzialità del media vengono ulteriormente enfatizzate, sollevando anche dubbi di natura artisticamente morale. Nel momento in cui parte delle competenze videoludiche vengono lasciate in mano all'intelligenza artificiale, ci si chiede che posto rimanga da coprire alla creatività umana e al suo ingegno: quando si parla di supporto e quando, invece, di sostituire l'artefatto umano con un autore sintetico? I rischi da tenere in considerazione sono diversi e meritano tutti uguale attenzione. Oltre alle questioni etiche legate alla definizione di creatività di un prodotto, non bisogna sottovalutare i sistemi di AI i cui risultati derivano dall'elaborazione di grandi quantità di dati, nei quali possono essere presenti errori, materiali protetti da copyright, incongruenze e bias destinati a riflettersi nel prodotto finale (Wu et al., 2023; Hemraj, 2025). Nella maggioranza dei casi, viene quindi preferito il controllo da parte degli sviluppatori rispetto ad una totale libertà del titolo, così da permettere al pubblico un'esperienza interattiva che possa sempre collocarsi entro certi limiti prestabiliti, garantendo la qualità del prodotto e dei suoi contenuti per rispettare gli standard commerciali. Al momento, i tentativi di sviluppare videogiochi interamente basati sull'AI hanno dimostrato che l'integrazione tra capacità artificiali e progettazione creativa umana rappresenta ancora la soluzione migliore (Hemraj, 2025).
Diverse visioni si possono avere, invece, nei confronti delle intelligenze artificiali puramente generative o delle Adaptive Artificial Neural Network (AANN). In quest'ultimo caso si tratta di strumenti che si basano sul riconoscimento autonomo di pattern grazie ad una fase di addestramento, senza una struttura pre-programmata (Aravind et. al., 2024). È stato dimostrato che il loro utilizzo ha avuto grande successo in diversi campi di studio, ma nel mondo videoludico potrebbe stravolgere alcune strutture preesistenti. Un'analisi condotta e pubblicata da Aravind, Bhatt, Jain, Kumar, Kuppusamy e Veena (2024) ha dimostrato che potrebbero dare ottimi risultati se impiegate in questo ambito, ma si tratta sempre di migliorare l'esperienza ludica del giocatore e non di sovrastarla fino a coprirla: per esempio, un'eccessiva "intelligenza" di un nemico potrebbe portarlo ad una capacità ludica praticamente perfetta che l'utente rischierebbe di trovare impossibile da contrastare, eliminando la componente di intrattenimento; oppure, lasciare che ogni NPC all'interno del titolo possa comportarsi in maniera del tutto indipendente potrebbe portare a delle conseguenze che né il giocatore né lo sviluppatore (tantomeno le capacità tecniche dell'hardware) possono prevedere e gestire, distruggendo l'architettura ludica e narrativa. Un loro adeguato utilizzo permetterebbe un'interattività e un'originalità di gameplay ampissime e sfaccettate, con dialoghi e rose di azioni disponibili tali da consentire una personalizzazione dell'esperienza quasi totale (un esempio può essere *Baldur's Gate 3*), ma il rischio è da tenere in considerazione. Quindi, anche in

questo caso, sembra che l'unione di ingegno e controllo umano assieme all'efficienza artificiale rimanga la soluzione più pratica, mantenendo l'apporto a livello di sostegno produttivo e di una creazione dinamica controllata ineguagliabile. Studi successivi e casi di maggiori impieghi di AI generativa o AANN nei titoli videoludici potranno dimostrare in maniera migliore i risultati di queste prime collaborazioni, aiutando anche a studiare le diverse nature delle interazioni tra uomo e intelligenza artificiale. Infatti, come già visto con il concetto di edutainment, se il gioco può essere un ottimo ponte per l'apprendimento, la dinamica ludica potrebbe aiutare a comprendere meglio come l'uomo scelga di porsi davanti alle diverse forme di intelligenza artificiale (Wu et al., 2023).
In quanto media a tutti gli effetti, ricco di produzioni bottom-up e con un bacino di utenza e community estremamente attivi, il videogioco emerge come potente strumento sociale. Le produzioni e le interazioni da parte del pubblico sono numerose e variegate, aiutate anche da elementi come Internet e i social che permettono una loro radicale espansione. Non a caso, le capacità comunicative audiovisive del media sono state rapidamente colte e sfruttate per creare qualcosa di nuovo, come dimostrano i numerosi festival cinematografici che si sono aperti a produzioni interamente tratte da materiali videoludici come il Milan Machinima Festival[1]. Un altro esempio della sua potenzialità sociale è *The Uncensored Library* su *Minecraft*, un progetto creato da Reporters Sans Frontières. Questa biblioteca virtuale si propone di aggirare la censura governativa in Paesi particolarmente oppressivi da questo punto di vista, rendendo accessibili alcuni documenti altrimenti introvabili. La sola esistenza di un tentativo di questo calibro dimostra come il videogioco possa trascendere l'intrattenimento per diventare un presidio di libertà di informazione, sfruttando mezzi di enorme diffusione (come Internet) non solo per mettere in contatto persone ma anche per condividere informazioni, materiali ed esperienze.
Rientrando in un contesto più tecnico, l'intelligenza artificiale si è dimostrata valida anche per l'aspetto dell'accessibilità. Infatti, Wu, Yi, Ma e Chen (2023) hanno riportato che l'utilizzo di architetture informatiche basate sul deep learning all'interno di un videogioco sono state capaci di riconoscere la gestualità umana in tempo reale, permettendo ad alcuni utenti affetti da disabilità di interagire con l'interfaccia. Si tratta di esempi che dimostrano quanto le potenzialità del mezzo siano ampie e positive, e come queste non si riducano solo all'aspetto ludico.

CONCLUSIONI

Il videogioco è un medium capace di integrare strutture letterarie canoniche (Propp) e cinematografiche in una cornice inedita e interattiva. Non deve essere considerato "visceralmente diverso" dalle forme culturali precedenti, ma un'evoluzione che aggiunge l'elemento dell'azione e della scelta personale alla fruizione del racconto. Che si tratti di reinterpretare classici letterari, sfidare la percezione della realtà con la meta-narrazione, difendere la libertà di stampa o fornire una visione più ludica ad elementi educativi, il videogioco si conferma come una forma di espressione artistica e culturale complessa, viva e in continua evoluzione grazie all'intimo rapporto che ha intrattenuto con la società fin dalla sua creazione.
Relazionandosi strettamente con quest'ultimo e continuando a svilupparsi accanto alle tecnologie più recenti, come l'intelligenza artificiale, si colloca in una posizione vantaggiosa anche se delicata, e solo riconoscendo al videogioco questa specificità è possibile comprendere tutte le sue potenzialità.

BIBLIOGRAFIA

Aarseth, E. (1997). Cybertext: Perspectives on Ergodic Literature. Baltimore: John Hopkins Univesity Press. 1. ISBN 0-8018-5576-0

Aravind, T., Bhatt, C., Jain, P., Kumar, K., Kuppusamy, U., & Veena, N. (2024). Game-changing intelligence: Unveiling the societal impact of artificial intelligence in game software in Entertainment Computing, 52. ISSN: 1875-9521 , https://doi.org/10.1016/j.entcom.2024.100862 (consultato 31/01/2026), https://www.sciencedirect.com/science/article/pii/S1875952124002301#b0115 (consultato 31/01/2026)

Backe, H., & Thon, J. (2019). Playing with Identity. Authors, Narrators, Avatars, and Players in The Stanley Parable and The Beginner's Guide. DIEGESIS. Interdisciplinary E-Journal for Narrative

[1] https://milanmachinimafestival.org/ (cons. 20/04/2026)

Research/Interdisziplinäres E-Journal für Erzählforschung, 8, 1-25. ISSN: 2195-2116, https://pure.itu.dk/ws/portalfiles/portal/84798663/357_1034_4_PB.pdf (consultato 27/01/2026)

Baldassarre, M., Sacco, M., & Di Vagno, R. (2023, aprile). Serious Games a servizio della cultura: per un apprendimento life long. QTimes Journal of Education Technology and Social Studies, 2, 11-25. Anicia editore. ISSN: 2038-3282 www.qtimes.it (consultato 28/01/2026) doi: 10.14668/QTimes_15203

Centro Sperimentale delle Arti Interattive. (2022, aprile). L'Intelligenza artificiale nei videogiochi: un mondo in continua evoluzione. Roma. https://csaiformazione.it/intelligenza-artificiale/ (consultato 27/01/2026)

Ensslin, A. (2014). Literary Gaming. Cambridge and London: The MIT Press. https://doi.org/10.7551/mitpress/9450.001.0001, ISBN: 9780262027151

Fulco, I. (2004). Per una cultura dei videogames. Teorie e prassi del videogiocare. M. Bittanti, Ed.), 65. Milano: Edizioni Unicopoli. ISBN: 88-400-0973-6

Hemraj, S. (2025). AI and the future of creative development: redefining digital media production. Springer, for AI and Ethics, 5, 5105-5119. https://doi.org/10.1007/s43681-025-00765-x (consultato 20/04/2026)

Jenkins, H. (2004). Game Design as Narrative Architecture, for First Person. New Media as Story, Performance, and Game. Wardrip-Fruin, N., Harrigan, P. (Eds.), 118-130. https://paas.org.pl/wp-content/uploads/2012/12/09.-Henry-Jenkins-Game-Design-As-Narrative-Architecture.pdf (consultazione 27/01/2026)

Propp, V. (1928). Morfologia della fiaba.

Salvador, M. (2013). Il videogioco. Brescia: Editrice La Scuola. 67. ISBN: 978-8835034964

Vandewalle, A. (2023, maggio). Video Games as Mythology Museums? Mythographical Story Collections in Games. International Journal of the Classical Tradition, 31, 90-112. https://doi.org/10.1007/s12138-023-00646-w , https://link.springer.com/article/10.1007/s12138-023-00646-w (consultato 27/01/2026)

Wu, Y., Yi, A., Ma, C., & Chen, L. (2023). Artificial intelligence for video game visualization, advancements, benefits and challenges. AIMS Press. Mathematical Biosciences and Engineering, 8, 15345-15373. Doi: 10.3934/mbe.2023686 (consultato 20/04/2026). http://www.aimspress.com/journal/mbe (consultato 20/04/2026)

[3/B] DH per storia, diritto e arte

Archivi ibridi di persona e preservazione forense dei materiali nativi digitali

MARIANGELA GIGLIO[1], ADELE GORINI[2]
[1]UNIVERSITÀ DI BOLOGNA
[2]UNIVERSITÀ DI BOLOGNA

ABSTRACT (ITALIANO)

Il contributo analizza un caso di studio relativo all'archivio dell'ingegnere Luciano Ravaglia (1923–2017), configurabile come archivio ibrido, in cui materiali cartacei inventariati coesistono con un corpus rilevante di supporti digitali legacy, in particolare floppy disk. La ricerca si concentra sulle fasi di ricognizione, preservazione e analisi forense dei materiali digitali, adottando protocolli di imaging orientati alla stabilizzazione dell'evidenza e all'esplorazione dei file system FAT e dello spazio non allocato, considerati come fonti documentarie. Attraverso l'impiego di strumenti di analisi di basso livello sviluppati ad hoc, il contributo riflette sul valore euristico dell'informazione residua e sui limiti interpretativi dell'intenzionalità autoriale, proponendo un approccio integrato tra metodologia archivistica e digital forensics nella costruzione di un archivio digitale criticamente fondato.
Parole chiave: archivi ibridi; preservazione forense; digital forensic; Fondo Luciano Ravaglia

ABSTRACT (ENGLISH)

Hybrid Personal Archives and Forensic Preservation of Born-Digital Materials.
This contribution examines a case study concerning the archive of the engineer Luciano Ravaglia (1923–2017), which can be characterized as a hybrid archive, in which well-inventoried paper materials coexist with a substantial corpus of legacy digital media, particularly floppy disks. The research focuses on the phases of survey, preservation, and forensic analysis of digital materials, adopting imaging protocols aimed at stabilizing the evidence and at exploring FAT file systems and unallocated space, treated as documentary sources in their own right. Using ad hoc low-level analytical tools, the study reflects on the heuristic value of residual information and on the interpretive limits imposed by authorial intentionality, proposing an integrated approach that brings archival methodology and digital forensics into dialogue in the construction of a critically grounded digital archive.
Keywords: hybrid archives; forensic preservation; digital forensics; Luciano Ravaglia Archive

1. INTRODUZIONE

Negli ultimi anni in Italia, numerosi progetti di ricerca si sono concentrati sul *born-digital* e sugli archivi ibridi (Allegrezza, Gorgolini, 2016; Carbé, 2023; Antonelli et al., 2025), mettendo in luce la complessità degli archivi di persona. Nonostante la crescente attenzione accademica gli interventi archivistici di riordino e inventariazione continuano generalmente a concentrarsi sulla documentazione cartacea, escludendo informazioni e supporti nativi digitali dalle attività di analisi e descrizione. Tale mancanza non è sempre riconducibile a una carenza di sensibilità o attenzione da parte degli archivisti incaricati, ma deriva piuttosto da una sensazione di smarrimento davanti agli oggetti digitali che, come ampiamente dimostrato, non possono essere attenzionati seguendo un approccio archivistico tradizionale (Gorini & Giagnolini, 2026; Giglio, 2025). In assenza di adeguate risorse e conoscenze tecniche/specialistiche, la componente digitale risulta, nella migliore delle ipotesi, oggetto di una descrizione sommaria, limitata alla registrazione di informazioni di contesto, come etichette e copertine dei supporti. Sebbene questa modalità non risponda a criteri di salvaguardia a lungo termine può risultare preziosa poiché agevola l'individuazione di materiali digitali e facilita la programmazione di successive attività di indagine.[1]
In questo contesto si inserisce la presente ricerca, concentrata sul recupero e la messa in sicurezza dei materiali nativi digitali presenti nell'archivio dell'Ingegnere Luciano Ravaglia di Forlì (1923-2017) già

[1] Su un'ipotesi di mappatura degli archivi nativi digitali di scrittori e letterati, si veda anche il progetto del gruppo di ricerca ALDiNa presentato al convegno BDCAM25: S. Allegrezza, F. Boschetti, E. Carbé, A. Del Bianco, L. Giagnolini, M. Giglio, A. Gorini, T. Mancinelli and F. Tomasi, "Mapping for Understanding: the ALDiNa Project", in BDCAM25: Born-Digital Collections, Archives And Memory, Londra, 03/04/2025. DOI: 10.5281/zenodo.15105904 (ultima consultazione gennaio 2026).

ampiamente descritto e inventariato nella sua parte cartacea nel corso del 2022 a cura della Società Cooperativa Ebla di Cesena.[2] Il fondo testimonia l'intensa attività di progettazione dell'ingegnere forlivese, svolta in ambito nazionale e internazionale dagli anni Quaranta del '900 fino al 2017.[3]
L'inventario realizzato dalla Cooperativa cesenate ha avuto il pregio di descrivere con estrema accuratezza l'eterogeneità tipologica della documentazione: relazioni, disegni tecnici lucidi e su carta, plastici, carteggio, materiali fotografici e audiovisivi in vari formati e supporti e volumi afferenti ad una biblioteca personale con monografie e periodici tecnici. Nonostante l'intervento archivistico fosse finalizzato esclusivamente al trattamento della parte cartacea, l'archivista incaricata ha provveduto alla segnalazione in inventario anche di numerosi supporti digitali, quali floppy disk, CD e DVD. Grazie a tale strumento è stato quindi possibile individuare le numerose scatole di cartone e plastica contenenti tali supporti e concentrare le attività sul recupero delle informazioni ivi custodite. L'eterogeneità dei materiali rende il fondo un esempio paradigmatico di archivio ibrido di persona.
Gli obiettivi della ricerca erano molteplici: da una parte, ampliare la comprensione complessiva del fondo, confermare il valore documentario del materiale digitale, nonché ribadire la natura ibrida dell'archivio; dall'altra, completare la ricostruzione del profilo del soggetto produttore, restituendo un'immagine di Luciano Ravaglia come figura profondamente moderna, capace di utilizzare strumenti informatici e software di progettazione – come AutoCAD – o i primi programmi di word processing, come WordStar.
La componente digitale del fondo consta complessivamente di 234 floppy disk (92 da 3,5" e 142 da 5,25"), circa 30 CD-ROM e 10 DVD. Il nuovo intervento si è concentrato in una prima fase su una ricognizione approfondita dei floppy disk, originariamente organizzati dal soggetto produttore in scatole di cartone o contenitori in plastica, spesso corredati da etichette manoscritte. È stato redatto uno strumento di corredo preliminare, finalizzato a favorire l'accesso e la futura attività di copia forense dei supporti, che riporta per ciascuna unità: il numero di corda attribuito all'unità archivistica nel precedente inventario e il relativo titolo (quando presente), un numero identificativo assegnato a ciascuna unità documentaria (intesa come singolo floppy) e la trascrizione integrale delle etichette.
La scelta di avviare l'attività dai floppy da 3,5" è stata motivata sia dalla fragilità di tali supporti, sia dalla più agevole reperibilità della strumentazione necessaria alla loro lettura e duplicazione. Attraverso questa fase di schedatura preliminare è stato possibile individuare un nucleo documentario originario e apparentemente coerente, dal quale avviare la successiva attività di analisi forense. In particolare, è stata selezionata l'unità archivistica n. 1748, intitolata "Progetto di Nuovo Complesso di San Francesco d'Assisi. Curia di Imola. Copia floppy lavoro", composta da 11 floppy disk da 3,5", tutti riconducibili all'ultimo progetto realizzato da Luciano Ravaglia tra la fine degli anni Novanta e il 2015[4]. Tale materiale digitale integra e completa la sottoserie cartacea di riferimento, costituita da 59 buste, 9 fascicoli e 32 quaderni**.**

2. PRESERVAZIONE FORENSE E OBSOLESCENZA DEI SUPPORTI MAGNETICI

Nell'ambito dei fondi *born-digital* legacy, la preservazione dei supporti magnetici obsoleti costituisce una condizione preliminare imprescindibile per qualsiasi operazione di analisi, descrizione e interpretazione dei documenti digitali. In particolare, i floppy disk, largamente utilizzati tra gli anni Ottanta e Novanta come dispositivi di memorizzazione per la produzione e la circolazione dell'informazione, presentano oggi un insieme di criticità che ne rendono l'accesso sempre più problematico (Durno, 2016; Waugh, 2014). A oltre trent'anni dalla loro massima diffusione, l'illeggibilità dei floppy disk va interpretata non solo come conseguenza della fragilità fisica del supporto, ma come effetto combinato del degrado materiale e della progressiva obsolescenza dell'ecosistema hardware e software necessario alla loro lettura (Gough, 2013; McKinley, 2015; Dicks & Iraci, 2017; Waugh, 2014). In questo quadro, l'acquisizione dei dati risulta fortemente condizionata da fattori tecnici eterogenei – dal formato e file system alle caratteristiche dell'unità di lettura – e richiede procedure di verifica iterative, confermando l'obsolescenza come un

[2] Sull'argomento si veda https://www.romagna.camcom.it/it/informazioni/la-camera-di-commercio-della-romagna/il-patrimonio-documentale/archivio-ravaglia-1 e https://www.eblacoop.it/. (ultima consultazione gennaio 2026).
[3] Donato per testamento olografo alla Camera di Commercio, Industria, Artigianato e Agricoltura di Forlì-Cesena – già sede dello studio dell'ingegnere – l'archivio è stato dichiarato di interesse storico particolarmente importante nel 2013 ai sensi del D.Lgs. 42/2004.
[4] https://www.sfrancesco.it/introduzione-alla-chiesa-di-san-pio/, anche archiviato presso https://web.archive.org/web/20251217141439/https%3A%2F%2Fwww.sfrancesco.it%2Fintroduzione-alla-chiesa-di-san-pio%2F.

processo stratificato che intreccia dimensioni materiali, logiche e tecnologiche (Durno, 2016; McKinley, 2015).
Alla luce di tale fragilità strutturale, il trasferimento delle informazioni da floppy disk verso sistemi di conservazione contemporanei si configura come un'operazione non procrastinabile (Waugh, 2014). Tuttavia, una semplice copia a livello di file risulta insufficiente, poiché espone il materiale a rischi di perdita informativa e cancella elementi strutturali potenzialmente rilevanti per la comprensione del contesto documentario. È in questo quadro che la comunità archivistica ha progressivamente adottato metodologie mutuate dall'ambito della *digital forensics*[5], individuando nella pratica dell'imaging del disco una risposta metodologicamente fondata alle esigenze di preservazione e analisi dei supporti obsoleti. L'imaging forense consiste nella realizzazione di una copia integrale bit per bit del supporto fisico, concepita non come duplicato funzionale orientato all'uso, ma come forma di stabilizzazione dell'evidenza originale (Lee et al., 2013). Tale procedura consente di ridurre drasticamente l'intervento diretto sul supporto, limitando i rischi connessi alla sua manipolazione, e di trasferire l'oggetto di lavoro entro un ambiente di conservazione e consultazione controllabile. Al tempo stesso, essa permette di preservare non solo i file logicamente accessibili, ma anche le strutture del file system e i dati residuali, spesso decisivi per la ricostruzione delle pratiche di produzione e gestione dei documenti digitali (Ries, 2018). Attraverso l'adozione di protocolli fondati su criteri di integrità, non invasività e completezza, e sull'attenzione alla tracciabilità delle operazioni, l'imaging forense si configura dunque come il primo passo per la salvaguardia dei materiali nativi digitali conservati su supporti magnetici obsoleti. Stabilizzando il dato e delimitando in modo controllabile il perimetro dei testimoni digitali, esso crea le condizioni necessarie per le successive attività di analisi strutturale e interpretazione documentaria, che costituiscono il nucleo dell'indagine presentata nelle sezioni seguenti.

3. ANALISI FORENSE ED ESTRAZIONE DI INFORMAZIONE STRUTTURALE

Una volta stabilizzato il materiale attraverso l'imaging forense, l'analisi dei supporti digitali del fondo Ravaglia si è concentrata sull'estrazione di informazione strutturale dai file system e dai documenti digitali. In questa fase, l'attenzione si è spostata dal supporto fisico all'organizzazione logica dei dati, trattando il file system e i formati digitali come fonti documentarie a pieno titolo.
Un primo livello di indagine ha riguardato l'esplorazione diretta dei file system FAT,[6] ampiamente diffusi nei floppy disk analizzati. In tali sistemi, la cancellazione di un file non comporta la rimozione immediata dei dati dal supporto, bensì la modifica delle strutture di indicizzazione che ne consentono l'individuazione da parte del sistema operativo. Il contenuto del file diviene così logicamente non più accessibile, pur potendo permanere materialmente sul supporto fino a un'eventuale sovrascrittura. L'analisi delle *directory entries* e delle tabelle di allocazione consente pertanto di individuare file cancellati, di ricostruirne in parte la struttura e, in alcuni casi, di recuperare informazioni relative al nome originario, alla dimensione e alla catena dei cluster precedentemente occupati. L'esplorazione di tali strutture è stata condotta mediante un software[7] sviluppato appositamente per l'analisi forense dei file system FAT basato sulla lettura diretta delle strutture di basso livello del disco. Lo strumento consente di analizzare separatamente il Volume Boot Record, le tabelle FAT primarie e secondarie e l'immagine completa del supporto, ricostruendo la geometria logica del volume e determinando il tipo di file system sulla base dei parametri strutturali. Attraverso il parsing sistematico delle *directory entries*, il software permette di individuare file logicamente cancellati a partire dai marcatori sintattici propri del file system e di includere opzionalmente tali elementi nel catalogo risultante. La gestione esplicita delle Long File Names consente, in numerosi casi, di ricostruire i nomi originari anche per file cancellati, evidenziando come la cancellazione in ambiente FAT non equivalga a una rimozione semantica completa, ma piuttosto a una sospensione dello statuto logico del file. In questo modo, l'analisi del file system diviene uno strumento di

[5] Definita come l'impiego scientifico di metodi comprovati per preservare, validare, identificare, analizzare, interpretare, documentare e presentare evidenze digitali provenienti da computer, telefoni, reti e servizi cloud. Sul tema si veda almeno Kaur, Ravneet, e Amandeep Kaur. 2012. «Digital Forensics». *International Journal of Computer Applications* 50 (5): 5–9. https://doi.org/10.5120/7765-0844.
[6] FAT, acronimo di *File Allocation Table*, designa una famiglia di file system sviluppata da Microsoft e ampiamente utilizzata, soprattutto tra gli anni Ottanta e Novanta, su floppy disk, hard disk e altri supporti rimovibili. La sua struttura si basa su una tabella di allocazione che registra la distribuzione dei cluster sul disco e le relazioni tra essi, consentendo al sistema operativo di localizzare i file e di ricostruirne la continuità logica.
[7] Il codice è disponibile su Github a https://github.com/mariangiglio/FAT_forensic_explorer (ultima consultazione aprile 2026).

accesso a informazioni cronologiche, strutturali e operative rilevanti per la ricostruzione delle pratiche di produzione e gestione dei documenti digitali.
Un secondo livello di analisi ha interessato lo spazio non allocato, ovvero le aree del supporto non associate a file logicamente attivi.[8] L'esame di tali porzioni si è rivelato particolarmente rilevante per l'individuazione di dati residuali riconducibili a documenti digitali nativi parzialmente cancellati o frammentati, sebbene non abbia condotto, nella maggior parte dei casi, al recupero completo e coerente dei contenuti originari. Anche in presenza di risultati quantitativamente limitati, questa fase ha fornito indicazioni significative sullo stato dei supporti e sulle pratiche di cancellazione e riutilizzo dello spazio di memorizzazione.
A differenza delle tecniche di *data carving* generaliste,[9] l'approccio adottato si è basato su uno strumento software sviluppato appositamente per il riconoscimento della semantica applicativa di specifici formati di interesse, in particolare i file di progettazione DWG.[10] Il software opera direttamente su immagini di spazio non allocato e su blob binari, individuando firme strutturali, marcatori interni e intestazioni riconducibili a versioni storiche di AutoCAD. Questo approccio mirato ha consentito di distinguere con maggiore attendibilità tra dati casuali e frammenti documentari significativi, riducendo il rischio di interpretazioni arbitrarie tipiche del carving generalista. Attraverso tale procedura sono stati effettivamente individuati ed estratti alcuni file DWG parzialmente ricostruibili. Tuttavia, tali materiali non sono stati oggetto di ulteriore analisi né esposti nel presente lavoro, al fine di non eccedere l'intenzionalità autoriale e di rispettare il carattere privato e operativo di documenti che, per loro natura, si collocano in una fase intermedia e non pubblica del processo progettuale. La loro individuazione mantiene comunque piena rilevanza metodologica, in quanto dimostra la persistenza di tracce documentarie significative nello spazio non allocato e conferma la validità dell'approccio adottato. Nel loro insieme, gli strumenti software impiegati non sono concepiti come soluzioni automatizzate o generaliste, ma come dispositivi euristici orientati alla riduzione della complessità del dato. Essi operano selezionando aree significative, evidenziando pattern ricorrenti, anomalie e marcatori applicativi che possono essere ulteriormente indagati attraverso l'analisi manuale. In questa prospettiva, il software non sostituisce l'interpretazione archivistica e storica, ma ne costituisce una condizione di possibilità, rendendo praticabile l'esplorazione di documenti digitali opachi e privi di metadati espliciti.
In questo quadro, anche la progettazione degli strumenti di analisi assume un rilievo etico non secondario. Il software sviluppato per l'individuazione di file DWG non è un dispositivo di recupero massivo, ma uno strumento di identificazione strutturale, orientato a riconoscere la presenza di tracce documentarie senza forzarne la ricostruzione integrale. Il riconoscimento delle firme di intestazione e dei marcatori interni consente di stabilire l'appartenenza semantica dei dati residui a uno specifico contesto applicativo, ma lascia deliberatamente aperta la questione del loro statuto documentario e della loro legittimità interpretativa, contribuendo a definire il perimetro di ciò che può essere legittimamente considerato oggetto di analisi. L'emergere di file DWG parzialmente recuperabili dallo spazio non allocato, pur confermando la persistenza materiale di tracce progettuali non più accessibili a livello logico, ha posto lo studioso di fronte alla necessità di distinguere tra possibilità tecnica e legittimità interpretativa. La decisione di non procedere a un'analisi approfondita di tali materiali non deriva dunque da un limite strumentale, ma da una scelta consapevole, volta a evitare che l'indagine forense si traduca in una violazione dell'intenzionalità autoriale o in una sovra-interpretazione di documenti che l'autore aveva esplicitamente abbandonato.

[8] Per "spazio non allocato" si intendono le porzioni del supporto di memorizzazione che il file system considera al momento libere, e dunque disponibili per la scrittura di nuovi dati. Tali aree, tuttavia, non sono necessariamente prive di contenuto: esse possono conservare tracce residue di file precedentemente presenti, cancellati solo sul piano logico ma non ancora sovrascritti sul piano materiale. Proprio per questo, lo spazio non allocato costituisce un ambito di particolare interesse per l'analisi forense, poiché può restituire frammenti testuali, strutture documentarie incomplete o altri elementi utili alla ricostruzione delle precedenti configurazioni del supporto.
[9] Per *data carving* si intende una tecnica di recupero dei dati che consente di individuare e ricostruire file o frammenti di file direttamente a partire dal contenuto binario del supporto, senza fare affidamento sulle informazioni di indicizzazione fornite dal file system. Nelle sue forme più comuni, il *data carving* si basa sul riconoscimento di firme caratteristiche, intestazioni, strutture interne o sequenze ricorrenti proprie di determinati formati. Le tecniche generaliste mirano di norma a identificare un ampio spettro di tipologie documentarie sulla base di criteri formali relativamente generici; l'approccio qui adottato, invece, si fonda su un riconoscimento più mirato della struttura e della logica interna di specifici formati, così da aumentare la precisione dell'analisi in relazione ai materiali effettivamente oggetto di studio.
[10] Lo script è visualizzabile su Github a https://github.com/mariangiglio/dwg_unallocated_carver (ultima consultazione aprile 2026).

4. CONCLUSIONI

L'analisi delineata nel paragrafo 2 ha consentito di approfondire in modo significativo la comprensione della componente digitale dell'archivio, permettendo di formulare prime valutazioni di ordine quantitativo e qualitativo, con particolare riferimento al nucleo archivistico con segnatura 1748, composto da 11 floppy disk per un volume complessivo di circa 10 MB, dei quali due risultano illeggibili nonostante i tentativi di acquisizione. La definizione degli estremi cronologici del nucleo presenta livelli di affidabilità differenziati: l'estremo post quem può essere collocato con ragionevole sicurezza al 1988, sulla base dei metadati e dei contesti applicativi dei file, mentre l'estremo ante quem risulta meno determinabile a causa di successive manipolazioni che hanno compromesso l'attendibilità delle date di modifica. Con le opportune cautele, l'ultima data considerabile attendibile può essere collocata intorno al 2004. La distribuzione dei file per tipologia di estensione riflette in modo coerente le pratiche di lavoro dell'autore. Il nucleo è caratterizzato dalla presenza prevalente di file di progetto in formato DWG, talvolta privi di estensione esplicita, e di archivi compressi ZIP, affiancati in misura minore da file temporanei, eseguibili e documenti di testo. Tale composizione conferma il carattere eminentemente operativo del materiale digitale e ne rafforza l'integrazione funzionale con la documentazione cartacea di riferimento, contribuendo alla ricostruzione delle modalità di produzione e gestione dei progetti professionali dell'ingegnere Ravaglia.
La presente ricerca intende dimostrare come, se applicate nel *respect des fonds*, alcune pratiche dell'informatica forense possano svolgere una funzione ancillare rispetto all'archivistica tradizionale. In questa prospettiva, l'analisi forense non si configura come uno strumento invasivo o totalizzante, ma come un supporto metodologicamente orientato, il cui impiego deve essere calibrato in funzione delle esigenze descrittivi e interpretative dell'archivistica. La decisione di interrompere l'analisi forense e l'indagine su informazioni cancellate, sovrascritte e su altri contenuti avantestuali – o, nel presente caso, "avangrafici" – dovrebbe sempre essere orientata al rispetto dell'intenzionalità autoriale. Nel caso esaminato da questa ricerca, i documenti analizzati erano ascrivibili ad un'opera compiuta, effettivamente costruita - quale il Complesso di San Francesco a Imola. Per tale motivo, una ricerca sistematica e indiscriminata di versioni precedenti, dettagli irrealizzati e idee primordiali avrebbe stressato l'intenzionalità ultima del progettista, che – a differenza di quella dello scrittore il cui esito finale si concretizza nel testo pubblicato – è ufficializzata dall'esecuzione fisica e cementizia. Si aggiunga poi che l'intenzionalità grafica e testuale dei documenti, così come apparivano salvati nei floppy analizzati, ha contribuito alla costruzione di legami non solo interni ai singoli file, ma anche in relazione al progetto nella sua totalità (K. Mitchell, 2008). Un'analisi eccessivamente invasiva, volta a scardinare e disarticolare ogni documento o supporto, avrebbe significato alterare il vincolo esistente tra i file e una violazione della volontà autoriale del progettista. Per questo motivo è importante saper dosare la potenza dell'analisi forense, calibrando in funzione di obiettivi specifici: da un lato, la ricostruzione di dinamiche d'interesse "filologico" ed evolutivo utili alla comprensione del "cantiere" di lavoro del soggetto produttore, le cui tracce nel contesto born-digital rimangono incorporate e invisibili senza il ricorso a strumenti di analisi più sofisticati; dall'altro, il recupero di informazioni essenziali alla leggibilità dei documento in presenza di lacune dovute a perdite o corruzioni di bit. Tale bilanciamento consente di preservare la forma, il contenuto e l'organizzazione che l'autore aveva intenzionalmente conferito al documento o al supporto.
In questa prospettiva, i dati ricavati dall'analisi forense si configurano come un livello informativo aggiunto: essi saranno integrati nella descrizione archivistica già presente sulla piattaforma xDams, contribuendo non solo a rendere lo strumento di corredo più completo ed esaustivo, ma anche alla valorizzazione di quegli elementi che non emergevano dalla sola documentazione tradizionale. Nonostante, in questo caso, molti materiali digitali trovino un corrispettivo diretto nelle unità archivistiche cartacee , l'analisi ha permesso di individuare elementi ulteriori, quali nomi di collaboratori coinvolti, tracce delle pratiche progettuali adottate e dell'ambiente tecnologico di creazione. Questi elementi ampliano la comprensione del contesto professionale dell'ingegner Ravaglia e, al contempo, estendono le possibilità di narrazione, interpretazione e riuso dei materiali, in linea con quanto evidenziato da precedenti ricerche nell'ambito degli archivi digitali di architettura (Lynn, 2014; Molenda, 2018).

RINGRAZIAMENTI

Questo studio è stato condotto nell'ambito del progetto di ricerca dottorale di Adele Gorini, finanziato dall'Unione europea – Next Generation EU, Investimento I.4.1, borsa di dottorato PNRR per il Patrimonio

Culturale, Decreto Ministeriale n. 118/2023. Le autrici desiderano inoltre ringraziare la Società Cooperativa EBLA per il supporto offerto alla ricerca e, in particolare, la Camera di Commercio, Industria, Artigianato e Agricoltura di Forlì-Cesena per aver agevolato l'accesso alla documentazione.

BIBLIOGRAFIA

Allegrezza, S., Boschetti, F., Carbè, E., Del Bianco, A., Giagnolini, L., Giglio, M., Gorini, A., Mancinelli, T., & Tomasi, F. (2025). Mapping for understanding: The ALDiNa Project. In *Born-digital collections, archives and memory* (Poster proceedings). London, United Kingdom, April 2–4, 2025.

Allegrezza, S., & Gorgolini, L. (2016). *Gli archivi di persona nell'era digitale: Il caso dell'archivio di Massimo Vannucci*. Il Mulino.

Antonelli, G., Giagnolini, L., & Milone, F. (2025). *The intangible papers: Authorial philology and born-digital texts*. Il Mulino.

Carbè, E. (2023). *Digitale d'autore: Macchine, archivi, letteratura*. Firenze University Press; USiena Press.

Dicks, C., & Iraci, J. (2017). *Computer hard disks and diskettes – FAQ*. Canadian Conservation Institute. https://www.canada.ca/en/conservation-institute/services/care-objects/electronic-media/computer-hard-disks-diskettes-faq.html#faq1

Durno, J. (2016). Digital archaeology and/or forensics: Working with floppy disks from the 1980s. *The Code4Lib Journal*, 34. https://journal.code4lib.org/articles/11986

Giglio, M. (2025). Metodologie computazionali per l'organizzazione di archivi nati digitalmente. In *Diversità, equità e inclusione: Sfide e opportunità per l'informatica umanistica nell'era dell'intelligenza artificiale* (Atti del XIV Convegno Annuale AIUCD, pp. 208–214). Verona, Italy, June 11–13, 2025.

Gorini, A., & Giagnolini, L. (2026). *Under the surface: Interpreting technical metadata as archival narrative*. In *Proceedings of the 21st International Conference on Digital Preservation (iPRES 2025)*. Wellington, Aotearoa New Zealand, https://phaidra.univie.ac.at/o:2198090.

Gough, L. (2013). *Project KryoFlux—Part 2: Why bother with it?* https://goughlui.com/2013/04/21/project-kryoflux-part-2-why-bother-with-it/

Kaur, Ravneet, e Amandeep Kaur (2012). «Digital Forensics». *International Journal of Computer Applications* 50 (5): 5–9. https://doi.org/10.5120/7765-0844.

Lee, C., Kirschenbaum, M., Chassanoff, A., Olsen, P., & Woods, K. (2012). BitCurator: Tools and techniques for digital forensics in collecting institutions. *D-Lib Magazine*, 18 (5–6). https://doi.org/10.1045/may2012-lee

Lynn, G. (Ed.). (2014). *The archaeology of the digital*. Sternberg Press; Canadian Centre for Architecture. ISBN: 9783943365801.

Mitchell, K. (2008). *Intention and Text : Towards an Intentionality of Literary Form*. Continuum. ISBN:978-1441198648.

McKinley, M. (2015). Imaging digital media for preservation with LAMMP. *The Electronic Media Review*, 3, 89–96.

Molenda, A. (2018). *Between creators and keepers: How HNI builds its digital archive*. In N. McGovern & A. Whiteside (Eds.), *Proceedings of the 15th International Conference on Digital Preservation, iPRES 2018* (Boston, MA, USA). iPRES.

Ries, T. (2018). The rationale of the born-digital *dossier génétique*: Digital forensics and the writing process: With examples from the Thomas Kling Archive. *Digital Scholarship in the Humanities*, 33 (2), 391–424.

Digital Humanities e studi neogreci: il progetto della Biblioteca Digitale dell'Osservatorio "Mario Vitti"

GEORGIOS KATSANTONIS
INDIPENDENT RESEARCHER, GEORGIOS.KATSANTONIS@ALUMNI.SNS.IT

ABSTRACT (ITALIANO)

Questo contributo si propone di presentare il progetto in corso per la realizzazione di una digital library dedicata alla letteratura neogreca in Italia dagli inizi del XIX secolo fino ai nostri giorni. L'Osservatorio Neogreco, finanziato dall'Università Sapienza di Roma, è una banca dati elettronica che raccoglie informazioni su pubblicazioni in lingua italiana relative alla cultura greca moderna. Il progetto si inserisce nell'ambito dei più recenti studi internazionali sulla filologia digitale, sull'archiviazione digitale dei materiali raccolti durante la ricerca sul campo, sulla loro condivisione scientifica nel sito tematico disponibile in rete. Sono queste premesse fondamentali allo sviluppo e al potenziamento di metodi di catalogazione online e allo studio dei materiali, anche in forme comparative. La relazione illustrerà le linee di intervento seguite e le motivazioni del progetto; discuterà inoltre gli obiettivi e i risultati della ricerca bibliografica e le attività principali dell'Osservatorio Neogreco, nonché le criticità infrastrutturali e di sostenibilità. La relazione metterà infine in evidenza come, attorno a questo nucleo fondamentale di una Biblioteca Digitale di Neogreco in italiano, trasferito dalla carta al digitale, si sia costruita una raccolta preziosa, un palinsesto variegato di volumi, articoli, traduzioni e recensioni apparsi su periodici scientifici e divulgativi, quotidiani e riviste che valorizzino la lunga tradizione di interesse che l'Italia ha manifestato nei confronti della produzione letteraria del mondo neogreco.
Parole chiave: Digital Humanities; biblioteche digitali; infrastrutture di ricerca; interoperabilità; studi neogreci

ABSTRACT (ENGLISH)

The Digital Library of the "Mario Vitti" Neo-Greek Observatory is presented as a case study for reflecting on research infrastructures in the field of Digital Humanities. Rather than being considered merely as a bibliographic repository, the project is analysed as an epistemic device capable of modelling the Italian scholarly production on Modern and Contemporary Greek studies. The project aligns with current international research in digital philology, focusing on the digital preservation of materials gathered during field research and their scholarly dissemination through an online platform. These efforts support the development of online cataloging methods and enable comparative studies of the collected materials. The paper outlines the intervention strategies and motivations behind the project, the objectives and outcomes of the bibliographic research, and the main activities of the Observatory, as well as infrastructure challenges and long-term sustainability considerations. It further highlights how, around the core of the Italian Neo-Hellenic Digital Library, a rich collection of books, articles, translations, and reviews published in academic journals, popular periodicals, and newspapers has been assembled, showcasing Italy's longstanding interest in Modern Greek literary production.
Keywords: Digital Humanities; digital libraries; research infrastructures; interoperability; Neo-Greek studies

1 INTRODUZIONE

Nel dibattito contemporaneo sulle *Digital Humanities*, l'attenzione si è progressivamente spostata dalla semplice digitalizzazione dei contenuti alla progettazione di infrastrutture della conoscenza, intese come ambienti digitali in grado di modellizzare, organizzare e rendere interrogabile il sapere umanistico. In questa prospettiva si colloca la Biblioteca Digitale dell'Osservatorio Neogreco "Mario Vitti", un progetto dedicato alla documentazione sistematica degli studi sulla cultura greca moderna e contemporanea in Italia. Il contributo presenta e analizza il progetto come infrastruttura digitale di ricerca, mettendone in luce le implicazioni epistemologiche e metodologiche nel contesto delle Digital Humanities. In particolare, la relazione illustra le linee di intervento seguite e le motivazioni che hanno guidato la progettazione

dell'Osservatorio, discutendo gli obiettivi e i risultati della ricerca bibliografica, le principali attività svolte, nonché le criticità infrastrutturali e le sfide legate alla sostenibilità a lungo termine.

1.1 Genesi dell'Osservatorio e del progetto digitale

L'*Osservatorio Neogreco "Mario Vitti"* è un'iniziativa culturale e scientifica finanziata e promossa sotto l'egida della Sapienza Università di Roma, istituita presso il Dipartimento di Studi Europei, Americani e Interculturali. Il progetto nasce con l'obiettivo di raccogliere, catalogare e rendere accessibili le fonti bibliografiche italiane relative alla cultura greca moderna e contemporanea, coprendo un arco cronologico che va prevalentemente dal XIX secolo fino ai giorni nostri. La Biblioteca Digitale raccoglie e organizza l'insieme delle pubblicazioni italiane dedicate alla letteratura neogreca e greco-cipriota — incluse monografie, articoli, traduzioni e recensioni apparse su riviste scientifiche, periodici e quotidiani — attraverso un sistema di schedatura bibliografica avanzata, concepito per favorire l'interrogazione, la riusabilità dei dati e l'integrazione con altri ambienti digitali di ricerca. L'Osservatorio prende il nome da Mario Vitti (1926–2023)[1]. Intitolare all'illustre filologo questo osservatorio non è un semplice atto commemorativo, ma una dichiarazione di intenti: continuare un percorso di ricerca, diffusione e valorizzazione della letteratura neoellenica attraverso gli strumenti digitali, promuovendo la conoscenza, l'accesso e l'interconnessione tra studiosi, studenti e appassionati di lingue e letterature moderne.
Il progetto si configura come un'importante risorsa scientifica per l'analisi, la documentazione e la consultazione delle pubblicazioni in lingua italiana che si occupano della Grecia moderna, dei suoi autori e delle sue reti culturali. L'Osservatorio non si limita a custodire tali materiali: li inserisce in una Biblioteca Digitale strutturata, finalizzata a renderli interrogabili attraverso metadati coerenti e standardizzati e a costituire un archivio per la storia degli studi neogreci. L'inserimento degli studi neogreci all'interno delle Digital Humanities rappresenta un passaggio strategico. Tradizionalmente collocati in un ambito filologico-letterario, gli studi sulla Grecia moderna hanno spesso sofferto di una marginalità disciplinare. La Biblioteca Digitale dell'Osservatorio "Mario Vitti" contribuisce a superare tale marginalità attraverso la formalizzazione digitale del sapere, trasformando la bibliografia in un oggetto di analisi comparativa e sistematica. La Biblioteca Digitale prende forma come risposta alla necessità di preservare e rendere accessibile questo patrimonio, trasformandolo in una risorsa interrogabile e contestualizzata, capace di supportare la ricerca storica e filologica.

1.2 Obiettivi consolidati della biblioteca digitale

Il nucleo operativo dell'Osservatorio è costituito da una biblioteca digitale tematica che raccoglie informazioni bibliografiche su pubblicazioni, studi, traduzioni e ricerche legate alla cultura neogreca in lingua italiana. L'archivio include riferimenti a monografie, articoli, traduzioni, tesi di laurea e di dottorato, recensioni e altri documenti di rilevanza accademica relativi alla modernità greca e alla sua ricezione culturale nel contesto italiano. Questa funzione di catalogazione specialistica permette a studenti, docenti e ricercatori di localizzare rapidamente materiali rilevanti e correlati tramite criteri di ricerca mirati, facilitando così il lavoro di studio e di analisi comparata tra opere e contributi critici.
L'Osservatorio non si limita a fungere da semplice archivio bibliografico, ma aspira a diventare una vera e propria infrastruttura digitale di ricerca. La sua missione principale è la creazione di una Biblioteca Digitale di Neogreco in italiano, ovvero un ambiente di consultazione e studio che valorizzi la produzione italiana sulla cultura neogreca e grecocipriota[2], mettendo a disposizione informazioni su volumi, articoli, traduzioni e recensioni apparsi su riviste accademiche, periodici divulgativi e quotidiani. Gli utenti possono effettuare ricerche bibliografiche semplici o incrociate attraverso maschere di interrogazione e consultare, in molti casi, anche il testo digitale integrale laddove non vi siano limitazioni legate ai diritti d'autore.

[1] La sua *Storia della letteratura neogreca* ha definito i fondamenti della disciplina, tracciando un percorso che va dall'età moderna ai principali poeti contemporanei. Pubblicata per la prima volta nel 1971 da ERI e successivamente aggiornata nella riedizione Carocci del 2001, l'opera integra storia letteraria, critica testuale e contestualizzazione culturale, diventando un riferimento imprescindibile per chi studia la letteratura neoellenica e le sue connessioni transnazionali (Vitti, 1971; 2001). Il libro è stato tradotto e pubblicato anche in greco con il titolo Ἱστορία τῆς νεοελληνικῆς λογοτεχνίας (Odysseas, 1978), rivelandosi un riferimento internazionale nei corsi universitari di filologia e letteratura greca moderna.
[2] https://lsn.seai.uniroma1.it/osservatorio/presentazione-obiettivi (cons. 22/01/2026)

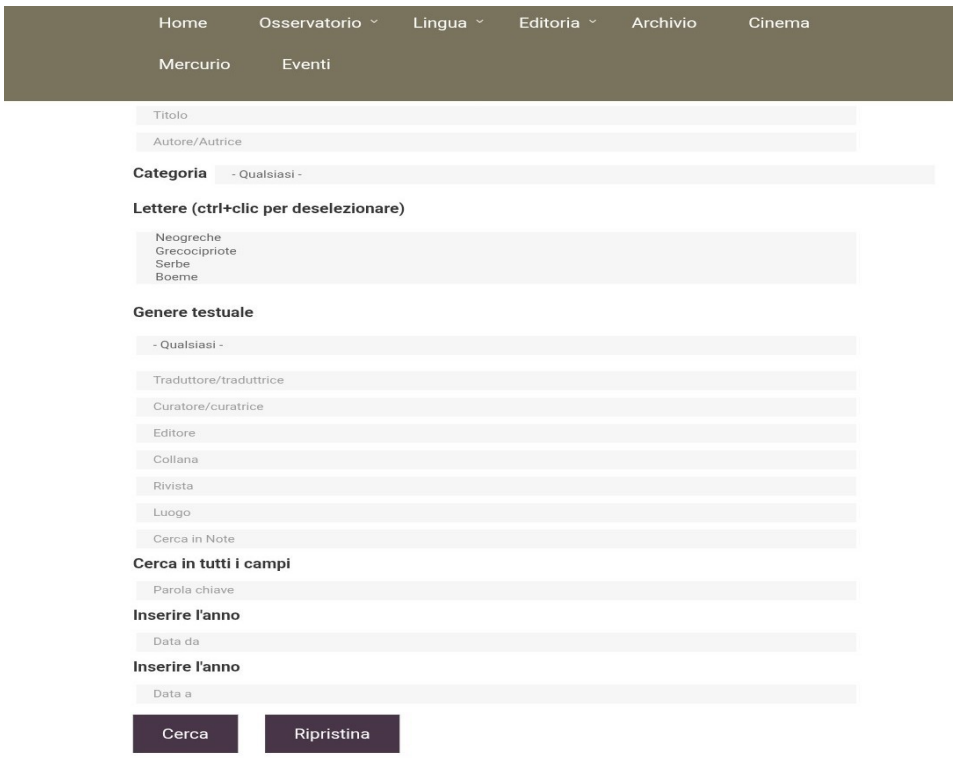


Figura 1. Maschera di ricerca della Biblioteca Digitale

Come illustrato in Figura 1, la piattaforma consente interrogazioni multi-criteriali basate su campi di metadatazione strutturati, facilitando l'accesso mirato alle risorse. Il modello di metadatazione adottato si basa su campi descrittivi standardizzati (autore, titolo, anno, tipologia documentaria, genere testuale, lingua, editore e parole chiave), in linea con i principali standard internazionali utilizzati nelle Digital Humanities, con particolare riferimento ai principi del Dublin Core. La piattaforma si configura come un database relazionale che consente l'organizzazione e l'interrogazione delle risorse attraverso metadati strutturati e una classificazione tipologica basata su un insieme definito e controllato di categorie documentarie (es. narrativa, saggio, letteratura di viaggio, recensione). L'architettura del sistema permette la connessione tra entità bibliografiche (autori, opere, traduzioni, riviste, editori), favorendo la navigazione tra record correlati e la ricostruzione di relazioni culturali e intellettuali tra le risorse. Sebbene non siano ancora implementate ontologie formali o modelli basati su Linked Open Data, l'uso di categorie coerenti e standardizzate assicura una descrizione ordinata e una piena interrogabilità del corpus.

In questo quadro teorico, Pierazzo sottolinea come nei progetti di digital scholarly editing la digitalizzazione non è un semplice trasferimento di contenuti su supporto elettronico, bensì *"una trasformazione delle pratiche editoriali e della comunicazione accademica"* (Pierazzo, 2015). In quest'ottica, sistemi come l'Osservatorio Neogreco non sono meri archivi, ma strumenti che promuovono nuove modalità di accesso, navigazione e confronto dei materiali. Analogamente, Mastandrea evidenzia che *"gli archivi elettronici e i sistemi digitali permettono opzioni multiple di ricerca intertestuale"* (Mastandrea, 2015), ponendo l'accento sul ruolo delle strutture dati nel favorire nuove forme di esplorazione bibliografica e comparazione tra testi, traduzioni e recezioni critiche.

Dal punto di vista metodologico, la Biblioteca Digitale dell'Osservatorio Neogreco si fonda su un principio chiave: il digitale non sostituisce la filologia, ma la potenzia. Ogni testo viene trattato come oggetto complesso, dotato di una storia editoriale, di varianti, di contesti di produzione e ricezione. La digitalizzazione è accompagnata da un'attenta descrizione filologica e bibliografica, che include informazioni sulle edizioni di riferimento, sulle traduzioni, sulle revisioni e sulle eventuali stratificazioni testuali. La struttura della biblioteca è pensata per essere modulare e aperta, in costante aggiornamento ed espansione. I materiali sono organizzati secondo criteri tipologici e cronologici e possono essere messi in relazione attraverso le funzionalità di ricerca della piattaforma. In questo modo, è possibile ricondurre e confrontare risorse eterogenee come traduzioni, recensioni e contributi critici che riflettono relazioni

culturali e storiche tra i testi. La biblioteca non si limita così a riprodurre un catalogo tradizionale, ma favorisce forme di esplorazione comparativa del corpus, in cui i materiali possono essere analizzati in relazione tra loro. In sintesi, l'Osservatorio Neogreco "Mario Vitti" persegue tre obiettivi principali, integrati tra loro:

Conservazione e accessibilità: raccogliere e tutelare materiali testuali fondamentali per gli studi neogreci, garantendo accesso aperto e disponibilità internazionale, superando limiti fisici e geografici.

Strutturazione e interoperabilità dei dati: creare un corpus digitale interrogabile tramite criteri tipologici, cronologici e descrittivi, con metadati strutturati che permettono analisi comparativa, statistica e intertestuale, e garantiscono collegamento con altre risorse. Dal punto di vista pratico, ogni scheda bibliografica è strutturata attraverso metadati dettagliati (autore, titolo, data, lingua, genere, edizione di riferimento) come già illustrato in Figura 1. Ad esempio, se inseriamo il nome di uno scrittore nella maschera di ricerca, emergono non solo i testi, ma anche traduzioni, saggi critici e altri materiali correlati, evidenziando la capacità del sistema di mettere in relazione contenuti eterogenei e di supportare forme di interoperabilità a livello applicativo. In questo modo, la biblioteca non restituisce semplicemente risultati isolati, ma consente una navigazione relazionale dei metadati, permettendo di esplorare dinamicamente il corpus attraverso le connessioni tra i record e di ricostruire le relazioni tra autori, opere e materiali critici. La ricchezza del progetto risiede anche nella capacità di collegare i dati dell'Osservatorio con risorse digitali esterne, consentendo una navigazione trasversale tra documenti e piattaforme diverse attraverso collegamenti ipertestuali. Un elemento centrale dell'esperienza dell'Osservatorio è stato lo sviluppo di una strategia di collegamento (link building) a risorse digitali esterne per aumentare la visibilità e la sostenibilità scientifica della piattaforma. Dal punto di vista operativo, i materiali dell'Osservatorio sono stati collegati a risorse digitali esterne, come archivi istituzionali universitari, biblioteche digitali internazionali e collezioni specialistiche, consentendo l'accesso diretto alle versioni digitalizzate dei testi, ove non sussistano restrizioni legate ai diritti d'autore. Tra le principali risorse digitali esterne collegate alle schede bibliografiche figurano archivi istituzionali, biblioteche digitali e repository accademici, tra cui OpenstarTs dell'Università di Trieste, l'archivio dell'Associazione Nazionale di Studi Neogreci, Edizioni Ca' Foscari, Internet Archive e il Centro Bibliotecario digitale dell'Università Aristotele di Salonicco.

Promozione della conoscenza e didattica: integrare ricerca e formazione, favorendo la consultazione comparata di testi, traduzioni e materiali critici, e offrendo strumenti per l'insegnamento e lo studio diretto delle fonti. La piattaforma digitale è pensata anche come strumento didattico: può essere utilizzata in contesti universitari per l'insegnamento della letteratura neogreca, della traduzione e della storia culturale, offrendo agli studenti la possibilità di lavorare direttamente sulle fonti.

2. FUNZIONI E MODALITÀ D'USO DELLA BIBLIOTECA DIGITALE

La Biblioteca Digitale dell'Osservatorio Neogreco "Mario Vitti" consente ricerche per autore, titolo, genere e anno, con funzioni avanzate per combinare criteri e accedere, quando disponibile, al testo integrale e ai materiali paratestuali. Queste funzionalità permettono a studiosi e studenti di consultare risorse bibliografiche altrimenti difficilmente accessibili, migliorando la fruizione di testi critici e traduttivi che documentano la ricezione della cultura greca moderna in Italia. La piattaforma comprende attualmente (aprile 2026) un corpus di 2.177 unità bibliografiche, suddivise in diverse tipologie documentarie: traduzioni (971), articoli in rivista (336), monografie (335), articoli in volume (118), atti di convegni (109) e recensioni (61), cui si affiancano materiali più specifici quali tesi di laurea e dottorato (circa 10), letteratura di viaggio (21), cataloghi di mostra (16), nonché ulteriori tipologie minori (libri fotografici, programmi di sala, edizioni anastatiche).
La distribuzione evidenzia la centralità delle traduzioni e della produzione critica, che costituiscono il nucleo principale del corpus, confermandone la rilevanza per lo studio della ricezione della cultura greca moderna in Italia.

Esempi di contenuti digitalizzabili:

Testi letterari neogreci (XIX–XX secolo)

Un settore centrale riguarda le traduzioni**,** fondamentali per lo studio della ricezione letteraria. La piattaforma include traduzioni di Seferis, Kavafis e Kazantzakis, insieme a studi linguistici e filologici e materiali utili per l'analisi dei testi. La Biblioteca valorizza sia il corpus critico e letterario neogreco, sia quei testi che hanno costituito pietre miliari per la formazione della disciplina, offrendo strumenti di ricerca completi e integrati. Un ulteriore ambito rilevante è costituito dalle pubblicazioni di studiosi che hanno contribuito in modo significativo allo sviluppo degli studi neogreci in Italia nel Novecento, tra cui Massimo Peri, Caterina Carpinato, Georgios Zoras, Lucia Marcheselli, Cristina Stevanoni, Costantino Nikas, Maria Perlorentzou e Gerasimos Zoras.

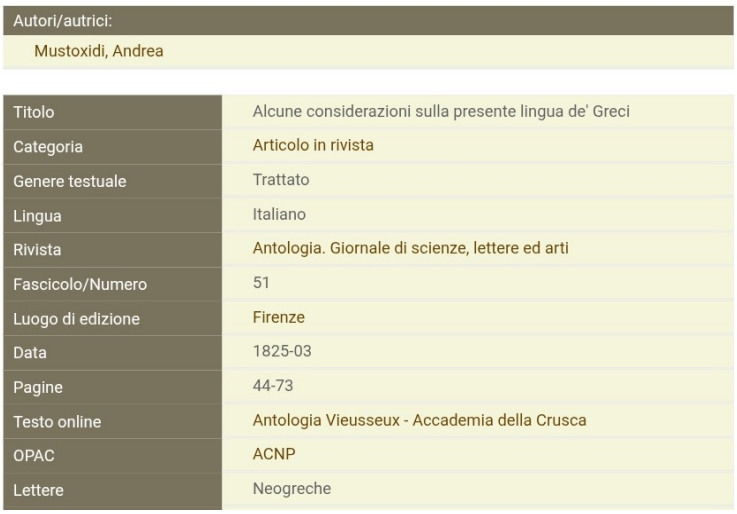
Critica letteraria e studi neogreci

La digitalizzazione di monografie e saggi fondamentali per la nascita e lo sviluppo degli studi neogreci in Italia, inclusi contributi oggi difficili da reperire, consente di ricostruire in modo dettagliato la storia intellettuale della disciplina. La piattaforma raccoglie opere di autori che hanno contribuito in maniera decisiva alla definizione del campo, come Niccolò Tommaseo, F. De Simone Brouwer, Andrea Mustoxidi, Mario Pieri e Luigi Ciampolini, rendendo così visibili genealogie critiche, dibattiti teorici e passaggi metodologici spesso accessibili solo a specialisti. All'interno di questa collezione trovano inoltre spazio testi che documentano le relazioni tra Grecia e Italia, analizzando questioni storiche, culturali e politiche, dinamiche internazionali e processi di globalizzazione e interculturalismo. Particolare attenzione è riservata a fenomeni quali migrazione transnazionale, diaspora, mobilità studentesca internazionale e studi etnici (ethnic studies). Questi materiali permettono di esplorare sistematicamente la circolazione delle idee e delle pratiche culturali tra Italia e Grecia, integrando al contempo riflessioni su tematiche contemporanee — dalla crisi economica greca alle disuguaglianze di genere, dalle dinamiche di razzismo e cittadinanza alla trasformazione degli spazi urbani — offrendo così un quadro polifonico, interdisciplinare e aperto alla comparazione tra fonti, autori e periodi storici.

Archivi, appunti e materiali didattici

Gli archivi d'autore includono corrispondenza, appunti, bozze, varianti manoscritte e traduzioni annotate, nonché documentazione accademica e istituzionale. La loro digitalizzazione permette di studiare la genesi dei testi, le reti intellettuali e la circolazione delle idee tra Grecia e Italia. La descrizione tramite metadati dettagliati e l'integrazione con risorse esterne favoriscono un approccio comparativo e transnazionale. La digitalizzazione di materiali didattici e preparatori restituisce la dimensione processuale della ricerca, trasformando la Biblioteca in uno spazio attivo di produzione di conoscenza, e non in un semplice archivio.

Riviste specialistiche
La Biblioteca ospita riviste centrali per gli studi neogreci, come *Antologia Vieusseux*, *Thesaurismata*, *Atene e Roma* e *Italoellinika*. L'integrazione digitale consente di analizzare la storia discorsiva della disciplina, le reti intellettuali e il ruolo delle riviste nella mediazione non solo culturale, ma anche nel tessuto politico e sociale. Ad esempio, l'Antologia Vieusseux tra il 1821 e il 1833 offre un ricco repertorio di articoli sul filellenismo italiano e sull'indipendenza greca dall'Impero Ottomano. Queste riviste, digitalizzate e indicizzate, illustrano come la piattaforma consenta lo studio della ricezione della cultura greca moderna nei periodici italiani dell'epoca.



Autori/autrici:	
Mustoxidi, Andrea	
Titolo	Alcune considerazioni sulla presente lingua de' Greci
Categoria	Articolo in rivista
Genere testuale	Trattato
Lingua	Italiano
Rivista	Antologia. Giornale di scienze, lettere ed arti
Fascicolo/Numero	51
Luogo di edizione	Firenze
Data	1825-03
Pagine	44-73
Testo online	Antologia Vieusseux - Accademia della Crusca
OPAC	ACNP
Lettere	Neogreche

Figura 2. Esempio di scheda bibliografica con metadati e collegamenti a risorse esterne

Come illustrato in Figura 2, alcune schede bibliografiche integrano collegamenti a risorse digitali esterne, rendendo accessibili contributi difficilmente reperibili in formato cartaceo; ad esempio, un articolo dell'*Antologia Vieusseux* è direttamente collegato alla versione digitalizzata disponibile tramite l'Accademia della Crusca, evidenziando la strategia di interoperabilità della piattaforma.

Paratesti, apparati e materiali didattici

Materiali paratestuali come introduzioni, prefazioni, apparati critici, bibliografie ragionate e percorsi tematici rafforzano la dimensione formativa e divulgativa della Biblioteca, rendendola uno strumento utile non solo per la ricerca avanzata, ma anche per la didattica universitaria e la formazione di nuovi studiosi.

3. CRITICITÀ E PROSPETTIVE FUTURE

Malgrado i progressi compiuti nella costruzione di una solida infrastruttura digitale, l'Osservatorio affronta alcune sfide tipiche dei progetti di Digital Humanities. Tra queste vi sono la gestione del copyright — in particolare opere moderne e traduzioni non sempre liberamente riusabili — la manutenzione tecnologica, la necessità di competenze interdisciplinari e la sostenibilità economica e scientifica nel lungo periodo. La sostenibilità è intesa non solo in termini finanziari, ma anche come continuità istituzionale e scientifica, garantita dall'integrazione con attività di ricerca, diffusione e promozione della piattaforma presso comunità accademiche internazionali. Anche l'interoperabilità con altre piattaforme digitali rappresenta una sfida, a causa di standard eterogenei e della complessità di connessione tra sistemi diversi.

Guardando al futuro, la Biblioteca Digitale dell'Osservatorio Neogreco "Mario Vitti" è concepita come un progetto aperto e in continua evoluzione. Tra gli sviluppi previsti vi sono l'ampliamento del corpus, il continuo aggiornamento dei contenuti e il potenziamento delle funzionalità di ricerca.

Con uno sguardo prospettico, la Biblioteca Digitale si configura come un polo di riferimento per le Digital Humanities a partire dagli studi neogreci, con l'intento di estendere progressivamente il proprio modello ad altri ambiti disciplinari affini, come la slavistica, e di integrare in futuro materiali cinematografici coerenti con la cultura neogreca e grecocipriota, ampliando così la piattaforma lungo nuove direttrici culturali e artistiche.

In conclusione, la Biblioteca Digitale rappresenta un modello in cui rigore scientifico, standard tecnici e apertura interdisciplinare convergono in un'infrastruttura al servizio della ricerca. Non si limita a catalogare fonti, ma emerge come una Biblioteca Digitale tematica, capace di coniugare conservazione, ricerca accademica e diffusione scientifica. Il progetto dimostra come le Digital Humanities siano non solo strumenti per la conservazione del passato, ma laboratori critici per il futuro della conoscenza umanistica.

BIBLIOGRAFIA

Bintoudis, C., & Zaccone, F. (2022). Osservatorio Neogreco "Mario Vitti". Sapienza Università di Roma, banca dati digitale.

Carpinato, Caterina (2024). Libri greci a Venezia (3). Mario Vitti, i suoi libri e la sua biblioteca. Edizioni Ca' Foscari (capitolo analitico sul fondo Vitti).

Carpinato, C. (2024). Mario Vitti, i suoi libri e la sua biblioteca. In E. Liosatou & F. Scalora (Eds.), Libri, storie, persone e parole fra Venezia e la Grecia. Un dialogo ininterrotto: miscellanea in memoria di Mario Vitti (pp. 15–35). Edizioni Ca' Foscari. https://doi.org/10.30687/978-88-6969-868-2/002

Mastandrea, P. (2015). Archivi Elettronici di Poesia Latina e opzioni multiple di ricerca intertestuale. Semicerchio, 53, 60–69.

Pierazzo, E. (2015). Digital Scholarly Editing: Theories, Models and Methods. Farnham, Surrey: Ashgate.

Vitti, M. (1971). Storia della letteratura neogreca. Torino: ERI.

Vitti, M. (1978). *Ἱστορία τῆς νεοελληνικῆς λογοτεχνίας*. Athina: Odysseas.

Vitti, M. (2016). Storia della letteratura neogreca. Venezia: Libreria Editrice Cafoscarina.

RISORSE DIGITALI

Osservatorio Neogreco "Mario Vitti". (s.d.). Osservatorio Neogreco "Mario Vitti". Sapienza Università di Roma.

https://lsn.seai.uniroma1.it/osservatorio/

artresearch.net: A Research Infrastructure for Exploring Interpretive Plurality in Art-Historical Photo Archives

MARILENA DAQUINO,[1] FRANCESCA MAMBELLI,[2] ALESSANDRO ADAMOU,[3] PIETRO MARIA LIUZZO,[4] STEFANIE SCHNEIDER,[5] SPYROS KOULOURIS,[6] ARTEM KOZLOV,[7] RAFAEL BRUNDO URIARTE[8]

[1] UNIVERSITY OF BOLOGNA, ITALY – MARILENA.DAQUINO2@UNIBO.IT
[2] FONDAZIONE FEDERICO ZERI - UNIVERSITY OF BOLOGNA, ITALY – FRANCESCA.MAMBELLI6@UNIBO.IT
[3] DHLAB - BIBLIOTHECA HERTZIANA – MAX PLANCK INSTITUTE FOR ART HISTORY, ITALY – ALESSANDRO.ADAMOU@BIBLHERTZ.IT
[4] FOTOTHEK - BIBLIOTHECA HERTZIANA – MAX PLANCK INSTITUTE FOR ART HISTORY, ITALY – PIETRO.LIUZZO@BIBLHERTZ.IT
[5] MARBURG UNIVERSITY, GERMANY – STEFANIE.SCHNEIDER@UNI-MARBURG.DE
[6] ATHENS SCHOOL OF FINE ARTS, GREECE – SPYROS.A.KOULOURIS@GMAIL.COM
[7] RS WORKS EE, GREECE – ARTEM@REM.SH
[8] KUNSTHISTORISCHES INSTITUT IN FLORENZ – MAX-PLANCK-INSTITUT, ITALY - RAFAEL.URIARTE@KHI.FI.IT

ABSTRACT (ENGLISH)

This paper presents artresearch.net as a digital research platform for art history that integrates heterogeneous metadata from art-historical photo archives into a shared semantic environment without resolving uncertainty or scholarly disagreement. Rather than enforcing uniform interpretations, the platform is designed to preserve ambiguous attributions, divergent institutional perspectives, and multiple interpretive positions as first-class elements of the data. By combining semantic modelling, explicit reconciliation policies, and interface design, artresearch.net supports cross-collection research while making uncertainty and interpretive plurality visible and methodologically productive.
Keywords: art history; photo archives; Semantic Web

ABSTRACT (ITALIANO)

artresearch.net: Un'infrastruttura di ricerca per l'esplorazione della diversità interpretativa negli archivi fotografici storico-artistici. Questo contributo presenta artresearch.net come una piattaforma di ricerca per la storia dell'arte che integra metadati eterogenei di archivi fotografici storico-artistici in un ambiente semantico condiviso senza eliminare l'incertezza e il disaccordo interpretativo. La piattaforma è progettata per preservare attribuzioni ambigue, prospettive istituzionali divergenti e pluralità interpretative come elementi costitutivi dei dati, anziché come anomalie da normalizzare. Attraverso la combinazione di modellazione semantica, politiche esplicite di riconciliazione e soluzioni di interfaccia, artresearch.net abilita la ricerca trasversale tra collezioni rendendo l'incertezza esplicita, navigabile e metodologicamente rilevante.
Parole chiave: storia dell'arte; archivi fotografici; Semantic Web

1. INTRODUCTION

Art-historical photo archives constitute a fundamental yet historically fragmented research resource for art history and photography studies. From the end of the 19th century and throughout the 20th century, photographic reproductions of artworks have supported attribution studies, provenance research, iconographic analysis, census and protection of cultural heritage. The collection in which they sedimented, taken as a whole, are able to recount the evolution of art history and the history of photography in support of that discipline. These archives have developed within localized institutional contexts, each of which has treated them on the basis of heterogeneous cataloging traditions, descriptive standards, and epistemic assumptions. As a consequence, large portions of this material remain siloed, difficult to access by metadata aggregators, and challenging to analyse across institutional boundaries. In recent years, Digital Humanities and GLAM communities have increasingly turned to semantic web technologies and Linked Open Data to address these issues, aiming to foster interoperability, cross-collection discovery, and reuse. However, the application of such technologies to art historical data raises fundamental methodological questions. Art history is characterised by uncertainty, which broadly includes

vague actors' identities (e.g. pseudonyms), competing attributions, evolving scholarly interpretations, and historically contingent descriptive practices. In the development of a shared research platform across multiple art-historical photo archives, attempts to enforce strict normalisation or one-to-one reconciliation risk flattening this complexity and obscuring the very interpretive richness that makes these sources valuable to researchers.
This contribution presents artresearch.net, an online research platform developed within the PHAROS association of art historical photo archives, as a case study in designing a digital research infrastructure that embraces heterogeneity. Rather than framing uncertainty management as a purely technical task, the platform treats it as a socio-technical process involving data modelling, institutional governance, Artificial Intelligence (AI) transparent empowerment, and interface design. The paper adopts a descriptive and methodological perspective, focusing on the design choices, compromises, and strategies that underpin the platform, and reflecting on their implications for digital art history and the broader Digital Humanities landscape.

2. STATE OF THE ART

In the GLAM sector, Linked Data and semantic research platforms have emerged to integrate heterogeneous collections. Platforms like ResearchSpace (Oldman and Tanase 2018) operationalize these principles, connecting datasets and enabling scholars to explore relationships across institutional data. Central to this integration is the CIDOC Conceptual Reference Model (CIDOC-CRM) (Doerr 2003), the gold standard ontology that formalizes cultural heritage concepts facilitating interoperability and complex querying, while preserving provenance and context.
Several large-scale initiatives have addressed the challenge of harmonising institutional records by constructing semantic knowledge bases to support cross-collection discovery. Among these are the Linked Open Data Infrastructure for Digital Humanities (LODI4DH) in Finland (Hyvönen 2020), the Smithsonian American Art Museum (Szekely et al. 2014) and Europeana, which aggregates heterogeneous cultural heritage metadata at a continental scale (Haslhofer and Isaac 2011). While highly effective for access and visibility, such platforms necessarily adopt strong normalisation strategies that may limit the explicit representation of institutional disagreement, uncertainty and interpretive plurality. However, historical and archival data are inherently uncertain (e.g. dates may be approximate, attributions disputed, actors only vaguely identified, and documentation incomplete) and disagreement cannot always be resolved. In linked datasets, such uncertainty can propagate across nodes, affecting analyses, visualizations, and scholarly interpretation. Addressing these challenges—through probabilistic reconciliation algorithms, confidence measures, explicit modeling of ambiguity, and shared policies for human revision—is essential (Daquino, Mambelli and Kozlov 2026). By acknowledging uncertainty, art historical platforms can balance computational rigor with the interpretive nature of historical scholarship, ensuring that digital infrastructures reflect the provisional and contested character of cultural knowledge. Against approaches that resolve heterogeneity through strong normalisation, artresearch.net treats uncertainty and institutional disagreement as integral to art-historical data. The platform adopts two complementary policies: one aimed at preserving diversity by making the provenance of information explicit and presenting multiple perspectives in parallel, and another focused on consolidation as a collaborative process, enabling providers to review and, where appropriate, refine or align overlapping records without enforcing automatic resolution.

3. THE RESEARCH INFRASTRUCTURE: ARTRESEARCH.NET

artresearch.net is a collaborative research platform developed by the PHAROS association, which brings together leading art-historical photo archives from Europe and North America (Caraffa et al. 2020). The platform currently aggregates over two million photographs depicting approximately 1.6 million artworks, drawn from institutional catalogues that span different periods, languages, and documentation traditions. The platform is built on ResearchSpace (https://researchspace.org), an open-source environment for semantic data management and exploration, and adopts an application profile of CIDOC-CRM to harmonise metadata across institutions. Data ingestion follows an Extract–Transform–Load (ETL) pipeline in which original records are transformed into Linked Open Data[1] while preserving their provenance and

[1] The endpoint is available at: https://artresearch.net/sparql.

institutional context. Rather than replacing local catalogues, artresearch.net operates as a shared semantic layer that enables cross-collection research while maintaining traceability to source data.
A central feature of the platform, in constant progress, is the reconciliation of entities—such as artists, photographers, artworks, places, and institutions—across datasets and against external authorities and datasets (e.g. Getty vocabularies, GND, Wikidata). In artresearch.net, reconciliation is not conceived as the elimination of differences to simplify data access, but as the construction of relationships that make similarities, divergences, and uncertainties explicit yet navigable. This approach reflects an understanding of research infrastructures not merely as technical systems, but as interpretive frameworks that actively shape how cultural data are explored and understood, and therefore require sustained involvement from domain experts.
With this regard, a crucial aspect of artresearch.net is the centrality of expert revision as a scholarly practice, rather than a generic human-in-the-loop mechanism. Reconciliation decisions, uncertainty thresholds, and modelling strategies are not validated *post hoc* by individual curators, but are collectively defined and periodically revised by domain experts from participating institutions, who are called to historicise their data to drive technical solutions. In this sense, expertise is not limited to correcting errors, but extends to shaping the epistemic rules under which entities are aligned, associated, or deliberately kept distinct. The platform thus embeds expert knowledge not only in data curation, but in the very policies that govern how uncertainty, disagreement, and ambiguity are represented and made actionable within the system.

4. MANAGING UNCERTAINTY AND INTERPRETIVE DIVERSITY IN ART HISTORY

Building on previous analyses of entity reconciliation in art-historical photo archives (Daquino, Mambelli and Kozlov 2026), this section briefly outlines the strategies adopted in artresearch.net to manage uncertainty across modelling, workflows, and user interfaces. While these strategies emerged from the specific context of PHAROS, they are indicative of broader challenges faced by digital humanities infrastructures dealing with heterogeneous cultural heritage data. These strategies are formalised as shared reconciliation policies, co-designed with participating institutions, in order to ensure consistency while addressing edge cases where scholarly disagreement persists or where authoritative sources diverge. Rather than treating such cases as exceptions to be resolved individually, artresearch.net formalises them as recurring scenarios for which mitigation strategies are agreed in the consortium. This includes defining when to prefer associative links over identity assertions, when to introduce umbrella entities, and when to preserve parallel interpretations without hierarchical prioritisation. In doing so, disagreement itself becomes a manageable and documented condition of the infrastructure, rather than a failure of reconciliation.

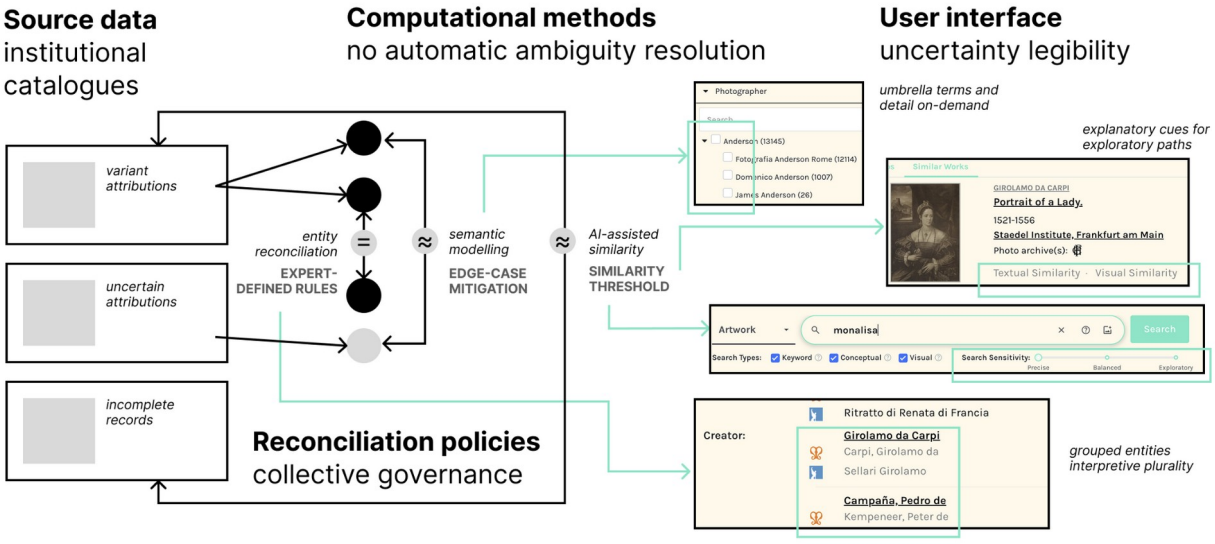


Figure 1. Managing uncertainty as a socio-technical process in artresearch.net.

Fig. 1 illustrates how uncertainty and interpretive plurality are preserved across four interconnected layers: source data, reconciliation policies co-defined by domain experts, computational processes (including semantic modelling and AI-assisted suggestions), and user interfaces designed to make degrees of certainty legible.

Expert-defined rules for reconciliation. In the ETL process, we distinguish between exact equivalences and looser associations. When the system cannot confirm that two entities are identical, because experts flagged the ambivalence of certain entities, it uses non-committal relationships (e.g. rdfs:seeAlso, skos:relatedMatch, and custom properties) instead of asserting full equivalence (owl:sameAs). This is a common situation in the reconciliation of artists to Getty ULAN. For instance, ULAN identifies Ercole Grandi with Ercole de' Roberti (ULAN 500124891),[2] but some institutions in PHAROS treat them as two distinct individuals. To avoid endorsing either position while preventing duplicates in the PHAROS authority file, we mint a URI for Ercole Grandi, linked to the ULAN candidate via rdfs:seeAlso. This choice reflects a policy-driven distinction between semantic alignment and scholarly endorsement, preserving uncertainty while maintaining meaningful connections and respecting divergent institutional traditions. Likewise, when historical sources cannot identify a person, workshop, or institution, e.g. "Rembrandt [circle of]",[3] the system creates bespoke entities for anonymous groups instead of forcing a false identification. Here too, reconciliation policies prioritise reticence over conjecture, allowing users to recognise the presence of an actor while acknowledging the epistemic limits of identification.

Edge-case mitigation with modelling solutions. To balance semantic precision with usability, artresearch.net employs *umbrella terms* to group heterogeneous or ambiguous entities (that could not be univocally identified) under a shared label, enabling users to explore related records without imposing artificial consensus. This is the case of photographers, whose legal status is not always clear from the name recorded by cataloguers (e.g. "Anderson" may designate the photographers Domenico Anderson, James Anderson or their studio),[4] or when even a rough attribution is impossible to determine (e.g. "Beyer, ?", where it is impossible to determine whether the reference is to Constantin, Günther, or Klaus Beyer). The adoption of umbrella terms follows explicit internal policies that define when semantic distinctions should be preserved at the data level but deliberately softened at the interface level, in order to facilitate retrieval and comparison without introducing misleading assertions. Such edge cases require bespoke modelling solutions that facilitate retrieval and friendly presentation in user interfaces, avoiding redundancy and excessive complexity, without neglecting correct semantic mappings. Rather than resolving ambiguity through forced normalisation, PHAROS policies favour the explicit representation of indeterminacy, even when this entails creating additional entities or higher-level groupings.

User-defined similarity threshold for exploration. In addition to these reconciliation strategies, artresearch.net designed policies for integrating AI-based components in the ETL, search and exploration interface to assist users in navigating uncertainty rather than resolving it automatically.

The ETL process employs a vector embedding-based similarity search, relying on open-weights models such as Qwen3-Embedding-8B (Yang et al. 2025), to select reconciliation candidates from external authorities like ULAN, Wikidata, GND, etc. The Large Language Model gpt-oss-120b (OpenAI 2025) is then used to re-rank and filter the candidates, selecting the best matches for expert review.

AI is also employed to retrieve results of a query based on an input text or image. The search is performed by using similarity search, exact match, and mixed methods, where the user is allowed to tune its sensitivity (i.e., whether to restrict the search parameters to an exact text search, semantic similarity search on metadata, or visual similarity in images) which sets the thresholds for the similarity algorithm. The system also highlights whether a result is returned through a confirmed equivalence, an associative, visual correspondence, or both—thereby making different levels of certainty legible to users.

When viewing individual artwork records, users can explore reconciled entries across institutions only when those reconciliations have been curated and reviewed by experts—highlighting both the reliability of confirmed alignments and the presence of scholarly disagreement where it exists. AI is used here to present visually similar artworks in cases where no explicit (curated) reconciliation has been established. To avoid epistemic overreach, the interface explicitly distinguishes AI-generated suggestions from curated reconciliations, framing them as prompts for further investigation by researchers and curators.

[2] See the example query on the SPARQL endpoint of artresearch.net: https://tinyurl.com/ulan-reconciliation.
[3] See the example query: https://tinyurl.com/rembrandt-groups.
[4] See the example query: https://tinyurl.com/anderson-umbrella.

The aforementioned metadata and image-similarity algorithms are employed to propose potential relationships between works across collections, especially where metadata alignment alone was insufficient or contradictory. These suggestions are deliberately presented as exploratory cues rather than assertions of identity: visual similarity does not imply equivalence, authorship, or scholarly agreement.
More generally, the use of AI in artresearch.net follows a policy of epistemic restraint. AI components are designed to support retrieval and sense-making, rather than to automate interpretive decisions.
The platform mediates the inevitable tension between philological precision and usability by abstracting complexity in the interface. In all aforementioned cases, users are presented with navigable summaries that can be expanded to reveal detailed provenance, institutional endorsement, variant forms of names as recorded in original catalogues, and qualifiers of uncertain attributions. In this sense, policies, modelling choices, and interface design are tightly interwoven: reconciliation decisions are made not only to improve data alignment, but to shape a research environment in which uncertainty and interpretive diversity remain visible and methodologically productive.

5. IMPLICATIONS FOR DIGITAL ART HISTORY

The design and implementation of artresearch.net suggest several implications for digital art history and, more broadly, for the development of Digital Humanities research infrastructures.
In this respect, artresearch.net positions itself alongside, but not in competition with, large-scale cultural heritage aggregators such as Europeana. While Europeana prioritises wide access, standardisation, and reuse across diverse domains, artresearch.net is conceived primarily as a research infrastructure tailored to the epistemic specificities of art-historical scholarship. Its focus is not on delivering a unified narrative or a single authoritative view, but on exposing institutional differences, scholarly disagreement, and varying degrees of certainty as first-class research objects. This complementary positioning highlights how different infrastructural models can coexist within the broader ecosystem of digital cultural heritage, addressing distinct research needs.
First, artresearch.net demonstrates how large-scale semantic aggregation can support art-historical research without imposing artificial consensus. By preserving variant attributions, competing reconciliations, and unknown entities, the platform enables scholars to work with institutional disagreement and metadata ambiguity as first-class research objects. artresearch.net exposes the plurality of institutional perspectives and scholarly traditions that underpin photographic documentation. This approach aligns with established art-historical practices, where uncertainty, attributional debate, and historiographical layering are integral to interpretation.
At the same time, the platform lowers the practical barriers to comparative research across photo archives. By reconciling artworks, artists, and photographs across institutions, artresearch.net enables queries and visual exploration that would be difficult or impossible within isolated catalogues. As a result researchers can move fluidly between close reading of individual records and broader patterns emerging across collections, being aware of the provenance and reliability of the underlying data.
Reconciliation decisions are shown to be inseparable from institutional governance, cataloguing legacies, and interface design. Choices about when to assert equivalence, when to maintain ambiguity, and how to visualise complex relationships directly shape what kinds of research questions a platform can support. From this perspective, reconciliation emerges as an ongoing infrastructural practice rather than a one-time data-cleaning task. The continuous ingestion of new datasets, the evolution of external authority files, and the emergence of new scholarly interpretations all require mechanisms for revision, review, and negotiation. artresearch.net addresses this by distributing responsibility across participating institutions and embedding human expertise into the workflow. This model suggests a viable path for other DH infrastructures that seek sustainability without enforcing centralised control or epistemic uniformity.
Finally, artresearch.net contributes to ongoing DH discussions about how uncertainty, vagueness, and interpretive plurality can be represented computationally without undermining usability. While the validation of solutions adopted is in progress (via focus groups and interviews), we can argue that the collective effort in designing artresearch.net sheds light on the fact that uncertainty does not need to be treated as a deficiency of data quality, but as an inherent property of humanities knowledge. Through strategies such as the use of non-committal semantic relations, umbrella terms, hierarchical interface facets, and explanatory summaries, artresearch.net tries to mediate complexity in ways that remain navigable while avoiding false precision.

6. CONCLUSION

artresearch.net is a collaborative research platform that operationalises a set of design principles for reconciling heterogeneous art-historical data, such as grouping ambiguous entities under umbrella terms, distinguishing between non-committal and definitive links, preserving anonymous groups, and mediating between data fidelity and user experience. Together, these principles allow the platform to reflect the interpretive complexity of cultural heritage sources while remaining usable and navigable for scholars. Importantly, these principles are conceived as scalable and extensible. As artresearch.net continues to grow through the ingestion of additional datasets, the same methodological framework will be progressively applied to new sources, ensuring continuity and coherence. At the same time, ongoing work explores the extension of these principles beyond core entities such as artworks, artists, and photographs, towards additional information, including artwork provenance and bibliography.
Rather than aiming for exhaustive integration, this incremental approach foregrounds dialogue between heterogeneous data domains, allowing new informational nuclei to be incorporated gradually and critically. In this sense, artresearch.net is conceived not as a closed system, but as an evolving research environment in which design principles, reconciliation policies, and scholarly practices can be continuously refined in response to both data growth and research needs.

ACKNOWLEDGEMENTS

PHAROS development was funded through grants from the National Endowment for the Humanities (United States), the Andrew W. Mellon and the Samuel H. Kress Foundations.

REFERENCES

Caraffa, C., Pugh, E., Stuber, T. and Wood Ruby, L. (2020). Pharos: A digital research space for photo archives. Art Libraries Journal 45, 1, 2-11. https://doi.org/10.1017/alj.2019.34

Daquino, M., Mambelli, F., & Kozlov, A. (2026). Many-to-many. Usability challenges of entity reconciliation in art history and photographic studies. Journal of Documentation. https://doi.org/10.1108/JD-09-2025-0284

Doerr, M. (2003). The CIDOC conceptual reference module: an ontological approach to semantic interoperability of metadata. AI Magazine 24, 3, 75–92. https://dl.acm.org/doi/10.5555/958671.958678

Haslhofer, B., & Isaac, A. (2011). data. europeana. eu: The europeana linked open data pilot. In Proceedings of the international conference on Dublin Core and metadata applications, 2011. https://doi.org/10.23106/dcmi.952135673

Hyvönen, E. (2020). Linked open data infrastructure for digital humanities in Finland. In Digital humanities in the Nordic countries. CEUR, 2612, 254-259. https://ceur-ws.org/Vol-2612/short10.pdf

Oldman, D., & Tanase, D. (2018). Reshaping the knowledge graph by connecting researchers, data and practices in ResearchSpace. Vrandečić, D., et al. (eds.) The Semantic Web – ISWC 2018. ISWC 2018. Lecture Notes in Computer Science, 11137. Springer, Cham. https://doi.org/10.1007/978-3-030-00668-6_20

OpenAI (2025). gpt-oss-120b & gpt-oss-20b model card. arXiv. https://doi.org/10.48550/arXiv.2508.10925

Szekely, P., Knoblock, C. A., Yang, F., Fink, E. E., Gupta, S., Allen, R., & Goodlander, G. (2014). Publishing the data of the Smithsonian American Art Museum to the linked data cloud. International Journal of Humanities and Arts Computing, 8, 152-166.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., & Qiu, Z. (2025). Qwen3 technical report. arXiv. https://doi.org/10.48550/arXiv.2505.09388

Legislative Changes Concerning the Reproduction of Images of Italian Cultural Heritage Assets between 2023 and 2025

MARCO CIURCINA[1], PIERGIOVANNA GROSSI[2]
[1] STUDIOLEGALE.IT — CIURCINA@STUDIOLEGALE.IT
[2]INDEPENDENT RESEARCHER — ME@PIERGIOVANNAGROSSI.IT

ABSTRACT (ENGLISH)

This article examines the significant legal reforms in Italy between 2023 and 2025 regarding the reproduction of images of state- and publicly owned cultural heritage assets. Historically, under Articles 107–108 of Legislative Decree No. 42/2004, such images were free to reuse for non-commercial, research, educational, and valorization purposes, with detailed regulation governed by circulars. However, the advent of digital publishing and new modes of dissemination rendered this framework outdated—particularly the lack of web guidance. Between 2022 and 2024 the Ministry of Culture enacted three key measures: the National Digitalization Plan and its accompanying Guidelines; the controversial Ministerial Decree 161/2023, which imposed onerous fees even on scholarly and public-interest uses; and the corrective Ministerial Decree 108/2024, which restored exemptions for research, education, open-access dissemination and granted discretionary fee waivers to custodial entities. Concurrently, the transposition of EU Directive 2019/790 introduced further complexity in national copyright legislation, most notably via Art. 32-quater of Law No. 633/1941. In 2025, an additional amendment further affected the framework by extending the term of protection for simple photographs to 70 years from production (Law No. 182/2025, Art. 47), with relevant implications for the reuse of reproductions. This contribution offers a critical synthesis of these legislative changes and outlines the resulting regulatory challenges and prospects.
Keywords: cultural heritage images; image reuse; Italian copyright law; open access; cultural heritage regulation

ABSTRACT (ITALIANO

Le modifiche alle normative sulle riproduzioni dei beni culturali italiani tra il 2023 e il 2025.
Questo articolo analizza le significative riforme legislative intervenute in Italia tra il 2023 e il 2025 in materia di riproduzione delle immagini dei beni culturali di proprietà dello Stato e di enti pubblici. Storicamente, ai sensi degli artt. 107–108 del Decreto Legislativo 42/2004, tali immagini erano liberamente riutilizzabili per finalità non commerciali, di ricerca, didattiche e di valorizzazione, mentre la regolamentazione di dettaglio era affidata a circolari. Tuttavia, l'avvento dell'editoria digitale e di nuove modalità di diffusione ha reso questo quadro superato, in particolare per l'assenza di indicazioni relative al web. Tra il 2022 e il 2024 il Ministero della Cultura ha adottato tre misure principali: il Piano Nazionale di Digitalizzazione e le relative Linee guida; il controverso Decreto Ministeriale 161/2023, che ha introdotto canoni onerosi anche per usi scientifici e di interesse pubblico; e il correttivo Decreto Ministeriale 108/2024, che ha ripristinato le esenzioni per ricerca, didattica e diffusione in open access e ha riconosciuto agli enti custodi la facoltà discrezionale di concedere l'esonero dal canone.
Parallelamente, il recepimento della Direttiva (UE) 2019/790 ha introdotto ulteriori complessità nella normativa nazionale sul diritto d'autore, in particolare attraverso l'art. 32-quater della Legge 633/1941. Nel 2025, un ulteriore intervento ha inciso sul quadro estendendo a 70 anni dalla produzione la durata di tutela delle fotografie semplici (Legge 182/2025, art. 47), con implicazioni rilevanti sul riuso delle riproduzioni. Il presente contributo propone una sintesi critica di tali cambiamenti legislativi e delinea le conseguenti sfide regolatorie e le prospettive che ne derivano.
Parole chiave: immagini dei beni culturali; riuso delle immagini; diritto d'autore; open access; normativa sui beni culturali

1. INTRODUCTION

Images of Cultural Heritage Assets owned by the State and by public territorial entities are regulated in Italy by the Code of Cultural Heritage and Landscape (Legislative Decree, D.Lgs. n. 42/2004), in Articles

107 and 108. Article 108, updated between 2014 and 2017, provides that images may be freely reused and disseminated for non-profit purposes, for valorization, and for research and teaching purposes. The provision sets out, in general terms, the cases in which the use of images is free of charge and those in which it is instead subject to the payment of a fee. Prior to 2023, detailed regulation was entrusted to a number of circulars and to the specific institution responsible for the custody of the assets (including: Circolare 33/2017, Circolare 39/2017, and Circolare 50/1995).
The need to regulate this practice at the national level has long been recognized. Although functional and consolidated over the years, this regulatory framework had become increasingly unable to keep pace with the rapid development of the digital world: the limits applicable to printed works (a print run of 2,000 copies and a cover price of 77.46, reduced to 70 euros for archival documents) and the lack of guidelines for web publications were by then outdated and made it uncertain where the boundary lay between free publication and publication subject to the payment of a fee, as in the case of many on-demand editions.
Against this background, between 2022 and 2024 the Ministry of Culture intervened through three main actions: the drafting of the National Digitalization Plan and its related Guidelines (PND 2022; PND, Linee Guida 2022), the issuance of Ministerial Decree No. 161 of 11 April 2023, which introduced a controversial tariff system extended also to scholarly and popular publications (D.M. 161/2023), and, finally, Ministerial Decree 108/2024, which corrected the previous measure by reintroducing exemptions for non-commercial purposes and expanding the decision-making autonomy of institutions (D.M. 108/2024).
Alongside these measures, further legislative updates in other areas have also been introduced, including the transposition of Directive (EU) 2019/790 on copyright, which resulted in the introduction of the new Article 32-quater of Law No. 633/1941, and the recent amendment to Article 92 of the same Law No. 633/1941, introduced by Article 47 of Law No. 182/2025, which extended the term of protection for simple photographs to 70 years.
Without claiming to be exhaustive, this short contribution offers a synthesis of the main updates that have affected the regulatory framework for the reproduction of cultural heritage assets, and an overview of current critical issues and prospects.1
Note: The Code of Cultural Heritage and Landscape (hereafter D.Lgs. n. 42/2004) and the related legislation apply exclusively to assets entrusted to the Ministry of Culture, to the Regions, and to public territorial entities. Privately owned cultural heritage assets – for instance those belonging to Dioceses, Foundations, or private individuals – do not fall within the scope of application of the Code, and their reproduction is regulated by private agreements between the holder of the asset and the requesting party.

2. TOWARDS MINISTERIAL DECREE 108/2024

The Digitalization Plan
In June 2022, the Central Institute for the Digitalization of Cultural Heritage published the National Digitalization Plan and the related Guidelines for the acquisition, circulation and reuse of reproductions of cultural heritage assets in the digital environment (PND, Linee Guida 2022). Produced through a wide-ranging consultation and intended to precede ministerial regulations on the reproduction of cultural heritage assets, these documents marked an important step forward in clarifying the distinction between the phase of image acquisition, which may entail reimbursement of costs, and their use, which is free of charge only when non-commercial.
The Guidelines, which are still in force, particularly encourage the dissemination of high-resolution images in order to support research and public dissemination. In short, although lacking normative force, they constitute an operational tool for public institutions and a concrete support for the free circulation of cultural images.
Ministerial Decree 161/2023
Ministerial Decree No. 161 of 11 April 2023 (D.M. 161/2023), however, introduced new tariffs for the use of images of the State's cultural heritage, including for academic purposes, breaking with the previous practice of free use or minimal costs. This prompted strong criticism from the academic and cultural community, due to its negative effects on research, dissemination, and public access to heritage. In addition to the costs, administrative burdens and certain regulatory inconsistencies were also contested. From the very moment of its issuance, the need to revise the decree became apparent (for an overview of the critical issues that emerged and of the protest actions undertaken, among others, see: AIB, 2023;

Volpe, 2023; Modolo, 2023; Manacorda, 2023; for two concrete examples of the burdens imposed on scholars and researchers, see: Corriere della Sera, 2023; Brugnoli, 2023).

3. MINISTERIAL DECREE 108/2024: CORRECTIONS AND OPENINGS

Prompted by numerous protests and in the face of evident regulatory shortcomings, the Ministry issued Ministerial Decree No. 108 of 21 March 2024 (D.M. 108/2024), introducing important corrective measures to the 2023 decree. The main innovations include:

– Exemption from payment of the fee for all scholarly and popular publications, regardless of format and print run.
– Explicit free-of-charge use for open access publications.
– Exemption for art catalogues with print runs of up to 4,000 copies.
– Exemption for journalistic publications within the scope of the right and duty to report.
– The possibility for public institutions to waive the fee autonomously, including on the basis of territorial specificities and heritage promotion strategies.

This represents a relevant change, strengthening a more open approach to access to cultural heritage for purposes of research, information, and promotion. However, it does not align with the indications that emerge from the documents produced by the Central Institute for the Digitalization of Cultural Heritage in June 2022: quantitative thresholds have been replaced by qualitative criteria, and free use remains subject to the assessment of the competent institution.
Note: D.M. 108/2024 applies exclusively to assets entrusted to the Ministry of Culture. Public territorial entities – Regions, Municipalities, and other local bodies – are not required to adopt it, with the risk of national unevenness in access to images of public heritage.

3.1 Critical Aspects of the New Decree

- Institutional Discretion

The decision on whether free use applies remains entrusted to the granting authority that has the asset in custody. In the absence of binding guidelines, this may lead to inconsistent applications: publications with identical purposes could receive different treatment. Such discretion, while introducing greater flexibility than the previous decree, has been regarded as problematic by some scholars, given the risks it poses to academic freedom and to equity of access.

- The Return to the Formal Request

The obligation to submit a formal request has been reintroduced even for publications with modest print runs and cover prices. This represents an additional bureaucratic burden for authors and publishers, and an increased workload for institutions, which are required to manage complex procedures even in the absence of economic returns. In some cases, the assessment of the scientific or popular nature of the publication, which is required in order to obtain the exemption, may introduce further uncertainty and delays.

3.2 Positive Developments: A Step towards Openness

- Expansion of Free-of-Charge Uses

For the first time, the decree explicitly lists the types of publications that may benefit from free use, including:
– Scholarly journals and Class A journals (according to ANVUR - National Agency for the Evaluation of Universities and Research Institutes).
– Exhibition catalogues with print runs of up to 4,000 copies.
– Journalistic publications for reporting purposes.
– Exhibition, educational, and outreach materials produced by public or private entities without profit motives.
This marks an important opening compared to the past, when free access for these publishing categories was subject to quantitative thresholds or to specific negotiations.

- Recognition of Open Access

It is finally clarified that publications that are freely accessible to anyone – in digital format and at no cost – fall among the works exempt from the fee. This reinforces alignment with European principles of open access and the free circulation of knowledge.

- Waiving the Fee

A noteworthy innovation is the power granted to the director of the institute holding the assets to reduce or waive the fee, depending on territorial specificities and opportunities to promote lesser-known cultural heritage. If applied coherently, this instrument may encourage the promotion of lesser-known contexts and support projects that contribute to the free dissemination of images and to the shared knowledge of heritage.

- Other Elements of Simplification and Remaining Critical Issues

The tariff structure has been reorganized to distinguish more clearly between commercial and non-commercial uses. However, the decree also introduced an increase in the minimum costs for digital reproductions (starting from 7 euros), excluding only lower-quality images (72 dpi). Moreover, a certain degree of complexity remains in authorization procedures, which could represent an obstacle for institutions and actors with limited resources (for an overview of the issue, see among others: Calculli, 2024; Liverani, 2024; Modolo, 2024; Volpe, 2023; Grossi, 2024) (Fig. 1).

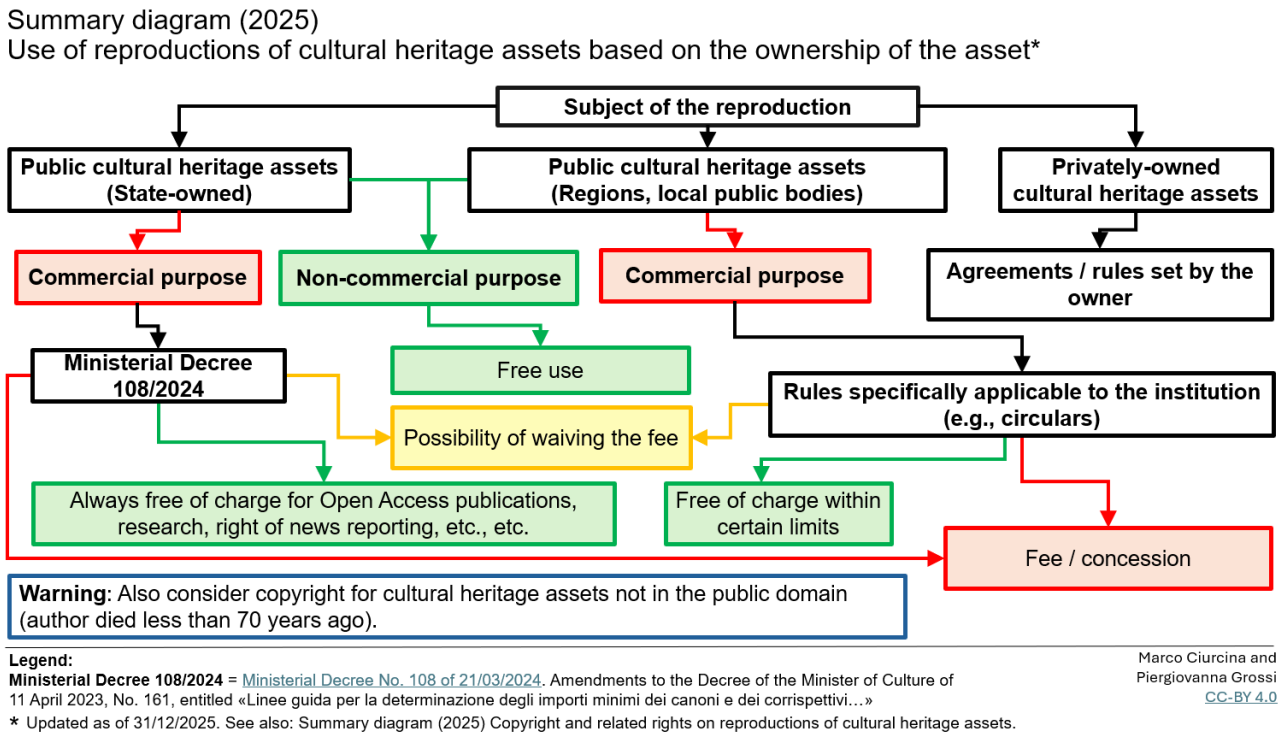


Figure 1. Use of reproductions of cultural heritage assets according to the ownership of the asset (2025).

4. EUROPEAN TRANSPOSITION AND REGULATORY CONTRADICTIONS. DIRECTIVE (EU) 2019/790 AND LAW NO. 633/1941, ARTICLE 32-QUATER

The decree is situated within a broader context, in which openness of data, digitalization, and the democratization of access to cultural heritage are becoming central. With Directive (EU) 2019/790 (art. 14), the European Union established that reproductions of works in the public domain must not be subject to copyright or related rights, except in the case of original re-elaborations. Italy transposed the provision in 2021 by introducing Article 32-quater into the copyright law (Law No. 633/1941), which nevertheless specifies that the provision does not concern the rules laid down in Articles 107 and 108 of the D.Lgs. n. 42/2004. From a theoretical and practical standpoint, the coexistence of copyright law and cultural heritage law generates a structural tension, particularly evident in digital contexts. While Article 14 of Directive (EU) 2019/790 promotes the free reuse of reproductions of public domain works, the Italian system—through Articles 107–108 of D.Lgs. n. 42/2004 and D.M. 161/2023—retains administrative control over images of publicly owned cultural heritage, potentially imposing conditions or fees even in the absence of copyright (Manacorda, 2021; Modolo, 2021; Volpe, 2021). This dual regime creates key inconsistencies: public domain images may still require authorization, at least in Italy. A paradigmatic example is the litigation concerning Leonardo's Vitruvian Man by Ravensburger: the

Tribunal of Venice prohibited commercial use without authorization, while German courts later allowed its commercialization, emphasizing the public domain status and territorial limits of Italian law. Further legal frictions have emerged in other cases that reveal a fragmented jurisprudential landscape, where the reuse of public domain images depends on jurisdiction and the balance between cultural heritage protection and copyright. For practitioners, this entails a combined assessment of intellectual property and administrative constraints (for a synthesis of the case law, see among others: De Angelis & Giardini, 2023; Dore & Priora, 2024).

4.1 Law No. 633/1941, Articles 32-quater and 92

Despite the new exception introduced by Article 32-quater, assessing the interaction between copyright, related rights, and the reproduction of cultural heritage assets remains a complex matter.
In particular, the related right in "simple photographs" under Article 87 of the Italian Copyright Law (Law No. 633/1941) covers "reproductions of works of figurative art and frames from cinematographic films." Article 92 of Law No. 633/1941 governs the duration of that right: originally set at 20 years from production, it was extended in 2025 to 70 years from production as a result of the amendment introduced by Law No. 182/2025, Article 47 (Fig. 2). It should be recalled, however, that Article 32-quater provides that the reproduction of a "work of the visual arts" in the public domain is not subject to copyright or related rights, unless the reproduction itself qualifies as an original work. Nonetheless, it cannot be overlooked that the two provisions employ different terminology: Article 32-quater refers to a "work of the visual arts," whereas Article 87 speaks of "works of figurative art". Although it may appear a bold argument, it remains possible that this difference could be invoked to contend that, for works of figurative art that are not works of the visual arts (assuming such a category exists), Article 32-quater would not apply.

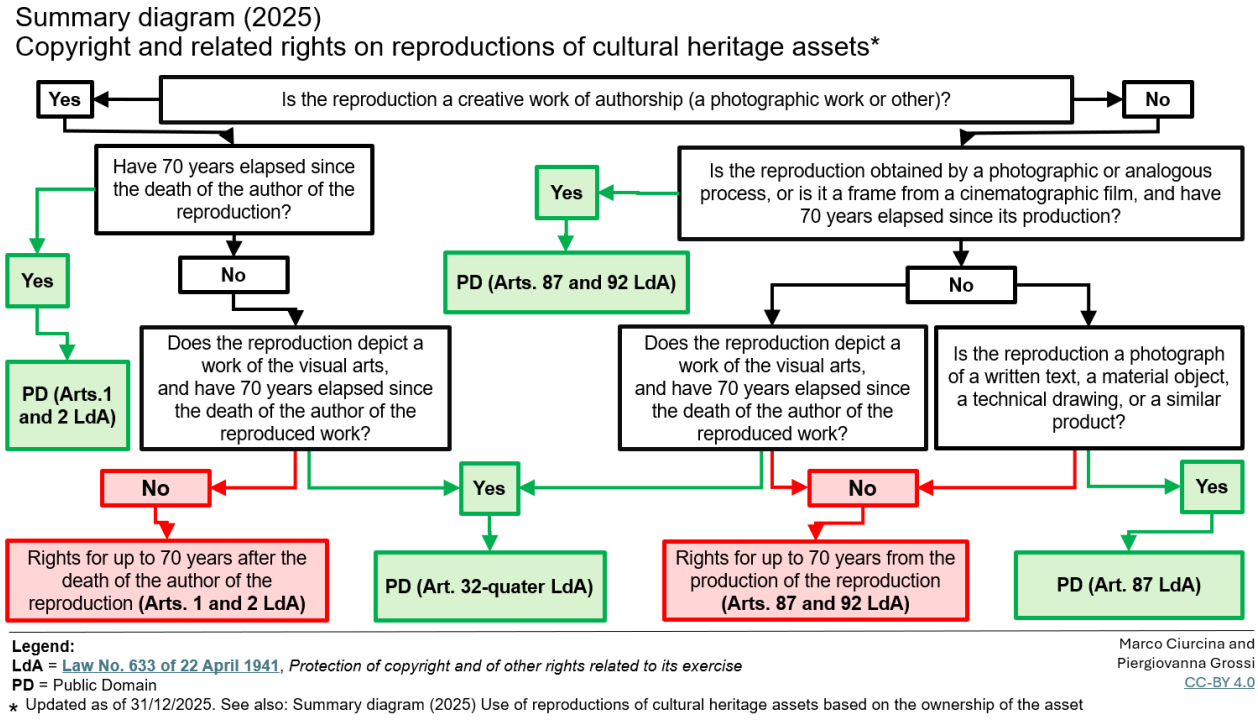


Figure 2. Copyright and related rights on the reproduction of cultural heritage assets (2025).

The 2025 reform extending the protection of "simple photographs" from 20 to 70 years from production significantly amplifies these tensions. The amendment to Article 92 of Law No. 633/1941 drastically reduces the number of images entering the public domain, impacting both commercial exploitation and scholarly reuse.
From an operational perspective, the extension introduces new complexity. Many reproductions of cultural heritage assets fall within the category of simple photographs, so materials previously reusable may now remain under exclusive rights for decades.

5. CONCLUSIONS: WHAT FUTURE FOR ACCESS TO AND REUSE OF IMAGES?

D.M. 108/2024, although not without limitations, marks a significant milestone. The exemptions introduced, the possibility of waiving the fee, and the recognition of open access represent important openings, moving in the direction of promoting knowledge and the valorization of heritage.
By extending the categories exempt from reproduction costs to scholarly, academic, educational, and outreach publications, as well as journals recognized by ANVUR, and by removing limits on print run and cover price, the decree constitutes a notable step towards free and open access.
However, despite the extension of exemptions, the decree introduces major complexity for publications that do not fall within the specified categories, leaving them subject to potentially prohibitive tariffs. Moreover, the need for formal recognition of the scholarly or educational nature of the publication by the granting authority remains, with resulting delays and bureaucratic uncertainties that further complicate access to the images needed for research and dissemination.
The decree also established an increase in costs for reproductions that do not fall within the exemptions, with a particular impact on digital images, for which the minimum cost was raised to 7 euros (5 euros for low resolution). This increase represents a significant obstacle for low-budget initiatives and for non-commercial uses, such as teaching and scholarly use.
The granting of broad discretion to institutions in authorizing the use of images raises questions about the uniformity and transparency of the decree's application. Such factors risk undermining, at least in part, the openness pursued by the measure, which moreover applies only to the State and not to territorial entities.

For the decree to represent a true paradigm shift, it will be necessary to accompany it with:
– Clear and shared guidelines for managing entities.
– Practices of transparency and uniformity in the management of authorizations.
– Cultural awareness-raising regarding the importance of the free circulation of images for research, education, and public participation.
The hope is that institutions will be able to seize this opportunity to develop a common-sense procedure, grounded in openness and efficiency (for example, by adopting the practice of waiving the fee). Only in this way will it be possible to overcome the economic and procedural obstacles that still exist, and to make the Italian cultural heritage truly accessible, in line with the principles of promoting culture and scientific knowledge.
From the perspective of those working concretely with images — researchers, curators, digital humanists, and publishers — the current framework can be described as a system characterized by partial openness combined with persistent procedural and interpretative barriers.
The recognition of open access and the extension of exemptions represent notable progress. However, their practical effectiveness depends largely on how granting authorities implement these principles. In the absence of standardized procedures, the risk is that regulatory benefits remain unevenly distributed in practice.
In particular, navigating different legal regimes, institutional policies, and administrative procedures requires new competencies beyond traditional scholarly skills. This includes the ability to assess legal risk, manage permissions, and design dissemination strategies compliant with both copyright and cultural heritage regulations.

REFERENCES

Associazione Italiana Biblioteche. (2023, November 28). *Lettera MAB a proposito del DM 11 aprile 2023, n. 161* [MAB letter regarding Ministerial Decree 11 April 2023, No. 161]. https://www.aib.it/documenti/lettera-mab-dm-161-2023

Brugnoli, A. (2023). Il d.m. 11 aprile 2023, n. 161 e il suo impatto sulla ricerca e sull'editoria: Brevi note a margine di un caso studio di pubblicazione in Open Access [Ministerial Decree 11 April 2023, No. 161 and its impact on research and publishing: Brief notes on an open-access case study]. *Aedon, 2*, 208–213. https://doi.org/10.7390/108957

Calculli, G. (2024). Il d.m. 21 marzo 2024, n. 108 del Ministero della Cultura: Un passo avanti, un passo indietro [Ministerial Decree 21 March 2024, No. 108 of the Ministry of Culture: One step forward, one step back]. *Aedon, 2*, 110–125. https://doi.org/10.7390/114917

Circolare 33/2017. Ministero dei Beni e delle Attività Culturali e del Turismo. (2017, September 7). *Circolare n. 33 (prot. 14106): Riproduzione di documenti archivistici effettuata da privati con mezzi propri* [Circular No. 33 (Prot. 14106): Reproduction of archival documents by private individuals using their own means]. http://2.42.228.207/archivi/index.php/strumenti/normativa-e-standard/item/34-normativa

Circolare 39/2017. Ministero dei Beni e delle Attività Culturali e del Turismo. (2017, September 29). *Circolare n. 39 (prot. 15630): Addendum alla circolare n. 33/2017* [Circular No. 39 (Prot. 15630): Addendum to Circular No. 33/2017]. http://2.42.228.207/archivi/index.php/strumenti/normativa-e-standard/item/34-normativa

Circolare 50/1995, Ministero per i Beni Culturali e Ambientali. (1995, June 7). *Circolare n. 50/1995: Servizi aggiuntivi e concessione d'uso di beni culturali* [Circular No. 50/1995: Additional services and concessions for the use of cultural heritage assets].

Corriere della Sera. (2023, October 30). *Le foto dei nostri beni culturali e le nuove gabelle per gli studiosi* [Photographs of our cultural heritage and the new levies for scholars]. https://www.corriere.it/cronache/23_ottobre_30/foto-nostri-beni-culturali-nuove-gabelle-gli-studiosi-ac0ca9ba-7753-11ee-9b27-58862827e9ef.shtml

De Angelis, D., & Giardini, G. (2023, July 10). *Tales of public domain protection in Italy*. Communia Association / Kluwer. https://communia-association.org/2023/07/10/tales-of-public-domain-protection-in-italy/

Directive (EU) 2019/790. *Direttiva (UE) 2019/790 del Parlamento europeo e del Consiglio del 17 aprile 2019 sul diritto d'autore e sui diritti connessi nel mercato unico digitale e che modifica le direttive 96/9/CE e 2001/29/CE* [Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market]. (2019). EUR-Lex. https://eur-lex.europa.eu/legal-content/IT/TXT/?uri=CELEX%3A32019L0790

D.Lgs. n. 42/2004. *Decreto legislativo 22 gennaio 2004, n. 42, Codice dei beni culturali e del paesaggio, ai sensi dell'articolo 10 della legge 6 luglio 2002, n. 137* [Legislative Decree No. 42 of 22 January 2004, Code of Cultural Heritage and Landscape]. (2004). Normattiva. https://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:decreto.legislativo:2004-01-22;42!vig=

D.M. 161/2023. Ministero della Cultura. (2023, April 11). *Linee guida per la determinazione degli importi minimi dei canoni e dei corrispettivi per la concessione d'uso dei beni in consegna agli istituti e luoghi della cultura statali (D.M. 11/04/2023, n. 161)* [Guidelines for determining minimum fees and charges for the granting of use of assets entrusted to state cultural institutes and venues (Ministerial Decree 11/04/2023, No. 161)]. https://www.beniculturali.it/comunicato/dm-161-11042023

D.M. 108/2024. Ministero della Cultura. (2024, March 21). *Modifiche al D.M. 11 aprile 2023, n. 161, recante "Linee guida per la determinazione degli importi minimi dei canoni e dei corrispettivi per la concessione d'uso dei beni in consegna agli istituti e luoghi della cultura statali" (D.M. 21/03/2024, n. 108)* [Amendments to Ministerial Decree 11 April 2023, No. 161 … (Ministerial Decree 21/03/2024, No. 108)]. https://www.beniculturali.it/comunicato/26075

Dore, G., & Priora, G. (2024, April 29). *The EU imperative to a free public domain: The case of Italian cultural heritage*. Communia Association / Kluwer. https://communia-association.org/2024/04/29/the-eu-imperative-to-a-free-public-domain-the-case-of-italian-cultural-heritage/

Grossi, P. (2024, April 18). *Nuove regole sulle riproduzioni dei beni culturali: Saremo più libere e liberi di usarle?* [New rules on reproductions of cultural heritage: Will we be freer to use them?]. Wikimedia Italia. https://www.wikimedia.it/news/nuove-regole-sulle-riproduzioni-dei-beni-culturali-saremo-piu-libere-e-liberi-di-usarle/

Law No. 633/1941. *Legge 22 aprile 1941, n. 633, Protezione del diritto d'autore e di altri diritti connessi al suo esercizio* [Law No. 633 of 22 April 1941, Protection of copyright and related rights]. (1941). Normattiva. https://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:legge:1941;633!vig=

Law No. 182/2025. *Legge 2 dicembre 2025, n. 182. Disposizioni per la semplificazione e la digitalizzazione dei procedimenti in materia di attività economiche e di servizi a favore dei cittadini e delle imprese.* (G.U. Serie Generale n. 291 del 14-12-2025). Normattiva. https://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:legge:2025-12-02;182

Liverani, P. (2024). Riproduzioni dei beni culturali statali: Il nuovo Decreto Ministeriale 108/2024 [Reproductions of state-owned cultural heritage: The new Ministerial Decree 108/2024]. *JLIS.it, 15*(2), 1–7. https://doi.org/10.36253/jlis.it-614

Manacorda, D. (2021). L'immagine del bene culturale pubblico tra lucro e decoro: Una questione di libertà [The image of publicly owned cultural heritage between profit and decorum: A question of freedom]. *Aedon, 1*, 24–29. https://doi.org/10.7390/100564

Manacorda, D. (2023). Un decreto inopportuno: Appunti di un archeologo [An ill-timed decree: Notes by an archaeologist]. *Aedon, 2*, 225–232. https://doi.org/10.7390/108960

Modolo, M. (2021). La riproduzione del bene culturale pubblico tra norme di tutela, diritto d'autore e diritto al patrimonio [Reproduction of publicly owned cultural heritage between protection rules, copyright, and the right to heritage]. *Aedon, 1*, 30–36. https://doi.org/10.7390/100565

Modolo, M. (2023). Il d.m. 11 aprile 2023, n. 161: Osservazioni e proposte [Ministerial Decree 11 April 2023, No. 161: Observations and proposals]. *Aedon, 2*. https://aedon.mulino.it/archivio/2023/2/modolo.htm

PND 2022. Istituto Centrale per la Digitalizzazione del Patrimonio Culturale. (2022, June). *Piano Nazionale di Digitalizzazione del Patrimonio Culturale (PND)* [National Digitalization Plan for Cultural Heritage] (Version 1.0). Docs Italia. https://docs.italia.it/italia/icdp/icdp-pnd-docs/it/v1.0-giugno-2022/index.html

PND, Linee Guida 2022. Istituto Centrale per la Digitalizzazione del Patrimonio Culturale. (2022, June). *Linee guida per l'acquisizione, la circolazione e il riuso delle riproduzioni dei beni culturali in ambiente digitale* [Guidelines for the acquisition, circulation, and reuse of digital reproductions of cultural heritage assets] (Version 1.0). Docs Italia. https://docs.italia.it/italia/icdp/icdp-pnd-docs/it/v1.0-giugno-2022/linee-guida/acquisizione-circolazione-riuso.html

Volpe, G. (2021, March 15). *Vietare l'uso delle foto dei beni culturali è un danno economico* [Banning the use of cultural heritage photos is an economic loss]. *L'Huffington Post (Italian ed.)*. https://www.huffingtonpost.it/entry/vietare-foto-beni-culturali-danno_it_604ebc9ec5b6cf72d094bea4

Volpe, G. (2023, October 22). *Le immagini del patrimonio culturale: Un'eredità condivisa ora a rischio* [Cultural heritage images: A shared legacy now at risk]. Archaeoreporter. https://www.archaeoreporter.com/it/2023/10/22/le-immagini-del-patrimonio-culturale-uneredita-condivisa-ora-a-rischio/

[3/C] Analisi computazionale del testo letterario

Flaiano postcoloniale? Un approccio computazionale alla *vexata quaestio* di *Tempo di uccidere*

SILVIA LILLI[1], DANIEL RAFFINI[2]

[1]UNIVERSITÀ DI ROMA TOR VERGATA, ITALIA - SILVIA.LILLI@UNIROMA2.IT
[2] SAPIENZA UNIVERSITÀ DI ROMA, ITALIA - DANIEL.RAFFINI@UNIROMA1.IT

ABSTRACT (ITALIANO)

Il romanzo *Tempo di uccidere* di Ennio Flaiano è al centro di un dibattito sulla posizione dell'autore rispetto all'impresa coloniale italiana in Etiopia, a causa di ambiguità legate alla costruzione narrativa, come la focalizzazione interna, l'inaffidabilità del narratore e il rovesciamento finale. Questo studio affronta la questione tramite un'analisi computazionale delle rappresentazioni di soldati e indigeni filtrate dal punto di vista del protagonista. L'ipotesi è che emergano tratti stilistici e tematici distintivi tra colonizzatori e colonizzati, utili a interpretare il romanzo in chiave coloniale, anticoloniale o postcoloniale. Il testo è stato annotato manualmente secondo diverse focalizzazioni (Io-narrante, Colonizzatore, Indigeno) e oggetti descritti (Paesaggio, Colonizzatore, Indigeno). Sono state poi applicate analisi computazionali esplorative (MFW, keyness, stilometria, parti del discorso) e tecniche di in-context learning con un Large Language Model. I risultati evidenziano una narrazione ambigua, che decostruisce i cliché linguistici del discorso coloniale, ma non giunge alla rappresentazione postcoloniale dell'alterità.
Parole chiave: Ennio Flaiano; letteratura postcoloniale; narratologia; focalizzazione; in-context learning

ABSTRACT (ENGLISH)

Postcolonial Flaiano? A Computational Approach to the Vexata Quaestio of "Tempo di uccidere". The novel *Tempo di uccidere* by Ennio Flaiano is at the center of a debate concerning the author's stance on the Italian campaign in Ethiopia, due to ambiguities related to its narrative construction, such as internal focalization, the unreliable narrator, and the reversal in the ending. This study addresses the issue through a computational analysis of the representations of soldiers and indigenous people, filtered through the protagonist's perspective. The hypothesis is that distinct stylistic and thematic features emerge between colonizers and colonized, contributing to an interpretation of the novel in colonial, anti-colonial, or postcolonial terms. The text was manually annotated according to different types of focalizations (narrating-I, colonizer, indigenous) and described objects (landscape, colonizer, indigenous). Exploratory computational analyses were then applied (MFW, keyness, stylometry, parts of speech), along with in-context learning techniques using a LLM. The results highlight an ambiguous narrative that deconstructs the clichés of colonial discourse, yet without fully achieving a postcolonial representation of otherness.
Keywords: Ennio Flaiano; postcolonial literature; narratology; focalization, in-context learning

1. INTRODUZIONE E STATUS QUESTIONIS

La vittoria del romanzo *Tempo di uccidere* di Ennio Flaiano nella prima edizione del Premio Strega del 1947 stupisce nel clima culturale dell'epoca: unico romanzo di un autore noto come critico e sceneggiatore, ambientato durante la Campagna d'Etiopia, ha al centro la vicenda di un debole, inetto ufficiale, moralmente ambiguo e oscillante tra l'ossequio e ironica dissacrazione della propaganda fascista. Mentre gli autori italiani testimoniavano la memoria dell'occupazione e della Resistenza, il successo del romanzo di Flaiano sembra contrastare il desiderio di chiusura con l'esperienza coloniale e con il suo relegamento a responsabilità del Fascismo (Del Boca, 1992: 119); tanto più che la presa di posizione antifascista nel romanzo rimane suggerita, ma non esplicitata. L'eccezionalità del romanzo si ridimensiona se consideriamo che l'ambientazione coloniale e lo svolgimento della trama – la "malattia" del protagonista a contatto con l'indigena e la crisi identitaria che ne consegue – sono elementi familiari ai lettori dell'epoca, abituati alla letteratura coloniale ed esotica (di cui *Heart of Darkness* non rappresenta che la punta di diamante) già prima che il Fascismo ne facesse un proprio genere di propaganda (Bongie, 1991; Casales, 2023). È probabile, in altre parole, che all'epoca della sua pubblicazione le convenzioni del

genere avessero la meglio sulla lettura politica, considerando che una riflessione matura sull'esperienza coloniale era ancora di là da venire (Del Boca 1998: 590).
Sull'onda del pensiero di Fanon e Spivak, solo dagli anni '90, con Orlandini (1992) e Domenichelli (1997), preceduti pionieristicamente da Pautasso (1992, 1ª ed. 1976), viene riconosciuto il significato storico dell'ambientazione africana, e l'opera viene interpretata come una chiara testimonianza della posizione anticolonialista e antifascista di Flaiano. Seguendo questa strada, nel corso degli anni 2000, sono fioriti numerosi studi che hanno collegato *Tempo di uccidere* alla critica postcoloniale, interpretando il romanzo come la prima opera postcoloniale nella letteratura italiana (Baraldi, 2004; Bazzocchi, 2012; Benvenuti, 2012; Brunetti, 2010; Maxia, 2010; Mengozzi, 2016; Nasson, 2012; Palumbo, 2002; Prisco, 2012; Re, 2017; Skocki, 2012), interpretazione avvalorata dalla lettura del diario di guerra scritto dall'autore durante l'esperienza della campagna d'Etiopia (Flaiano, 2020).
Nonostante questa diffusa lettura, nel romanzo è assente una presa di posizione esplicita, e lo statuto morale del protagonista risulta tutt'altro che definito. A questo effetto concorrono vari elementi narrativi:

- La narrazione è condotta da un narratore auto e intradiegetico, alternato raramente da un narratore in prima persona extradiegetico (posizionato a una certa distanza di tempo). La focalizzazione è interamente concentrata sull'io-agente, anche dove la presenza del narratore extradiegetico supporrebbe una conoscenza e un giudizio sui fatti. L'affidabilità del narratore (Booth, 1961; Sternberg and Yacobi, 2015) è quindi minata, anche in virtù dell'atteggiamento tipico dell'inetto, con l'accordo imperfetto e contraddittorio tra pensieri, intenzioni e azioni.
- L'io-agente mostra un atteggiamento ambiguo verso il colonialismo: da una parte ripete gli stereotipi della propaganda coloniale, impostati sulla superiorità bianca e maschile (Stefani, 2007), dall'altra prova disgusto nei confronti dell'apparato militare, nella sua mediocrità volgare e opportunistica.
- Lo svolgimento della trama, che fa suppone una direzione escatologica – dal crimine al riconoscimento della colpa, alla confessione e alla guarigione grazie al padre della vittima – viene ribaltato quasi completamente nel capitolo finale, in cui la colpa individuale viene minimizzata da un sottotenente (interpretabile come alter ego dell'autore), che attribuisce al destino e all'imprevedibile sequenza degli eventi la responsabilità dei crimini commessi, per i quali, in effetti, il tenente non verrà mai accusato.

Il testo lascia perciò aperte diverse domande: dietro la scelta di Flaiano di ambientare il romanzo in Etiopia vi è l'intenzione consapevole di porsi criticamente verso il colonialismo, o vi è semplicemente l'ossequio alla tradizione di un genere letterario? E ancora: la decostruzione del finale escatologico rassicurante vuole suggerire il trasferimento della colpa dall'individuo alla società che ha pianificato e attuato l'occupazione coloniale? O è invece un modo per dichiarare la non colpevolezza per i crimini commessi in quanto semplici incidenti, e lo svolgimento della trama va inteso piuttosto come un percorso interiore nella coscienza (dove l'"altro" è solo una componente del paesaggio)? Anche se gli elementi testuali e contestuali scoraggiano fortemente l'interpretazione del romanzo nella cornice tradizionale del genere letterario coloniale, la domanda rimane: perché l'autore avrebbe omesso di chiarire la sua posizione sul colonialismo, nascondendo la critica dietro l'ambigua e sfuggente posizione morale del suo personaggio?

2. OBIETTIVI

Questo studio si colloca all'interno del dibattito critico sulla posizione di Flaiano nei confronti del colonialismo in *Tempo di uccidere*. L'obiettivo principale è di contribuire al dibattito introducendo nuovi elementi derivati dall'analisi computazionale. In particolare, la ricerca approfondisce la complessità e l'ambiguità narrativa concentrandosi sul meccanismo della focalizzazione messo in atto nel romanzo (Genette, 1980; Bal, 2009), non solo nella sua bipartizione tra io-agente e io-narrante extradiegetico, ma anche nella forma di *embedded focalizations* (O'Neill, 2002). Nel romanzo, infatti, il protagonista mostra spesso di trasferire la propria prospettiva immedesimandosi psicologicamente nei personaggi che incontra (indigeni o italiani) – immaginandone, cioè, sentimenti, aspettative, intenzioni – secondo quelle che possono essere definite "finestre di focalizzazione" (Jahn, 1996). L'ipotesi da cui muove la ricerca è che proprio l'analisi del complesso meccanismo di focalizzazioni di primo (io-narratore), secondo (io-agente) e terzo grado (personaggi) possa fornire elementi per l'interpretazione del romanzo in senso coloniale o

anticoloniale. Se giungere a una soluzione dirimente non è plausibile per la costruzione stessa del *modus narrandi* di Flaiano – così come emerge anche dai suoi scritti giornalistici, in cui la componente ironica è un tratto distintivo – nondimeno l'analisi mira a supportare la critica interpretativa fondata sulla lettura diretta del testo con elementi testuali linguistici valutati anche nella loro componente quantitativa. Un secondo obiettivo, metodologico, intende poi verificare se le conclusioni raggiunte tramite l'analisi quantitativa tradizionale siano ottenibili anche sfruttando l'in-context learning di un LLM. Tale approccio, condotto in forma sperimentale ed esplorativa, è mirato a valutare l'efficacia dei LLM per compiti di analisi e interpretazione letteraria e a verificare la coerenza tra osservazioni umane, analisi computazionali quantitative e analisi generata da LLM, in un processo reciprocamente valutativo.

3. METODOLOGIA

La ricerca è condotta in tre fasi cronologicamente ordinate.
1) Nella prima fase, si è proceduto all'annotazione manuale del testo integrale del romanzo, secondo categorie definite in base agli obiettivi sopra esposti (per l'annotazione dei livelli narrativi, si veda Kearns, 2021). L'annotazione è stata condotta a livello di frase e senza possibilità di sovrapposizione delle etichette, da parte di un annotatore principale e di un secondo annotatore che ha condotto l'annotazione su una porzione di testo (4 capitoli randomici su 7 totali) per validare il risultato della prima annotazione mediante il calcolo dell'Inter Rater Reliability (IRR, k = 0,82). Le categorie di annotazione sono:

- "Narrator": focalizzazione sull'io-narrante, che interrompe occasionalmente la focalizzazione dell'io-agente. Questa categoria è suddivisa in tre sottocategorie, in base agli elementi marcati (linguistici e semantici) che rendono rilevabile il cambiamento di prospettiva: "Indexicals": riferimenti di luogo/tempo o altro non riconducibili al mondo narrativo dell'io-agente (Klauk et al., 2012); "Knowledge": grado di conoscenza superiore a quello dell'io-agente; "Judgement": commenti o giudizi espressi sui pensieri e le azioni dell'io-agente.
- "Conqueror": focalizzazione sui soldati, filtrata attraverso l'identificazione dell'io-agente.
- "Indigenous": focalizzazione sugli indigeni, filtrata attraverso l'identificazione dell'io-agente.

Alle categorie sopra elencate se ne aggiunge una quarta:

- "Description": all'interno di questa categoria sono distinte descrizioni di "Landscape", "Indigenous" e "Conqueror", a seconda dell'oggetto descritto, con l'interesse di isolare per l'analisi le ultime due categorie e ricavare gli elementi di proiezione dell'interiorità del protagonista sui personaggi esterni secondo la dinamica colonizzati/colonizzatori.

Pur con uno statuto narratologico differente, questa categoria risulta informativa in quanto la descrizione è focalizzata sul protagonista agente, esprimendone le rappresentazioni del mondo circostante. Nella tabella seguente (vd. Tab. 1), sono riportate le statistiche descrittive delle annotazioni per le categorie principali.

Categoria	**Num. annotazioni**	**Parole totali**	**Parole per segmento (media)**
Nessuna annotazione	4831	63841	13,21
Indigenous	367	7435	20,26
Description	233	5092	21,85
Conqueror	224	4099	18,30
Narrator	100	1563	15,63

Tabella 1. Distribuzione delle annotazioni per categoria principale

2) Nella seconda fase, le frasi etichettate sono state sottoposte a varie analisi esplorative, volte a individuare elementi linguistici informativi in relazione agli obiettivi.[1] Nello specifico:

- Parole più frequenti (MFW): analisi delle 80 parole più frequenti per ciascuna categoria, seguita da un'analisi qualitativa dei risultati. Questo approccio è stato preferito ad altre misure di

[1] I risultati completi di questa fase sono pubblicati in Lilli & Raffini (2026).

rappresentatività statistica (come Log-Likelihood Ratio e TF-IDF) tendenti a rappresentare parole legate al singolo contesto (ad es. nomi di persona).

- Analisi stilometrica: si è tentato un approccio non supervisionato usando il pacchetto Stylo (Eder et al., 2016), per verificare se le diverse categorie di annotazione presentassero differenze linguistiche rilevabili dall'analisi dei cluster, e a quale livello di frequenza lessicale (parole grammaticali o parole di contenuto). È stata esclusa la categoria "Narrator" per le occorrenze limitate; le frasi annotate sono state segmentate in blocchi di 100 parole, combinati casualmente in 5 gruppi, e l'analisi è stata eseguita su un corpus bilanciato (8 blocchi per categoria) o completo, e testata con vari parametri: i risultati ottimali sono stati ottenuti con l'intero corpus e il coseno Delta come misura della distanza.
- Distribuzione delle parti del discorso (POS): l'annotazione grammaticale è stata effettuata tramite la libreria Python Stanza (Qi et al., 2020). La significatività delle distribuzioni è stata valutata con il test del chi-quadro con correzione di Bonferroni, e l'effetto è stato misurato con V di Cramér. Le parole delle tre categorie grammaticali a maggiore contenuto semantico (sostantivi, verbi e aggettivi) più distintive per ciascuna categoria di annotazione (in base al rapporto frequenza_categoria_A + smoothing / frequenza_categoria_B + smoothing) sono state oggetto di un'analisi qualitativa.

3) Nella terza fase, si è sperimentata l'analisi del corpus di annotazioni dei passaggi focalizzati e descrittivi illustrati sfruttando le capacità dell'in-context learning di un Large Language Model (Chat-GPT 5.2). Il file contenente il romanzo annotato (in formato .xlsx) è stato fornito al LLM. Un prompt strutturato ha definito il protocollo di analisi, basato su aderenza ai dati per ogni risposta.[2] In seguito, si è verificata l'efficacia del LLM nel leggere le annotazioni attraverso 10 domande contenenti compiti di estrazione di passaggi annotati e delle etichette. In seguito alla verifica della lettura delle annotazioni, sono state fornite al modello le definizioni delle varie etichette, attraverso un prompt di guida all'analisi semantica. Si è dunque proceduto all'analisi del romanzo. A tale scopo, si è costruito un set di 20 domande relative alla tecnica di focalizzazione nel romanzo e alla rappresentazione dei colonizzati e dei colonizzatori attraverso la focalizzazione e la descrizione. Le prime 12 domande ricalcavano i task già analizzati nelle analisi computazionali, con un approccio quantitativo basato su richiesta esplicita di calcolo, mentre ulteriori 8 domande avevano un approccio più interpretativo-discorsivo. Queste ultime sono state sottoposte al LLM in un'altra finestra conversazionale fornendo solamente il testo del romanzo, senza annotazioni. In questo modo è stato possibile confrontare la differenza di prestazioni nel caso in cui di fornisca al modello un testo annotato o un testo non annotato.

4. DISCUSSIONE DEI RISULTATI DELL'ANALISI QUANTITATIVA

L'analisi esplorativa delle annotazioni ha rivelato in primo luogo la consistenza delle annotazioni a livello linguistico. Sia l'analisi dei cluster non supervisionata, sia l'analisi della distribuzione delle POS[3] hanno confermato che vi è una differenza statisticamente significativa a livello di parole più frequenti e di categorie grammaticali tra le diverse categorie di annotazione, confermando che esse identificano sezioni linguisticamente distinte della narrazione. L'analisi semantica delle parole più frequenti e delle parole più distintive per categoria grammaticale ha permesso di approfondire come si realizzi tale differenza a livello semantico, consentendo di formulare osservazioni circa la domanda di ricerca iniziale.
Nello specifico, interessante è l'incidenza della congiunzione avversativa "ma" nella categoria "Narrator", suggerendo l'intervento correttivo del narratore rispetto alle affermazioni o azioni del personaggio agente, mentre ben rappresentate in tutte le categorie sono espressioni modali, tramite avverbi ("forse") o modi verbali (condizionale), che sottolineano l'effetto di inaffidabilità del narratore come evidenziato nell'introduzione. Tipicamente attribuito alla categoria del "Conqueror" è invece il verbo di volontà ("voleva"), statisticamente rilevante anche quando confrontato con le occorrenze di "volevo" nelle parti non annotate (38 occorrenze contro 61), suggerendo che l'io-agente riconosce ai membri dell'esercito con cui entra in contatto un'agentività dominante anche in rapporto alla propria volizione, pur riconoscendosi

[2] I prompt e la conversazione con il LLM sono disponibili al link: https://github.com/slilli23/Lilli_Raffini_AIUCD2026.git.
[3] La cluster analysis non supervisionata ha raggiunto una quasi corretta distinzione delle parti annotate per categorie già con i parametri 100 MFW, 0% culling, classic Delta, e un'assegnazione perfetta con 300 MFW. Il test di significatività effettuato sulle distribuzioni delle POS ha rivelato 11 categorie grammaticali su 13 mostrano una correlazione significativa con le categorie di annotazione.

una predominanza epistemica (misurata in base alle occorrenze di "sapevo", 53 contro 23). Un altro elemento di distinzione tra le categorie è relativo al trattamento del genere: la parola "donna", infatti, compare estensivamente nei passaggi focalizzati sugli indigeni, mentre la parola "uomo" compare esclusivamente nelle descrizioni in relazione ai soldati italiani.[4] Questo sembra avvalorare l'uso della caratterizzazione di genere non come identificatore, ma piuttosto come classificatore che risente degli stereotipi della propaganda: l'indigeno associato al ruolo passivo (donna), il soldato al ruolo attivo (uomo).
Ciononostante, più sfumata è la rappresentazione degli stereotipi come emerge dall'analisi qualitativa delle parti del discorso. Il processo di depersonalizzazione investe in pieno i soldati (identificati solo con i loro ruoli gerarchici e professionali nell'esercito, ma anonimi) rispetto agli indigeni (identificati con i nomi propri, tutti di ascendenza biblica). Inoltre, i verbi e i sostantivi associati ai soldati identificano primariamente oggetti concreti e suggeriscono una rappresentazione tipizzata e mediocre (con termini che rimandano alla pigrizia, al machismo, all'arrivismo), laddove associati agli indigeni sono verbi che identificano attività quotidiane e sostantivi astratti o altamente generici ("tempo", "presenza", "lavare", "villaggio"). Ancora, si osserva una preponderanza di verbi transitivi nel caso dei colonizzatori ("amare", "insegnare", "convincere") e di verbi relativi a stati interiori nel caso degli indigeni ("riconoscere", "sorridere"), alternati a verbi chiaramente esprimenti la dinamica relazionale fondata su rapporti di potere ("seguire", "ammettere", "respingere").

5. DISCUSSIONE DEI RISULTATI DELL'ANALISI TRAMITE LLM

Dall'analisi sviluppata tramite LLM emergono risultati rivelanti sia per la capacità di lettura di testi annotati, che per le capacità interpretative. Si espongono di seguito le principali emergenze per quanto riguarda la lettura dei testi annotati:

- Il LLM è in grado di leggere in maniera efficace le annotazioni, riconoscere le colonne e le righe correttamente, comprendere autonomamente la tassonomia delle etichette, individuare i collegamenti tra porzioni di testo ed etichette, ricostruire porzioni di testo.
- Il LLM si dimostra poco incline a seguire il protocollo di aderenza ai dati stabilito all'inizio, rendendo necessario un richiamo ad esso almeno ogni 2-3 prompt, al fine di evitare allucinazioni. Dopo il richiamo, il protocollo viene seguito correttamente.
- L'interazione dimostra l'efficacia della richiesta "ragiona almeno 30 secondi" nel prompt, al fine di assicurare il rispetto del protocollo ed evitare allucinazioni

Dopo aver stabilito il protocollo di analisi e l'efficacia nella lettura del file, si è passato all'analisi dei contenuti, con particolare attenzione alla questione della focalizzazione. Per fare ciò, sono state fornite al modello le definizioni delle etichette e indicazioni su come collegarle ai segmenti testuali per una corretta interpretazione. In particolare, è stato chiesto di utilizzare esplicitamente le etichette come parte dell'analisi, supportare l'interpretazione con riferimenti a segmenti testuali specifici, distinguere le differenze semantiche tra le etichette, verificare la coerenza tra contenuto testuale ed etichettatura.
Sono state quindi fornite le domande di analisi. Le domande del primo gruppo ricalcano le analisi qualitative condotte nella fase precedente e si basano sulla richiesta esplicita di conteggio delle parole per analizzare le focalizzazioni e le differenze linguistiche tra le etichette per quanto riguarda sostantivi, verbi ed aggettivi. Sui task di calcolo il LLM si è rivelato poco efficace, in alcuni casi fornendo conteggi sbagliati e in altri dichiarando l'impossibilità di analizzare il corpus completo (come stabilito da protocollo).[5] Non riuscendo a fornire calcoli esatti, il modello ha proposto analisi per campionatura, supportate da passaggi significativi. In questo modo il LLM è riuscito a cogliere alcune caratteristiche già emerse dall'analisi computazionale classica. Fermo restando che alcuni elementi possono essere dedotti dal modello perché intuitivi o presenti nelle definizioni fornite, altri elementi meno scontati dimostrano come il meccanismo di attenzione dei LLM, pur non riuscendo ad esprimere conteggi esatti, tuttavia sia in grado di cogliere aspetti peculiari del testo. Il LLM riesce ad esempio individuare la maggior frequenza delle descrizioni dei paesaggi rispetto a quelle dei personaggi; lega correttamente i personaggi colonizzatori alle aree

[4] L'assenza del termine donna in relazione ai soldati può ascriversi a un'effettiva assenza di donne nell'ambito militare, tuttavia non è vero il contrario, dal momento che sono frequenti anche relazioni con uomini indigeni.
[5] Nello specifico, il modello lega l'impossibilità all'espletamento dei task di conteggio all'impossibilità di operare una tokenizzazione, lemmatizzazione e POS tagging sull'intero file.

semantiche dell'organizzazione, dell'azione e controllo e non a quelle di eroismo e aggressività, plausibili a livello teorico, ma non effettivamente rappresentate nel romanzo; individua la predominanza nelle parti relative alle popolazioni locali di termini generici e della dimensione corporea e quotidiana. Un limite evidente è la tendenza alla ripetizione di schemi: una volta individuate alcune parole chiave, il LLM tende a rispondere in base a quelle anche nelle risposte successive. Ciò determina che le caratteristiche individuate per i colonizzatori e per gli indigeni nell'analisi dei sostantivi siano replicate in maniera quasi identica nell'analisi dei verbi e degli aggettivi. Possiamo dunque concludere che le domande impostate in senso quantitativo risultano inefficaci per ottenere conteggi reali, abbastanza efficaci quando il modello passa dalla richiesta di conteggio alla campionatura, ma troppo tendenti a uniformare il discorso e a ripetere schemi.
Si è passato poi a sottoporre al modello 8 domande impostate su un approccio qualitativo e discorsivo. Questa modalità si è rivelata più efficace e non ha prodotto le allucinazioni dell'approccio quantitativo, determinando una maggiore coerenza espositiva. Molto efficace in questo approccio è la capacità di supportare le affermazioni con esempi testuali, come richiesto dal protocollo di analisi. Buona risulta anche la capacità di analisi e commento dei singoli segmenti testuali estratti dal romanzo. Un dato estremamente interessante è che alla domanda "quale delle focalizzazioni ti sembra più frequente?" il modello ha fornito autonomamente i conteggi corretti, che invece aveva sbagliato quando gli era stato chiesto specificatamente di contare. Alcuni elementi emersi dall'analisi quantitativa sono confermati in questa fase anche dal LLM: la predominanza nei colonizzatori di "una lingua più orientata all'agency e alla volontà, spesso applicata a figure socialmente definite: soldato, generale, contrabbandiere", e dunque l'appiattimento sui ruoli; mentre per gli indigeni nota una "lingua più mediata, congetturale, corporea e osservativa" e l'attribuzione di attività quotidiane. Elementi nuovi emergono, supportati dalle evidenze testuali, come la tendenza degli indigeni a esprimere agency principalmente nel rifiuto. In alcuni casi l'analisi è meno accurata: il linguaggio del dubbio – rappresentato dall'avverbio "forse" – viene legato dal LLM ai passaggi legati agli indigeni, mentre le analisi quantitative dimostravano una distribuzione uniforme, che determina come abbiamo visto più una postura del narratore che una caratterizzazione di specifici gruppi di personaggi. L'ultima domanda riguardava il posizionamento dell'autore rispetto ai concetti di colonialismo e postcolonialismo: la risposta mette in luce la rappresentazione problematica e irrisolta del soggetto colonizzatore ma allo stesso tempo la mancanza di attribuzione di soggettività ai colonizzati, risolvendosi in "una critica del colonialismo dall'interno della sua stessa prospettiva, senza però giungere a una piena riconfigurazione postcoloniale dell'alterità".
Come ultima fase sperimentale, sono state sottoposte le 8 domande qualitative fornendo il testo senza annotazioni. Il LLM individua "una costellazione di focalizzazioni interne: percettiva, riflessiva, immaginativa e visionaria" e le descrive puntualmente. Sebbene riconducibili effettivamente al romanzo, queste categorie non definiscono tanto il concetto di focalizzazione in senso narratologico, quanto delle modalità dello sguardo di un unico narratore. Il concetto di "focalizzazione immaginativa" si avvicina alla "focalizzazione *embedded*" (O'Neill, 2002), perché definito come l'attribuzione da parte del protagonista di "pensieri, intenzioni e psicologie agli altri personaggi". Assente invece la distinzione tra narratore-agente e narratore postumo. I personaggi vengono divisi tra "interni e simmetrici" e "altri e opachi", corrispondenti in parte con le categorie da noi etichettate come "Conqueror" e "Indigenous". Interessante l'attribuzione di un ruolo a parte al personaggio della donna indigena, per la quale la focalizzazione sarebbe "radicalmente interna e inaffidabile, perché l'alterità è massima ma il narratore tenta comunque di appropriarsene". Il LLM individua la focalizzazione più frequente come quella "interna riflessiva e soprattutto immaginativa", unendo la focalizzazione del narratore e la focalizzazione *embedded*. Uno spunto interessante viene dal collegamento tra focalizzazione immaginativa (*embedded*) e lingua ipotetica e oscillante, suggerendo la possibilità di attribuire la componente dubitativa non tanto alla dicotomia tra "Conqueror" e "Indigenous" quanto a quella tra pensieri propri e pensieri attribuiti. Alla domanda sulle differenze tra rappresentazione degli italiani e degli indigeni, si nota un appiattimento sulla dinamica di distanza e vicinanza, in cui gli italiani sono comprensibili dal narratore e gli indigeni sono indecifrabili. Viene meno, quindi, la problematicità del rapporto tra il narratore e i colonizzatori. In questo caso l'analisi del file annotato aveva invece rilevato caratteristiche più complesse e profonde. A parte alcuni casi, come ad esempio l'individuazione dell'agency per rifiuto degli indigeni e l'assenza di eroismo dei conquistatori, l'analisi senza annotazione mostra meno profondità e anche meno capacità di collegamento in generale e di collegamento tra analisi e segmenti testuali specifici. L'ultima domanda – sul rapporto con colonialismo e postcolonialismo – ribadisce la messa in discussione del colonialismo dell'interno ma anche la mancanza di un "ribaltamento di voce" che avrebbe reso il testo postcoloniale.

6. CONCLUSIONI

Lo studio condotto ha carattere esplorativo, ma permette di trarre alcune conclusioni, sia a livello metodologico che di interpretazione del romanzo. A livello metodologico il confronto tra analisi computazionali classiche e analisi con LLM ha mostrato una buona convergenza di risultati, ma anche la necessità di approcciare diversamente all'analisi tramite LLM rispetto alle analisi computazionali classiche basate su conteggio di parole. I limiti dei LLM stanno nella minore affidabilità e verificabilità; tuttavia, hanno il pregio di fornire dei risultati più direttamente interpretabili rispetto alle metodologie di Digital Humanities classiche. Dal confronto tra l'analisi tramite LLM con testo annotato e non annotato emerge che fornire testi annotati migliora in particolare la correttezza dei concetti narratologici, la capacità di riferirsi a precisi segmenti testuali, oltre a un maggiore approfondimento ed esattezza dal punto di vista interpretativo. Resta, comunque, la necessità di replicare la sperimentazione, confrontando la performance anche di altri modelli linguistici.
A livello di contenuti, le analisi quantitative e qualitative hanno permesso di definire meglio le modalità di focalizzazione in relazione ai diversi gruppi di personaggi. Ciò che emerge, conferma l'interpretazione di *Tempo di uccidere* come una lettura critica del colonialismo, che mette in discussione la rappresentazione eroica del colonizzatore, ma che non può dirsi ancora pienamente postcoloniale.

DICHIARAZIONI

Gli autori dichiarano che strumenti di IA generativa sono stati utilizzati per questo studio nelle seguenti fasi: programmazione in Python, analisi statistica e formale, revisione stilistica, copy-editing, metodologia di analisi testuale e oggetto di studio. L'ideazione, la progettazione, la conduzione della sperimentazione e la revisione complessiva del testo sono stati svolti in piena collaborazione dai due autori. La scrittura materiale è stata invece così suddivisa: Lilli paragrafi 1, 2, 3.1, 3.2, 4; Raffini 3.3, 5, conclusioni.

BIBLIOGRAFIA

Bal, M. (2009). Narratology: Introduction to the Theory of Narrative (3rd edition). Toronto: University of Toronto Press.

Baraldi, M. (2004). Il cuore di tenebra di un uomo ridicolo. Quaderni del '900, 4, 97-104.

Bazzocchi, M. A. (2012). Il corpo e le piaghe: L'Africa di Flaiano. Narrativa, 33-34, 301-309.

Benvenuti, G. (2012). Da Flaiano a Ghermandi: riscritture postcoloniali. Narrativa, 33–34, 311-321.

Bongie, C. (1991). Exotic Memories: Literature, Colonialism and the Fin de Siecle. Stanford: Stanford University Press.

Booth, W. C. (1961). The Rhetoric of Fiction. Chicago: University of Chicago Press.

Brunetti, B. (2010). Modernità malata. Note su Tempo di uccidere di Ennio Flaiano. Fuori centro. Percorsi postcoloniali nella letteratura italiana. Derobertis R. (Ed.). 57-71. Roma: Aracne Editrice.

Casales, F. (2023). Raccontare l'Oltremare. Storia del romanzo coloniale italiano (1913-1943). Milano: Mondadori.

Del Boca, A. (1992). L'Africa nella coscienza degli italiani. Miti, memorie, errori, sconfitte. Roma-Bari: Laterza.

Del Boca, A. (1998). Il colonialismo italiano tra miti, rimozioni, negazioni e inadempienze. Italia contemporanea, 212, 589–603.

Domenichelli, M. (1997). Il contagio della terra straniera. Lo straniero. Atti del Convegno di Studi (Cagliari, 16-19 novembre 1994). Domenichelli M. and Fasano P. (Eds.). 2, 645-660. Roma: Bulzoni.

Dong, Q. Li, L., Dai, D., Zheng, C., Ma, J. Li, R. Xia, H. Xu, J., Wu, Z., Chang, B., Sun, X., Li, L. and Sui, Z. (2024). A Survey on In-context Learning. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. 1107-1128. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.64.

Eder, M., Rybicki, J., & Kestemont, M. (2016). Stylometry with R: a package for computational text analysis. R Journal, 8, 1, 107-121. https://doi.org/10.32614/RJ-2016-007.

Flaiano, E. (2020). Aethiopia. Appunti per una canzonetta. Milano: Adelphi (ebook).

Genette, G. (1980). Narrative Discourse. An Essay in Method, Foreword by Jonathan Culler. Ithaca (NY): Cornell University Press.

Hart, C. (2010). Critical Discourse Analysis and Cognitive Science: New Perspectives on Immigration Discourse. London: Palgrave Macmillan. https://doi.org/10.1057/9780230299009_2.

Jahn, M. (1996). Windows of Focalization: Deconstructing and Reconstructing a Narratological Concept. Style, 30, 2, Rhetoric and Poetics, 241-267. https://www.jstor.org/stable/45063726.

Kearns, E. (2021). Annotation guidelines for narrative levels, time features, and subjective narration styles in fiction (SANTA 2). Journal of Cultural Analytics, 6, 4, 53-68. https://doi.org/10.22148/001c.30699.

Klauk, T., Köppe, T. & Onea, E. (2012). Internally focalized narration from a linguistic point of view. Scientific Study of Literature, 2, 2, 218-242. https://doi.org/10.1075/ssol.2.2.03kla.

Leeuwen, T.V. (2008). Discourse and Practice: New Tools for Critical Discourse Analysis. Oxford: Oxford University Press.

Lilli, S. & Raffini, D. (2026). Decoding the Conqueror's Gaze. A Computational Approach to Ennio Flaiano's (Post)Colonialism. Computational Humanities Research, 1–24. https://doi.org/10.1017/chr.2024.1.

Maxia, S. (2010). Una stagione in Etiopia. Tempo di uccidere di Ennio Flaiano. Appunti di lettura. L'occhio e la memoria. Miscellanea di studi in onore di Natale Tedesco. 2, 225-249. Caltanissetta: Editori Del Sole, Edizioni Lussografica, Salvatore Sciascia Editore.

Mengozzi, C. (2016). Lo sguardo e la colpa: Tempo di uccidere di Ennio Flaiano e la dialettica servo-signore alla prova del colonialismo. MLN, 131, 1, 174-195. https://doi.org/10.1353/mln.2016.0010.

Nasson, L. (2012). Alla base di ogni espansione, il desiderio sessuale. Negotiating Exoticism and Colonial Conquest in Ennio Flaiano's *Tempo di uccidere*. Italian Studies in Southern Africa/Studi d'Italianistica nell'Africa Australe, 25, 1, 40-58.

O'Neill, P. (2002). Points of origin: On focalization in narrative. Canadian Review of Comparative Literature / Revue Canadienne de Littérature Comparée, 19, 331-350.

Orlandini, R. (1992). (Anti)colonialismo in Tempo di uccidere di Ennio Flaiano. Italica, 69, 4, 478-488.

Palumbo, P. (2002). National Integrity and African Malaise in Ennio Flaiano's *Tempo di uccidere*. Forum Italicum, 36, 53-68.

Pautasso, S. (1992). Attualità di un romanzo. La critica e Flaiano. Sergiacomo L. (Ed.). 107-110. Pescara: Associazione Culturale Ennio Flaiano-Ediars.

Prisco, M. (2012). Il significato dello svelamento dei crimini del colonialismo italiano in *Tempo di uccidere*, *Regina di fiori e di perle* ed *Oltre Babilonia*. Quaderni del '900, 12, 103-108.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J. & Manning, C.D. (2020). Stanza: A Python natural language processing toolkit for many human languages. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Celikyilmaz, A. & Wen, T.-H. (Eds.). 101-108. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-demos.14.

Skocki, T. (2012). Il soldato italiano in colonia e il suo rapporto con l'Altro. *Il Giovane maronita* di Alessandro Spina e *Tempo di uccidere* di Ennio Flaiano. Narrativa, 33-34, 333-342.

Stefani, G. (2007). Colonia per maschi. Italiani in Africa Orientale: una storia di genere. Verona: Ombre Corte.

Sternberg, M. & Yacobi, T. (2015). (Un)reliability in narrative discourse: A comprehensive overview. Poetics Today, 36, 4, 327-498. https://doi.org/10.1215/03335372-3455114.

Verso una rappresentazione del macrotesto: filologia d’autore e analisi quantitativa nei racconti di Fausta Cialente

EMMANUELA CARBÈ
UNIVERSITÀ CA’ FOSCARI DI VENEZIA

ABSTRACT (ITALIANO)

La narrativa breve di Fausta Cialente è in gran parte pubblicata su rivista negli anni Trenta e parzialmente confluita nelle sillogi *Pamela o la bella estate* (1962) e *Interno con figure* (1976), due macrotesti che sono l’esito di processi di selezione, riscritture e rimaneggiamenti. Il contributo propone un’integrazione tra analisi filologiche e quantitative attraverso pratiche di modellazione e rappresentazione delle micro- e macrostrutture testuali, con l’obiettivo di interrogare geneticamente i racconti di Cialente senza orientare l’indagine alla produzione di un’edizione scientifica digitale. A partire da un modello di metadati che distingue tra unità testuali astratte e attestazioni storiche, e dalla costituzione di un *corpus* di testi, si presentano alcuni esperimenti esplorativi per osservare i percorsi editoriali dei racconti e le logiche di costruzione delle due raccolte. Si verifica così se e in che misura sia possibile rappresentare la dinamicità del macrotesto e formulare ipotesi di lavoro per indagini filologico-critiche integrate, applicabili soprattutto a testi della tradizione novecentesca ancora soggetti a vincoli di diritto d’autore. In prospettiva, il modello è pensato per includere l’analisi variantistica di materiali preparatori cartacei, ibridi e *born-digital*.
Parole chiave: filologia digitale; critique génétique; metodi quantitativi; macrotesto; letteratura italiana contemporanea

ABSTRACT (ENGLISH)

Towards a Representation of the Macrotext: Authorial Philology and Quantitative Analysis from the Short Stories of Fausta Cialente. Fausta Cialente’s short fiction was largely published in periodicals during the 1930s and later partially reassembled in the collections *Pamela o la bella estate* (1962) and *Interno con figure* (1976), two macrotexts resulting from processes of selection, rewriting, and reworking. The paper proposes an integration of philological and quantitative analyses through practices of modelling and representation of textual micro- and macro-structures, with the aim of investigating Cialente’s short stories from a genetic perspective without directing the inquiry toward the production of a Digital Scholarly Edition. Based on a metadata model that distinguishes between abstract textual units and individual witnesses, and on the construction of a *corpus* of texts, the paper presents a set of exploratory experiments designed to observe the editorial trajectories of the short stories and the compositional logics underlying the two collections. The aim is to assess whether, and to what extent, it is possible to represent the dynamic nature of the macrotext and to formulate hypotheses for integrated philological-critical investigations, particularly in relation to twentieth-century literary texts that remain subject to copyright restrictions. In the longer term, the model is conceived as extensible to the variant analysis of preparatory materials, including paper-based, hybrid, and born-digital documents.
Keywords: digital philology; genetic criticism; quantitative methods; macrotext; Italian contemporary literature

1. INTRODUZIONE

Una raccolta di racconti può essere intesa, sotto alcune condizioni, come una configurazione macrotestuale la cui coerenza è ottenuta grazie a scelte autoriali di selezione, organizzazione e composizione dei materiali (Corti 1975), configurandosi così come un dispositivo testuale complesso, in cui possono intervenire processi di riuso e riscrittura di testi già pubblicati in sedi diverse. È questo il caso della narrativa breve di Fausta Cialente (1898-1994): i racconti risalgono prevalentemente agli anni Trenta, mentre le sillogi *Pamela o la bella estate* (1962) e *Interno con figure* (1976) raccolgono e rielaborano a distanza di decenni una selezione di testi. Questa stratificazione cronologica si presta a un’analisi filologico-critica orientata ai processi di costruzione del macrotesto, e mette alla prova le ipotesi di lavoro di una filologia d'autore supportata da metodi computazionali.

In ambito DH l'analisi assistita dei testi letterari ha storicamente conosciuto sviluppi discontinui (Ciotti 2017; 2023); applicata solo di rado alla letteratura contemporanea (Hoover 2008; 2013), ha tuttavia trovato nell'informatica umanistica esempi significativi e forse troppo sottovalutati: il pensiero va anzitutto ai lavori di Giuseppe Gigliozzi, di cui si ricorda almeno l'analisi di *Memoriale* di Paolo Volponi (Gigliozzi 1996; Gigliozzi 1997). Le ragioni per cui queste pratiche non abbiano trovato un terreno fertile, né abbiano prodotto maggiori ibridazioni in altri campi come quello filologico-critico (ma si veda Boschetti 2021), sono molteplici: un fattore rilevante può essere individuato nella nascita dell'iniziativa TEI e nel progressivo consolidamento di modelli per le edizioni scientifiche digitali, che hanno indirizzato una parte degli studi verso la formalizzazione dei testi e della variantistica attraverso modelli basati prima su SGML e poi su XML (la versione TEI P3 risale al 1994). Pur avendo avuto un ruolo fondamentale nella costruzione di metodi e progetti di grande importanza, tale centralità può avere contribuito a privilegiare problemi di modellazione e visualizzazione dell'edizione, lasciando relativamente sullo sfondo l'analisi computazionale del testo.
Un rinnovato interesse per l'analisi quantitativa, spesso declinato nelle esperienze del *distant reading* (Moretti 2005; 2013; 2022) e della *macroanalysis* (Jockers 2013), ha riaperto il dibattito sui rapporti tra computazione e critica letteraria. Questa ripresa si è tuttavia concentrata soprattutto su grandi corpora o su approcci teorici a campo lungo più che su verifiche puntuali legate all'esplorazione filologica del testo. Rimane dunque ancora parzialmente inesplorata, sebbene non manchino contributi teorici fondativi (Fiormonte 2003),[1] l'integrazione tra analisi quantitativa e filologia d'autore come spazio teorico e metodologico a sé stante. L'ipotesi di fondo è che la prospettiva computazionale possa sostenere indagini filologiche non solo e non necessariamente finalizzate all'edizione, ma anche all'esplorazione analitica di dossier genetici digitali.

2. I RACCONTI DI FAUSTA CIALENTE: MODELLO E CORPUS

La narrativa breve di Cialente è rimasta a lungo poco accessibile alla critica per la difficoltà di condurre spogli sistematici su periodici pubblicati all'estero e in particolare in Egitto, dove la scrittrice visse stabilmente tra il 1921 e il 1947. La dispersione delle fonti e la limitata accessibilità dei materiali hanno contribuito a rendere parziale e frammentario il quadro complessivo della sua produzione, incidendo non solo sulla puntuale ricostruzione cronologica degli scritti, ma anche sulla possibilità di osservare i processi di rielaborazione e riscrittura dei suoi racconti, in parte confluiti nelle raccolte *Pamela o la bella estate* (1962) e *Interno con figure* (1976).
La campagna di digitalizzazione del "Giornale d'Oriente" (El Beih 2024), organo di informazione della colonia italiana in Egitto, consente oggi di affrontare in modo più solido la definizione del *corpus*, pur nella consapevolezza che il censimento degli interventi di Cialente su riviste e periodici non può ancora dirsi concluso. Dallo spoglio e dalla trascrizione del materiale (supportata da OCR e successiva revisione) sono stati individuati 129 testi a firma della scrittrice, di cui 28 racconti pubblicati tra il 1930 e il 1940.[2]
Integrando questi dati con quanto già noto, il quadro si attesta su 32 racconti complessivi, riconducibili salvo poche eccezioni agli anni Trenta. La sede principale è "Il Giornale d'Oriente", con pubblicazioni circoscritte sull'"Italia letteraria" e su "Occidente". Negli anni Quaranta si registrano rare ripubblicazioni e un solo nuovo racconto su "Fronte Unito" e sul "Mattino della Domenica", giornali antifascisti diretti da Cialente (Carbé 2021: 18-20). Dopo il rientro in Italia nel 1947, alcuni racconti vengono rimaneggiati e riproposti, accanto a pochi inediti, in sedi come "l'Unità" e "Noi Donne", fino alla selezione e sistemazione nella silloge del 1962 e al suo ampliamento nel 1976.
Per il modello si è scelto di assumere come unità primaria l'evento (la pubblicazione, la riscrittura, l'inclusione in raccolta), distinguendo poi il racconto come *work*, entità testuale astratta identificata da un nucleo narrativo riconoscibile, dalle sue diverse attestazioni storiche (*witness*) corrispondenti alle singole pubblicazioni su periodico, in antologia o in volume. Le attestazioni sono dunque intese come output di eventi situati in un tempo e in uno spazio precisi: ogni *witness* è descritto attraverso un set di metadati

[1] Non mancano, come è noto, progetti e ambienti di interfaccia per l'analisi computazionale dei testi, ad esempio *Voyant Tools* di Sinclair e Rockwell, così come esperimenti orientati alla promozione della lettura, tra cui, nel panorama italiano, *Macchine per leggere* (Ciotti e Baldi 2022).
[2] La trascrizione degli scritti rinvenuti nel "Giornale d'Oriente" è stata realizzata da un gruppo di ricerca costituito da Wafaa El Beih, Laura Cannavacciuolo, Nunzio Ruggiero, e da chi scrive. Nel presente contributo, circoscritto alla sola narrativa breve, sono stati esclusi due estratti dal romanzo *Cortile a Cleopatra* (1936), apparsi rispettivamente il 1° gennaio 1931 e il 30 giugno 1936 con i titoli *Partenza di Marco* e *Il pianto di Dinah*.

che registrano titolo della manifestazione, sede editoriale, estensione e data di pubblicazione modulata sulla granularità della fonte (al giorno per le uscite su periodico, all'anno per quelle in volume); per le pubblicazioni in due puntate è registrata anche la data della seconda uscita.
L'obiettivo è la costruzione di diversi livelli di astrazione, capaci di rendere analizzabili: 1. le relazioni tra le attestazioni di un medesimo *work*, attraverso l'individuazione delle varianti tra *witness* mediante procedimenti di collazione manuale o assistita; 2. l'evoluzione diacronica dei testi tramite analisi computazionali supportate da tecniche NLP; 3. le differenze stilistiche tra i *subcorpora* della fase degli anni Trenta e quelli delle raccolte, per collocare le trasformazioni anche sul piano stilistico e lessicale; 4. la struttura dei rapporti tra i testi, le loro ricorrenze, i loro spostamenti e le loro ricontestualizzazioni nel tempo. Il corpus si configura così come una rete di relazioni genetiche tra *work*, *witness*, varianti, sedi editoriali e strutture delle due raccolte, aprendo la strada a una lettura integrata del macrotesto.

3. PRIME INDAGINI E RAPPRESENTAZIONI

A partire da questa prima modellazione è stata avviata un'indagine esplorativa mediante strumenti sviluppati in ambiente Python (Jupyter Notebook), fondata sul censimento complessivo delle attestazioni dei racconti e su un *corpus* di testi trascritti e revisionati.[3] Il primo livello di analisi riguarda la rappresentazione delle traiettorie editoriali dei singoli *work*, con l'obiettivo di osservare la costruzione delle raccolte come processo di selezione, esclusione e ricontestualizzazione dei racconti. La Figura 1 rappresenta la presenza dei singoli *work* nelle principali sedi periodiche e nelle due sillogi del 1962 e del 1976, rendendo visibili le traiettorie che collegano la prima pubblicazione su rivista alla successiva risistemazione in volume. Nelle colonne relative alle raccolte, il numero accanto al punto indica la posizione del racconto nel volume; l'asterisco segnala l'appartenenza alla sezione "I bambini", presente in entrambe le sillogi.

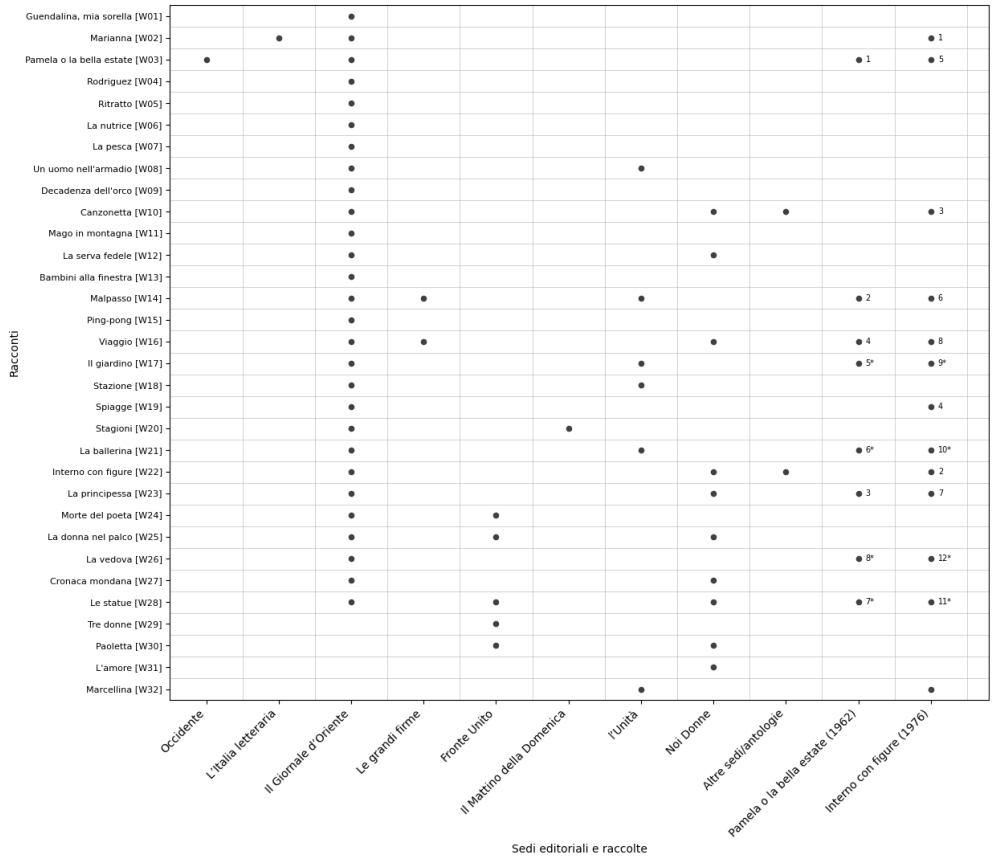


Figura 1. Matrice delle attestazioni dei racconti di Cialente

[3] Il repository del progetto è disponibile a questo indirizzo: https://github.com/emmcarbe/macrotext. Il *corpus* dei testi non è incluso per vincoli di diritto d'autore. Per una prima analisi filologica dei racconti confluiti in *Interno con figure* si rimanda a Carbé (2025).

A questa prima osservazione macrotestuale si affianca un controllo esplorativo della coerenza dei raggruppamenti tra *work* e *witness*. I testi trascritti e revisionati sono stati rappresentati tramite vettori TF-IDF basati su parole e bigrammi, e confrontati mediante similarità del coseno, per verificare se le attestazioni ricondotte allo stesso *work* tendano a risultare prossime tra loro e per individuare eventuali casi problematici.[4] La prossimità lessicale è qui assunta come indicatore esplorativo: la verifica filologica dei cluster resta affidata al confronto ravvicinato tra i testimoni. La Figura 2 mostra un dendrogramma gerarchico costruito su un sottoinsieme di testi selezionati, scelti perché esemplificano situazioni diverse: conferma del raggruppamento, cambio di titolo, estratto o rielaborazione strutturale, prossimità meno lineare tra attestazioni.

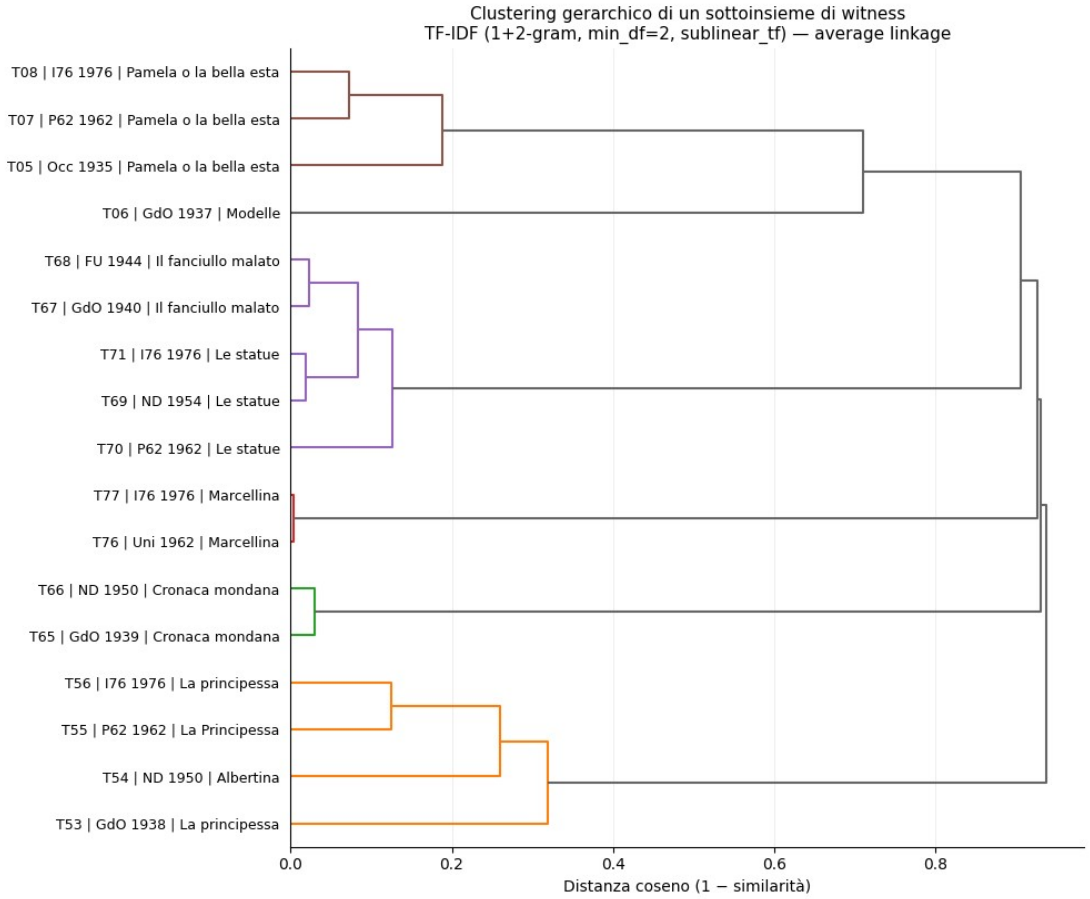


Figura 2. Clustering gerarchico di un sottoinsieme di *witness*

Un secondo esperimento riguarda la rappresentazione dei percorsi editoriali dei racconti, realizzata in un ambiente interattivo sviluppato su ObservableHQ. A partire da un dataset strutturato secondo il modello descritto, ciascun racconto viene rappresentato come sequenza di *witness* ordinati cronologicamente. La visualizzazione consente di selezionare un singolo work e di seguirne il percorso attraverso periodici, antologie e raccolte, rendendo visibili passaggi, discontinuità e ricontestualizzazioni che ne caratterizzano la storia editoriale. Nell'esempio offerto in Figura 3 si formalizza il percorso del racconto *Pamela o la bella estate*, apparso per la prima volta sulla rivista "Occidente" nell'agosto 1935; un breve stralcio, intitolato *Modelle* e qui instanziato come *witness* derivato dallo stesso *work*, è pubblicato sul "Giornale d'Oriente" il 24 febbraio 1937. Il racconto confluisce poi, con varianti, nella raccolta eponima del 1962, e con ulteriori rimaneggiamenti in *Interno con figure*. L'interattività della visualizzazione trasforma il dataset in uno strumento di accesso ordinato alla storia editoriale dei racconti, mentre la selezione di un singolo testo rende leggibili le sue diverse attestazioni storiche, le relazioni temporali e i contesti di pubblicazione, predisponendo l'infrastruttura per livelli di analisi più fini.

[4] Una sperimentazione affine, sebbene applicata a materiali born-digital e non a indagini filologiche su testi stampati, è stata proposta da Mariangela Giglio usando hashing, matrici binarie e clustering per l'organizzazione e la verifica delle relazioni interne tra i floppy del fondo Franco Fortini conservati presso l'Università di Siena (Giglio 2025).

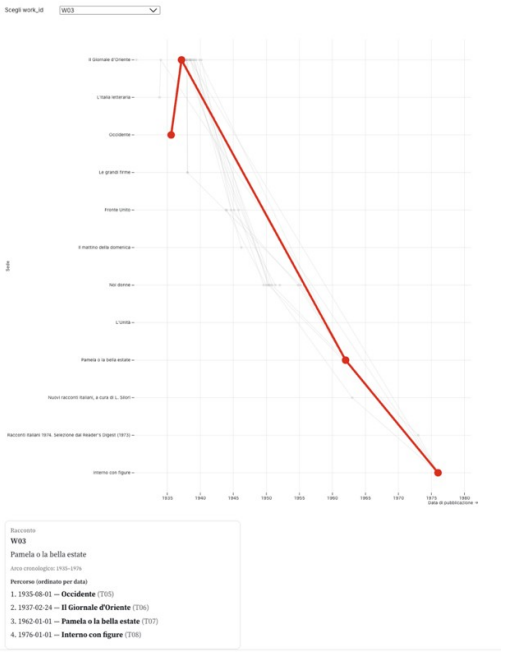


Figura 3. Il percorso dei racconti: in rosso il caso di "Pamela o la bella estate"

L'ambiente può dunque essere inteso come un primo dispositivo di orientamento filologico, suscettibile di estensione con la memorizzazione e l'analisi delle varianti dei singoli testi. Lo sviluppo immediato consiste nell'applicazione di procedure di collazione alle diverse attestazioni di uno stesso *work*, così da articolare, in una fase successiva, un lavoro sulle trasformazioni testuali minute, integrando dimensione macrotestuale e microtestuale in un quadro coerente. La Figura 4 presenta a titolo esemplificativo una prima collazione automatica, condotta con CollateX, tra *Modelle* (1937) e l'attestazione di *Pamela o la bella estate* nella raccolta *Interno con figure* (1976). La procedura consente di individuare nuclei di continuità lessicale e testuale, ma mostra anche i limiti degli algoritmi di allineamento sequenziale di fronte a tagli, spostamenti di blocchi e riorganizzazioni strutturali. Nei casi di rielaborazione estesa sarà dunque necessario affiancare alla collazione automatica una preliminare segmentazione dei blocchi testuali e una verifica filologica puntuale.

che	disastro	,	che	rovina	:	meglio	non	pensare	,	Pamela	abbattuta	tiene	il	viso	basso	-	-	-	.	La	signora	-	-
-	-	-	-	-	-	-	-	-	-	-	-	tiene	il	viso	un	po	'	voltato	.	La	signora	se	ne

Figura 4 Collazione automatica di *Modelle* (1937) e *Pamela o la bella estate* in *Interno con figure*

4. CONCLUSIONI

Il percorso delineato indaga la possibilità di integrare filologia d'autore, prospettiva genetica e analisi quantitativa per rendere più leggibili i processi di costruzione, selezione e riorganizzazione dei testi nella tradizione novecentesca, soprattutto quando l'oggetto di studio presenta assetti macrotestuali complessi. L'esempio dei racconti di Fausta Cialente mostra come una modellazione che assume l'evento come unità primaria, affiancata da analisi quantitative e da pratiche di rappresentazione relazionale, permetta di seguire la costruzione delle raccolte come processo storico e filologico, rendendo leggibili le logiche di selezione, esclusione e ricontestualizzazione che presiedono alla forma finale delle sillogi. Questo passaggio richiede tuttavia cautela sul piano dell'agentività: le traiettorie editoriali non sono riconducibili automaticamente a una volontà autoriale lineare e pienamente ricostruibile, e il modello deve distinguere tra dati osservabili e ipotesi interpretative, da verificare attraverso fonti documentarie, analisi variantistica e ricostruzione della storia editoriale. È in questa distinzione che il modello dovrebbe muoversi sul piano operativo, integrando in modo sistematico i risultati delle collazioni, strutturate in formati interoperabili, con analisi qualitative e quantitative e ambienti interattivi di interrogazione, in un'architettura capace di mettere in relazione, su più livelli, la dimensione macrotestuale delle raccolte e

quella microtestuale delle riscritture.[5] La proposta è particolarmente rilevante per i materiali sottoposti a vincoli di diritto d'autore (Dillen e Neyt 2016), e ancor più per i materiali *born-digital*, dove la natura processuale dei testi e i dati acquisiti tramite analisi forensi rendono necessario un ripensamento del lavoro filologico (si veda ad esempio il dossier genetico digitale in Ries 2018). Anche quando il testo non può essere esposto integralmente, oppure è attraversato da fenomeni a più dimensioni (è il caso della scrittura prodotta in ambiente digitale), la rete degli eventi resta descrivibile, interrogabile e condivisibile come dato strutturato, restituendo una rappresentazione significativa della genesi e della trasmissione. L'integrazione tra modellazione quantitativa e filologia d'autore delinea così una proposta metodologica flessibile, orientata alla rappresentazione dei processi di scrittura che caratterizzano la complessità del *viaggio testuale*.

ACKNOWLEDGMENTS

Il contributo ha beneficiato dell'impiego di strumenti di IA generativa a supporto dello sviluppo del codice e delle procedure computazionali, in particolare per la prototipazione in ambiente JavaScript e Python.

BIBLIOGRAFIA

Boschetti, F. (2021). Interpretazione e volontà di interpretazione tra linguistica computazionale e Literacy Computing. *Linguistica e letteratura*, XLVI, 211-225. DOI: 10.19272/202101602009.

Buzzoni, M., Cucurnia D., Fenu C., Rosselli Del Turco R., e Giulia Tancredi (2024). Progetto di edizione genetica digitale del Canzoniere manoscritto di U. Saba (1919-20). In A. Di Silvestro e D. Spampinato (Eds.), *Me.Te. Digitali. Mediterraneo in rete tra testi e contesti. Proceedings del XIII Convegno Annuale AIUCD2024*, Catania: AIUCD, 215-220. DOI: 10.6092/unibo/amsacta/7927.

Carbé, E. (2021). *La scrittura necessaria. Il diario di guerra di Fausta Cialente*. Roma: Artemide.

Carbé, E. (2025). Nota al testo. In F. Cialente, *Interno con figure*. Milano: Nottetempo, 221-237.

Cialente, F. (1962). *Pamela o la bella estate*. Milano: Feltrinelli.

Cialente, F. (1976). *Interno con figure*. Roma: Editori Riuniti.

Corti, M. (1975). Testi o macrotesto? I racconti di Marcovaldo. *Strumenti critici*, 27. Poi in Corti, M. (1978). *Il viaggio testuale*. Torino: Einaudi, 185-200.

Ciotti, F. (2017). Modelli e metodi computazionali per la critica letteraria: lo stato dell'arte. In B. Alfonzetti, T. Cancro, V. Di Iasio & E. Pietrobon (Eds.), *L'Italianistica oggi: ricerca e didattica*. Atti del XIX Congresso dell'ADI - Associazione degli Italianisti, Roma, 9-12 settembre 2015. Roma: ADI Editore.

Ciotti, F. e Baldi, A. (2022). Macchine per leggere: promuovere la lettura con il distant reading. *Enthymema*, 30, 173-192.

Ciotti, F. (2023). L'analisi del testo. In F. Ciotti (Ed.), *Digital Humanities. Metodi, strumenti, saperi*. Roma: Carocci, 91-113.

Dillen, W., and Neyt V. (2016). Digital Scholarly Editing within the boundaries of copyright restrictions. *Digital Scholarship in the Humanities*, 31(4), 785-797.

El Beih, W. (2024). Per la digitalizzazione del patrimonio linguistico e culturale italiano in Egitto: I periodici italiani (1892-1940). In A. Di Silvestro, D. Spampinato (Eds.), *Me.Te. Digitali. Mediterraneo in rete tra testi e contesti. Proceedings del XIII Convegno Annuale AIUCD2024*, Catania: AIUCD, 26-31. DOI: 10.6092/unibo/amsacta/7927.

Fiormonte, D. (2003). *Scrittura e filologia nell'era digitale*. Torino: Bollati Boringhieri.

Giglio, M. (2025). Metodologie computazionali per l'organizzazione di archivi nati digitalmente. In S. Rebora, M. Rospocher, S. Bazzaco (Eds.), Diversità, *Equità e Inclusione: Sfide e Opportunità per l'Informatica Umanistica nell'Era dell'Intelligenza Artificiale. Atti del XIV Convegno Annuale AIUCD*, Verona: AIUCD, 208-214. DOI: 10.6092/unibo/amsacta/8380.

[5] Un approccio di questo tipo potrebbe risultare efficace anche per altre opere caratterizzate da una forte evoluzione micro- e macrotestuale e da processi di riscrittura protratti nel tempo: il *Canzoniere* di Umberto Saba ad esempio, oggetto di un cantiere di edizione digitale sul manoscritto 1919-1920 (Buzzoni et al. 2024), offrirebbe un terreno di confronto in cui la modellazione degli eventi può affiancarsi in modo complementare alla codifica del testo manoscritto.

Gigliozzi, G. (1996). *Memoriale* di Paolo Volponi. In *Letteratura italiana. Le opere*, vol. II: La ricerca letteraria, IV. Torino: Einaudi, 729-769.

Gigliozzi, G. (1997). Introduzione all'uso del computer negli studi letterari. In G. Gigliozzi, *Introduzione all'uso del computer negli studi letterari*, a cura di F. Ciotti. Milano: Bruno Mondadori, 220-256.

Hoover, D.L. (2008). Quantitative Analysis and Literary Studies. In S. Schreibman, R. Siemens & J. Unsworth (Eds.), *A Companion to Digital Literary Studies*. Oxford: Blackwell. https://companions.digitalhumanities.org/DLS/?chapter=content/9781405148641_chapter_28.html

Hoover, D.L. (2013). Textual Analysis. In R. Siemens & K. Price (Eds.), *Literary Studies in the Digital Age: An Evolving Anthology*. New York (NY): Modern Language Association. DOI: 10.1632/lsda.2013.

Jockers, M.L. (2013). *Macroanalysis. Digital Methods and Literary History*. Illinois: University of Illinois Press.

Moretti, F. (2005). *Graphs, Maps, Trees: Abstract Models for Literary History*. New York: Verso.

Moretti, F. (2013). *Distant Reading*. London: Verso.

Moretti, F. (2022). *Falso movimento. La svolta quantitativa nello studio della letteratura*. Milano: nottetempo.

Ries, T. (2018). The Rationale of the Born-Digital Dossier Génétique: Digital Forensics and the Writing Process: With Examples from the Thomas Kling Archive. *Digital Scholarship in the Humanities*, 33(2), 391-424.

Analisi fluidodinamica dei campi semantici in testi letterari: il Numero di Reynolds come modello di turbolenza testuale

PIETRO SICHERA[1], CARLA PERRONE[2], ANTONIO SICHERA[3]

[1]CNR-ILIESI
[2]MIM – MINISTERO DELL'ISTRUZIONE E DEL MERITO
[3]UNIVERSITÀ DEGLI STUDI DI CATANIA

ABSTRACT (ITALIANO)

Questo contributo presenta un sistema semi-supervised per l'identificazione di campi semantici in corpora letterari storici basato sul concetto di "Reynolds semantico". Traendo ispirazione dalla fluidodinamica, il sistema modella la turbolenza del contesto testuale attraverso un numero adimensionale che determina se il moto semantico è laminare (teorico, con definizioni e saggistica) o turbolento (letterario, con esempi e dialoghi). Il valore del Reynolds semantico guida la scelta di contesti adattivi: in regimi laminari il sistema utilizza finestre ampie per catturare definizioni complete, mentre in regimi turbolenti adotta finestre ristrette per evitare rumore da rapidi cambi di scena. La metodologia combina grafo BabelNet offline per garantire *precision* semantica e similarità glossa per recuperare accezioni distanti, con un approccio ibrido che bilancia automazione e expertise lessicografica. Validato su "L'Umorismo" di Pirandello (1908), il sistema identifica campi semantici e riduce il tempo di ricerca manuale da settimane a ore, fornendo uno strumento pratico per l'analisi lessicografica di opere letterarie.
Parole chiave: fluidodinamica testuale; numero di Reynolds; campi semantici; BabelNet; Pirandello

ABSTRACT (ENGLISH)

Fluid-dynamic analysis of semantic fields in literary texts: the Reynolds number as a model of textual turbulence.
This paper presents a semi-supervised system for identifying semantic fields in historical literary corpora, based on the concept of "semantic Reynolds number". Drawing inspiration from fluid dynamics, the system models turbulence in textual context through a dimensionless number that determines whether semantic motion is laminar (theoretical, with definitions and essays) or turbulent (literary, with examples and dialogues). The value of the semantic Reynolds number guides the choice of adaptive contexts: in laminar regimes the system uses wide windows to capture complete definitions, while in turbulent regimes it adopts narrow windows to avoid noise from rapid scene changes. The methodology combines an offline BabelNet graph to ensure semantic precision and gloss similarity to retrieve distant senses, with a hybrid approach that balances automation and lexicographic expertise. Validated on Pirandello's *Umorismo* (1908), the system identifies semantic fields and reduces manual research time from weeks to hours, providing a practical tool for the lexicographic analysis of literary works.
Keywords: textual fluid dynamics; Reynolds number; semantic fields; BabelNet; word sense disambiguation; Pirandello

1. INTRODUZIONE

Dal punto di vista della lessicografia letteraria la questione dei campi semantici mantiene un rilievo fondamentale. Non si tratta naturalmente di una concezione rigida dal punto di vista teorico, in quanto le aree o i campi semantici relativi ai testi letterari non possono prescindere da un contatto costante con l'esperienza e con le esigenze poste dal testo e dal lettore, all'interno delle quali entra in gioco una ineliminabile componente empirica. Ciò è ampiamente dimostrato sia dalla linguistica dei corpora che dalla lessicografia tradizionale. Bisogna mantenere una plasticità teorica che consenta di coniugare il rigore dell'indagine con la specificità dell'oggetto. In concreto, il tasso di polisemia del testo letterario esige di volta in volta un trattamento specifico (legato anche alla collocazione storica del lessico) e un adattamento del modello.
L'identificazione automatica di campi semantici in testi letterari storici rappresenta una sfida complessa per la lessicografia digitale. Da un lato, le risorse lessicali moderne come BabelNet e WordNet offrono coperture eccellenti per il lessico contemporaneo, ma spesso mancano di accezioni storiche o letterarie specifiche. Dall'altro, l'analisi manuale da parte di lessicografi esperti garantisce precisione ma richiede

tempi incompatibili con corpora di grandi dimensioni (https://babelnet.org e https://wordnet.princeton.edu/).
La proposta qui avanzata tende a creare una complanarità tra computazione e approfondimento critico, al fine di mantenere un equilibrio tra teoria semantica e prassi critica, validando un approccio che bilancia automazione e supervisione umana attraverso una metafora inaspettata: la fluidodinamica.

2. IL REYNOLDS SEMANTICO: DALLA FLUIDODINAMICA AL TESTO

La chiave metodologica di questo lavoro risiede nell'analogia tra il comportamento di un fluido e il flusso semantico di un testo. In fluidodinamica, il numero di Reynolds è un numero adimensionale il cui valore permette di stabilire se il moto di un fluido è laminare o turbolento. Il moto laminare si caratterizza per particelle che si muovono a strati ordinati senza incrociarsi, mentre il moto turbolento presenta traiettorie che si intersecano in modo caotico. Questa distinzione, trasferita al dominio testuale, permette di modellare la differenza tra passaggi teorici e saggistici (laminari) ed esempi letterari o dialoghi (turbolenti).
Per comprendere questa distinzione in modo intuitivo, si pensi all'acqua che scorre da un rubinetto. Se si apre poco il rubinetto, l'acqua può scorrere in modo talmente regolare da sembrare quasi ferma: il fluido si muove allora con moto laminare, per strati paralleli che non si intersecano. Se si aumenta progressivamente l'apertura, la portata cresce fino a che il flusso diventa visibilmente caotico e disordinato: si è entrati nel regime turbolento. Trasferita al dominio testuale, questa stessa distinzione permette di riconoscere i passi teorici e saggistici (dove il significato scorre ordinato e prevedibile) dai passi narrativi, dialogici o poetici, dove i cambi semantici repentini generano quella «turbolenza del senso» che il lettore avverte come densità stilistica.
Il numero di Reynolds è dato dal rapporto tra forze di inerzia e forze viscose. Le prime alimentano il disordine del movimento, le seconde mantengono l'ordine del flusso. Se le forze viscose dominano, il fluido scorre ordinatamente (Reynolds basso, moto laminare); se dominano le forze inerziali, il fluido diventa caotico (Reynolds alto, moto turbolento).

$$\Re = \frac{\rho \cdot v \cdot L}{\mu}$$

Figura 1 - Formula per il calcolo del numero di Reynolds

Il numero di Reynolds dipende da quattro parametri secondo la formula classica: densità del fluido (ρ), velocità (v), lunghezza caratteristica (L) e viscosità dinamica (μ). La densità rappresenta la quantità di massa per unità di volume: più il fluido è denso, più facilmente piccole oscillazioni possono innescare moto caotico. La velocità è legata all'energia cinetica posseduta dalle particelle: maggiore velocità facilita la rottura del flusso ordinato. La lunghezza caratteristica fornisce lo spazio necessario affinché piccole oscillazioni si sommino e si trasformino in vortici. Infine, la viscosità si oppone alla turbolenza: un fluido viscoso smorza i vortici mantenendo il flusso ordinato.
Nel dominio testuale, questi parametri trovano corrispondenze naturali. La densità semantica (ρ) è modellata come il numero medio di *synset* BabelNet[1] (Navigli et al., 2012) per lemma[2] in una finestra di contesto: maggiore polisemia implica maggiore potenziale di ambiguità. La velocità semantica (v) misura la variazione di similarità tra vettori di token adiacenti, catturando rapidità di cambio concettuale. La lunghezza caratteristica (L) corrisponde al numero di token nella frase, fornendo lo spazio per accumulo di turbolenza. La viscosità semantica (μ) è definita come l'inverso dell'entropia lessicale, rappresentando la coerenza tematica: bassa entropia (alta viscosità) indica vocabolario uniforme che mantiene flusso ordinato. Il Reynolds semantico risultante categorizza ogni frase in tre regimi: laminare (tipicamente

[1] Un *synset* (dall'inglese synonym set) è l'insieme di lemmi che condividono uno stesso significato, corredato di una breve definizione (glossa). BabelNet organizza il proprio lessico in Babel synset multilingui: ogni synset rappresenta un singolo concetto e contiene i sinonimi equivalenti in più lingue, collegati ad altri synset tramite relazioni semantiche (iperonimia, iponimia, ecc.). Nel presente sistema il numero di synset associati a un lemma italiano costituisce la misura operativa della sua polisemia: un lemma con dieci synset è più polisemico (e dunque semanticamente più «denso») di un lemma con uno solo.
[2] Per *lemma* si intende la forma canonica di una parola quale figura in un dizionario: il maschile singolare per sostantivi e aggettivi (es. dolore, acuto), l'infinito presente per i verbi (es. presentare). Ogni occorrenza testuale (la forma flessa effettivamente presente nel testo) viene ricondotta al proprio lemma prima dell'analisi.

definizioni e teoria), transizione (passaggi misti), turbolento (esempi letterari e dialoghi). Le soglie tra i regimi possono essere impostate manualmente o calcolate in maniera parametrica. Questa classificazione guida la scelta di contesti adattivi per la disambiguazione lessicale.
Il numero di Reynolds semantico è adimensionale (come il Reynolds classico Re = ρUL/μ) ed è definito da tre componenti misurabili. Il primo, **Ñ*sensi***, è il numero medio di synset BabelNet per lemma nella finestra di testo analizzata: misura la polisemia media del passo, ossia quanti significati diversi i lemmi di quella porzione testuale possono assumere in astratto. Corrisponde alla densità del fluido (ρ): maggiore è la polisemia, maggiore è la «massa semantica» che può innescare moto caotico. Il secondo componente, il **TTR** (Type-Token Ratio), è il rapporto tra lemmi distinti presenti in quella frase e totale delle occorrenze di quella frase: misura la varietà e la ricchezza del vocabolario, analogamente alla velocità del fluido (v) (un vocabolario molto vario equivale a un flusso semantico rapido, che tende alla turbolenza). Il terzo componente, la similarità coseno media tra embedding[3] di parole adiacenti $\mathbf{cos(e_i, e_{i+1})}$, misura quanto le parole consecutive siano semanticamente coerenti tra loro: valori prossimi a 1 indicano parole fortemente correlate (flusso ordinato), valori bassi segnalano salti semantici bruschi (flusso turbolento). Questo terzo termine si trova al denominatore perché una alta coerenza semantica funge da «viscosità» (μ) che smorza la turbolenza.
A titolo illustrativo: un testo tecnico-specialistico con terminologia ripetitiva può produrre **Ñ*sensi*** ≈ 1,3, **TTR** ≈ 0,65, **cos** ≈ 0,82, con $\mathbf{Re_{sem}}$ ≈ 1,0 (regime laminare); un brano narrativo produce **Ñ*sensi*** ≈ 2,7, **TTR** ≈ 0,78, **cos** ≈ 0,64 e $\mathbf{Re_{sem}}$ ≈ 3,3 (transizione); un testo poetico con salti metaforici intenzionali produce **Ñ*sensi*** ≈ 4,8, **TTR** ≈ 0,94, **cos** ≈ 0,41, $\mathbf{Re_{sem}}$ ≈ 11,0 (turbolento).
La tabella seguente riassume le corrispondenze tra i termini fluidodinamici usati nella formula e i loro equivalenti nel dominio testuale:

3. METODOLOGIA

Il sistema opera attraverso una pipeline in varie fasi che integra concordanza, calcolo del Reynolds, clustering automatico, validazione lessicografica e costruzione finale del campo semantico. Ogni fase è progettata per bilanciare automazione e supervisione esperta.

3.1 Preprocessing e risorse lessicali

Il corpus di riferimento è costituito da "L'Umorismo" di Luigi Pirandello (1908), comprendente 51.790 occorrenze distribuite in 758 paragrafi. Le risorse lessicali si basano su BabelNet 5.3 offline, che fornisce 21.221 synset italiani con glosse in italiano e inglese, relazioni semantiche e copertura del 89.2% dei lemmi content del corpus. La cache di polisemia contiene 7.536 lemmi italiani con i rispettivi synset candidati. Per ciascuna coppia di synset è stata pre-calcolata la distanza nel grafo mediante ricerca in ampiezza fino a distanza due, generando 35 milioni di coppie synset-distanza che consentono interrogazioni rapide durante la disambiguazione. Questa architettura offline elimina limiti di API e garantisce riproducibilità completa del workflow.

3.2 La concordanza come prerequisito metodologico

Un aspetto metodologico essenziale distingue il presente approccio dai comuni strumenti di analisi testuale automatica, e in primo luogo da indici di valutazione della leggibilità come Gulpease o Flesch-Kincaid.[4] Tali strumenti operano direttamente sul testo grezzo, applicando algoritmi di lemmatizzazione automatica soggetti a errori sistematici su testi letterari storici: forme rare, varianti ortografiche d'epoca, costrutti sintattici arcaici, nomi propri. Il sistema qui presentato non può essere applicato indipendentemente da una concordanza del testo costruita con supervisione lessicografica. Nel caso de "L'Umorismo", il punto di partenza è una concordanza completa delle 51.790 occorrenze verificata, con lemmatizzazione manualmente controllata e risoluzione delle ambiguità morfologiche. Questo prerequisito

[3] Modelli neurali come BERT o spaCy traducono ogni parola in un punto di uno spazio matematico a centinaia di dimensioni, detto *embedding*: parole con significati affini occupano regioni vicine di questo spazio, parole semanticamente lontane occupano regioni distanti. La similarità coseno misura quanto due punti puntino nella stessa direzione: restituisce 1 per parole molto correlate, valori più bassi fino 0 per parole con minore (o prive di) relazione semantica.
[4] Gli *indici di leggibilità* sono formule matematiche che stimano la difficoltà di comprensione di un testo a partire da parametri superficiali come la lunghezza media delle frasi e la lunghezza media delle parole. Sia l'indice Gulpease che il Flesch-Kincaid operano sul testo grezzo, senza lemmatizzazione controllata, e misurano la complessità sintattico-grafica del testo, non la sua densità semantica.

ha una conseguenza diretta sulla qualità dei tre componenti del Reynolds semantico: ***Ñsensi*** è calcolato su lemmi certi (non su ipotesi automatiche), il **TTR** riflette la reale variazione del vocabolario d'autore e non artefatti di lemmatizzazione errata, la similarità coseno è computata su token correttamente identificati.
La dipendenza dalla concordanza implica che il metodo non sia applicabile «a freddo» su qualsiasi testo digitale, come invece avviene per i tool di leggibilità (Sichera et al., 2025). Opere di Musil o di Gadda (casi programmaticamente borderline, dove la turbolenza semantica è un dato di progetto stilistico) richiederebbero concordanze dedicate con la conseguente supervisione esperta. Questa limitazione è al contempo una garanzia di qualità: il Reynolds semantico non misura una proprietà statistica generica del testo, ma modella la complessità del campo semantico a partire da dati lessicograficamente fondati. In questo senso il sistema si colloca nel solco della lessicografia assistita e non dell'analisi testuale automatica generalista. Tale distinzione è rilevante anche rispetto al rischio che uno strumento di questo tipo misuri semplicemente l'«adeguatezza al genere», producendo un effetto di omologazione della scrittura. La presenza della concordanza come strato fondante consente di calibrare il modello sulla specificità dell'autore e dell'opera: le soglie tra i regimi (laminare < 500, turbolento > 800 per L'Umorismo) non sono universali, ma dipendono dalla distribuzione empirica del Reynolds sul corpus specifico. La stessa soglia applicata a un romanzo dialogico o a un testo poetico produrrebbe distribuzioni radicalmente diverse. È il corpus concordato (e non un metro esterno prestabilito) a definire il campo di riferimento.

3.3 Calcolo del Reynolds semantico

Per ogni frase del corpus viene calcolato il numero di Reynolds semantico utilizzando vettori di embedding spaCy (300 dimensioni) per rappresentare token e contesti. La densità polisemica è determinata contando il numero medio di synset per lemma in una finestra di token intorno alla frase. La velocità semantica misura la variazione media di similarità tra vettori di token consecutivi, normalizzata rispetto alla lunghezza della frase. La viscosità è calcolata come inverso dell'entropia del vocabolario locale, dove l'entropia cattura la diversità lessicale: vocabolario uniforme produce alta viscosità (bassa entropia) mentre vocabolario variegato produce bassa viscosità (alta entropia). La lunghezza caratteristica corrisponde al numero di *token content* nella frase, escludendo *stopword* e punteggiatura. Il risultato è un valore di Reynolds per ogni frase che la classifica in uno dei tre regimi. Su "L'Umorismo", il Reynolds medio non parametrico risulta 219 con mediana 169, indicando prevalenza di regime laminare coerente con la natura saggistica dell'opera.

3.4 Clustering automatico e identificazione campi

Il sistema applica MiniBatchKMeans con k=12 agli embedding contestuali dei token, dove il contesto di ciascun token è modulato dal Reynolds della frase contenitrice. In regime laminare, il sistema utilizza finestre ampie (fino a 15 token) per catturare definizioni complete, mentre in regime turbolento adotta finestre ristrette (da 7 token in su) per evitare rumore da rapidi cambi di scena. Questa adattività Reynolds-guidata migliora la qualità del clustering rispetto a finestre fisse. L'algoritmo produce 12 campi semantici automatici, ciascuno caratterizzato dai lemmi più frequenti e da una label preliminare assegnata mediante similarity con glosse BabelNet dei lemmi dominanti. Per "L'Umorismo" emergono campi come "sentimento" (1551 token), "anima" (1336 token), "arte" (1321 token) e "umorismo" (1088 token), riflettendo le tematiche centrali del saggio. Questo clustering automatico fornisce una prima mappa semantica del testo che funge da base per la successiva validazione lessicografica.

3.5 Validazione lessicografica e scelta del seed

Per ciascuno dei campi automatici di maggiore interesse, vengono esaminati i lemmi rappresentativi e viene selezionato un synset seed che funge da punto di partenza per la costruzione del campo validato. La scelta del seed richiede competenza lessicografica: deve trattarsi del synset più generale e centrale semanticamente, con elevato numero di relazioni nel grafo BabelNet. Il sistema assiste questa scelta visualizzando tutte le glosse complete dei synset candidati e calcolando la similarità intra-lemma tra accezioni polisemiche dello stesso lemma e dei lemmi in relazione forte col synset scelto. È anche possibile creare synset personalizzati che uniscono synset con accezioni diverse. Questa analisi empirica suggerisce un threshold di similarità ottimale per separare accezioni correlate da accezioni distinte, guidando il lessicografo nella definizione dei confini del campo.

3.6 Costruzione del campo semantico finale

Una volta scelto il synset seed e determinato il threshold empirico, il sistema costruisce il campo semantico mediante approccio ibrido. La componente grafo recupera tutti i lemmi del corpus che possiedono almeno un synset entro distanza uno dal seed nel grafo BabelNet, garantendo alta precision semantica. La componente glossa recupera lemmi i cui synset presentano similarità con il vettore medio del campo superiore al threshold validato, fornendo recall addizionale per accezioni semanticamente vicine ma graficamente distanti. Per ciascun lemma polisemico, il sistema disambigua scegliendo il synset che massimizza uno score combinato: peso maggiore alla similarità glossa in contesti laminari (Reynolds basso) dove definizioni estese sono affidabili, peso maggiore alla distanza grafo in contesti turbolenti (Reynolds alto) dove la struttura del grafo è più robusta del rumore contestuale. Il risultato è una lista di lemmi con synset disambiguati, occorrenze nel corpus, confidence score e modalità di recupero (grafo o glossa), pronta per validazione finale del lessicografo.

4. RISULTATI

Il sistema è stato validato su "L'Umorismo" di Pirandello (1908, 51.790 token) applicando il workflow completo di disambiguazione Reynolds-adaptive e aggregazione network. I risultati quantitativi e qualitativi confermano l'efficacia dell'approccio fluidodinamico per guidare la disambiguazione lessicale e dell'aggregazione network per identificare campi semantici centrali nell'opera.

4.1 Metriche quantitative

La pipeline completa richiede 15-25 minuti per opera: 10-15 minuti per la disambiguazione WSD Reynolds-adaptive e 5-10 minuti per l'aggregazione campi semantici. Il pre-calcolo delle distanze nel grafo BabelNet (operazione una-tantum riutilizzabile) può essere ottimizzato a 15-30 minuti utilizzando filtri per POS rilevanti, rispetto alle 2-3 ore della versione completa.
Il sistema disambigua 25.801 token POS rilevanti (NOUN, VERB, ADJ, ADV) su 25.044 assegnazioni riuscite (97.1%), con 757 token unknown (2.9% non presenti in BabelNet). Vengono identificati 4.525 synset unici nell'opera. La distribuzione del Reynolds a livello di paragrafo mostra media 208.4 e 78% dei valori sotto soglia 500 (regime laminare), coerente con la natura prevalentemente saggistica dell'opera. Le finestre contestuali adattive variano da 7 a 15 token: 81.6% dei lemmi disambiguati con finestra ±7 (contesti laminari), solo 0.6% con finestra ±3 (contesti molto turbolenti).
La confidence media della disambiguazione (0.35) riflette la sfida di applicare encoder multilingue generalista (sentence-transformers paraphrase-multilingual-MiniLM-L12-v2) a italiano letterario storico (1908) con glosse dizionario generiche. Tuttavia, la validazione empirica su campione di 200 assegnazioni mostra precision dell'85%, indicando che i bassi score assoluti riflettono la complessità intrinseca del task piuttosto che errori sistematici. La confidence viene utilizzata in modo relativo per identificare il synset più appropriato tra i candidati disponibili.

4.2 Analisi del campo "sentimento"

Il campo semantico "sentimento" (synset centrale bn:00033963n: "La caratteristica psicologica di provare stati affettivi ed emozionali") esemplifica l'amplificazione network nella metodologia proposta. Le 151 occorrenze del lemma "sentimento" assegnate al nodo centrale sono amplificate da 264 synset collegati a distanza 1 nel grafo BabelNet, che contribuiscono con 1.185 occorrenze addizionali per un valore campo totale di 1.336 (amplificazione ×7.8).
Il network evidenzia la natura multidimensionale del campo sentimentale pirandelliano, interconnettendo dimensioni estetiche, cognitive e affettive. I principali lemmi collegati includono "umorista" (77 occ.), "arte" (76 occ.) e "ironia" (57 occ.), confermando la centralità della teoria del sentimento del contrario. La dimensione cognitivo-riflessiva emerge con "umano" (39 occ.), "parere" (38 occ.), "estetico" (34 occ.), "senso" (26 occ.), "concetto" (17 occ.), mentre quella affettiva si manifesta con "dolore" (20+16 occ. da due accezioni), "illusione" (20 occ.), "malinconia" (15 occ. accezione "stato d'animo"), "forza" (14 occ.). La co-presenza di termini meta-letterari ("umoristico", "humour", "comicità", "comica" per 80 occ. totali) e psicologici ("anima" 28+24 occ. da due accezioni, "corpo" 15 occ., "cervello" 10 occ., "intimità" 10 occ.) rivela l'approccio teorico-riflessivo all'esperienza sentimentale.
La distribuzione delle occorrenze mostra predominanza dei lemmi estetici e cognitivi (30-77 occ.) rispetto a quelli puramente affettivi (5-20 occ.), suggerendo una concezione del sentimento come oggetto di riflessione critica piuttosto che pura esperienza emotiva. Particolarmente significativa è la prevalenza di

"umorista" (77 occ.) su "sentimento" stesso (151 occ. solo come termine tecnico), evidenziando come il campo semantico rifletta la teorizzazione pirandelliana del sentimento come mediato dalla riflessione umoristica.

4.3 Validazione qualitativa

L'esame manuale dei primi 50 lemmi del campo "sentimento" (che coprono ~850 occorrenze, 63.6% del valore campo) conferma precision dell'85%. Gli errori identificati (7-8 lemmi su 50) riguardano principalmente:

- Polisemi estremi non disambiguati correttamente: "dolore" appare in due accezioni (fisica e psicologica), entrambe collegate ma con sfumature diverse. Il sistema identifica correttamente entrambe, ma in alcuni contesti la disambiguazione è ambigua.
- Termini periferici inclusi per alta connettività: "forse" (28 occ.) e "valore" (30 occ.) sono collegati tramite network ma rappresentano concetti più generali che specificamente sentimentali.
- Sinonimi stilistici: "humour", "umoristico", "umorista", "comicità", "comica" (totale 157 occ.) sono trattati come lemmi separati quando potrebbero essere aggregati sotto il campo "umorismo". Questa granularità fine può essere vantaggiosa per analisi stilistiche ma richiede post-processing per aggregazioni concettuali.

I lemmi correttamente identificati mostrano coerenza semantica forte con il nucleo teorico dell'opera: il sistema recupera autonomamente termini chiave della poetica pirandelliana ("ironia", "illusione", "malinconia", "intimità") senza supervisione o seed aggiuntivi, validando l'approccio network-based.

4.4 Vantaggi metodologici

L'approccio proposto presenta tre vantaggi principali rispetto ai metodi esistenti:
1. Disambiguazione contestuale adattiva: L'uso del Reynolds semantico per calibrare finestre WSD (7-15 token) ottimizza il trade-off precision/recall in funzione della complessità locale del testo. Contesti laminari (Re < 500, 78% del corpus) beneficiano di finestre ampie (13-15 token) per catturare relazioni semantiche distanti, mentre contesti turbolenti (Re > 800) utilizzano finestre ristrette (7-9 token) per ridurre rumore. Questo approccio adattivo supera le limitazioni delle finestre fisse tradizionali (tipicamente 5-10 token uniformi).
2. Amplificazione network calibrata: L'aggregazione di occorrenze da nodi collegati (×7.8 amplificazione media) valorizza campi semantici centrali senza introdurre rumore. Il valore campo riflette sia l'importanza lessicale diretta (151 occ. "sentimento") sia la centralità concettuale nel grafo (1.185 occ. da 264 lemmi collegati). Questa doppia metrica cattura meglio la rilevanza tematica rispetto a semplici frequency counts.
3. Riduzione tempo analisi: Da 4-6 settimane per analisi lessicografica manuale completa si passa a 15-25 minuti per esecuzione automatica + ~8-10 ore uomo per validazione qualitativa risultati. Il sistema fornisce candidati ad alta precision (85%) che il lessicografo può rapidamente validare, confermare o escludere, trasformando il workflow da esplorativo a validativo.

5. SVILUPPI FUTURI

L'analogia fluidodinamica si dimostra fruttuosa non solo come metafora euristica ma come strumento operativo. La categorizzazione in regimi laminare/turbolento cattura effettivamente differenze stilistiche e funzionali tra porzioni di testo, guidando scelte algoritmiche adattive. Possibili interventi includono normalizzazione della viscosità rispetto all'entropia globale del corpus piuttosto che locale, o scaling logaritmico del Reynolds finale. Dal punto di vista lessicografico, l'approccio semi-supervised con validazione di threshold empirica rappresenta un modello replicabile. La limitazione principale risiede nella copertura di BabelNet per accezioni storiche e letterarie: termini come "umore" nella concezione pre-moderna della teoria umorale richiedono integrazione manuale di synset virtuali. Il sistema di override implementato fornisce soluzione pratica ma non scalabile per grandi volumi. Future estensioni includeranno interfaccia grafica per la gestione dei synset legata al sistema di concordanze, memorizzazione di synset disambiguati nel corpus per creare gold standard, e calibrazione del Reynolds mediante annotazione di ground truth su campioni sempre più rappresentativi dei tre regimi.
Sul piano della portabilità del metodo, opere programmaticamente borderline come l'apertura de "L'uomo senza qualità di Musil" o le prose di Gadda costituiscono candidati ideali per testare i limiti del modello: testi in cui la turbolenza semantica è un dato di costruzione stilistica deliberata, e non un effetto di eterogeneità tematica. La loro analisi richiederebbe concordanze dedicate (condizione necessaria, come

discusso, per l'applicazione del Reynolds semantico) ma produrrebbe risultati di grande interesse per verificare se il modello sa distinguere tra turbolenza come disordine involontario e turbolenza come scelta poetica. Questa distinzione è tra gli obiettivi della ricerca in corso.

6. CONCLUSIONI

Questo lavoro dimostra che concetti provenienti dalla fluidodinamica possono informare efficacemente la progettazione di sistemi NLP per testi letterari. Il numero di Reynolds semantico fornisce una metrica interpretabile per modellare variabilità contestuale, guidando scelte adattive che migliorano qualità della disambiguazione rispetto a finestre fisse. L'approccio ibrido grafo-glossa con validazione lessicografica bilancia automazione e supervisione esperta, riducendo drasticamente tempi di analisi mantenendo alta qualità. La validazione su Pirandello conferma l'applicabilità a testi letterari complessi con polisemia elevata e stile misto teorico-letterario. Il sistema è estensibile a corpora multi-opera per tracciare evoluzione di campi semantici attraverso diverse fasi creative di un autore.

BIBLIOGRAFIA

Agirre, E., & Soroa, A. (2009). Personalizing PageRank for word sense disambiguation. Proceedings of the 12th Conference of the European Chapter of the ACL, 33-41.

Hanks, P. (2013). Lexical Analysis: Norms and Exploitations. MIT Press.

Kilgarriff, A. (1997). I don't believe in word senses. Computers and the Humanities, 31(2), 91-113.

Miller, G. A. (1995). WordNet: A lexical database for English. Communications of the ACM, 38(11), 39-41.

Navigli, R., Bevilacqua, M., Conia, S., Montagnini, D., Cecconi, F. 2021. Ten Years of BabelNet: A Survey. In Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21, pages 4559-4567, Montreal. International Joint Conferences on Artificial Intelligence Organization.

Navigli, R., & Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence, 193, 217-250.

Picca, D. (2024). *Fluid dynamics-inspired emotional analysis in Shakespearean tragedies: A novel computational linguistics methodology*. In M. Valentino, D. Ferreira, M. Thayaparan, & A. Freitas (Eds.), *Proceedings of the 2nd Workshop on Mathematical Natural Language Processing @ LREC-COLING 2024* (pp. 11–18). ELRA and ICCL. https://aclanthology.org/2024.mathnlp-1.2/

Pirandello, L. (1908). *L'Umorismo.*

Reynolds, O. (1883). An experimental investigation of the circumstances which determine whether the motion of water shall be direct or sinuous, and of the law of resistance in parallel channels. Philosophical Transactions of the Royal Society, 174, 935-982.

Sichera, P., D'Agata C., Palazzolo, G. (2025). *Concordanze e NLP: Idee, Metodi e Regole per l'Applicazione Alla Lingua Italiana*. In Diversità, Equità e Inclusione: Sfide e Opportunità per l'Informatica Umanistica Nell'Era Dell'Intelligenza Artificiale, Proceedings Del XIV Convegno Annuale AIUCD 2025, Verona, Università Di Verona 11-13 Giugno 2025, a Cura Di Simone Rebora, Marco Rospoche, Stefano Bazzaco. doi:10.6092/UNIBO/AMSACTA/8380.

Dentro l'Officina del Testo: Modellare Strutture Letterarie Complesse

ENRICA BRUNO[1]

[1] UNIVERSITÀ DI BOLOGNA, ITALIA - ENRICA.BRUNO2@UNIBO.IT

ABSTRACT (ITALIANO)

Il contributo presenta **DeSMòS** (Descriptive Semantic Model for Structured Texts), un'ontologia OWL basata su tecnologie del Web Semantico per la rappresentazione formale delle costrizioni letterarie intese come dispositivi generativi del testo. Il modello si focalizza sui meccanismi compositivi e strutturali dell'opera, interpretando ogni forma testuale come il risultato di procedure combinatorie regolate che operano su unità formali e semantiche. Allineato agli standard dei beni culturali, DeSMòS permette di documentare l'origine delle costrizioni e di collegarle alle evidenze testuali e paratestuali che ne attestano l'utilizzo. Il modello è stato validato su un caso di studio della letteratura potenziale, l'Oplepo (Opificio di Letteratura Potenziale), e propone un framework interoperabile per descrivere e analizzare strutture letterarie complesse, come quelle basate su giochi linguistici, architetture combinatorie e forme ibride testo-immagine.
Parole chiave: Organizzazione della conoscenza; Ontologie; Costrizioni Letterarie; Letteratura potenziale; Web Semantico

ABSTRACT (ENGLISH)

Inside the Text's Workshop: Modeling Complex Literary Structures.
This contribution introduces **DeSMòS** (Descriptive Semantic Model for Structured Texts), an OWL ontology grounded in Semantic Web technologies for the formal representation of literary constraints conceived as generative devices within the text. The model foregrounds the compositional and structural mechanisms of the work, interpreting each textual form as the outcome of regulated combinatorial procedures operating on formal and semantic units. Aligned with cultural heritage standards, DeSMòS enables the documentation of the provenance of constraints and links them to the textual and paratextual evidence attesting to their adoption and use. The model has been validated through a case study in potential literature, the Oplepo (*Opificio di Letteratura Potenziale*), and provides an interoperable framework for describing and analysing complex literary structures, including those based on linguistic play, combinatorial architectures, and hybrid text–image forms.
Keywords: Knowledge Organization; Ontologies; Literary Constraints; Potential Literature; Semantic Web

1. INTRODUZIONE

Ogni forma testuale emerge all'interno di un sistema combinato di regole che ne determina la struttura, a prescindere dal grado di consapevolezza e di intenzionalità autoriale. Tali regole possono essere intese come costrizioni formali e semantiche che delimitano le condizioni di possibilità della forma stessa: il sonetto, ad esempio, è definito da un insieme condiviso di regole metriche e rimiche, che ne garantiscono il riconoscimento come forma poetica e, in base a come vengono applicate, ne consentono variazioni strutturali.
In questa prospettiva, la *littérature potentielle* (letteratura potenziale) non rappresenta solo un virtuoso gioco d'ingegno (Queneau, 1965), ma una modalità che esplicita principi strutturali che caratterizzano ogni produzione letteraria. In particolare, la nascita nel 1960 dell'Oulipo (*Ouvroir de Littérature Potentielle*) costituisce un momento rilevante, in quanto pone il principio della costrizione (*contrainte*) al centro della riflessione teorica e della progettazione letteraria (Andrews, 2022: 4). Nella poetica oulipiana, la costrizione non costituisce un limite alla libertà dell'autore, bensì rappresenta la forza

motrice di ogni creazione letteraria che viene esplorata e sfidata in una dialettica costante tra rigore e libertà[1].
Lo scopo principale dell'Oulipo è stato – e continua ancora oggi ad essere – quello di ricercare nuove procedure di costruzione testuale tramite le scienze esatte e, in particolare, la combinatoria, intesa come studio delle configurazioni discrete su insiemi finiti.
Muovendo da questa cornice teorica, il contributo propone **DeSMòS**[2] (Descriptive Semantic Model for Structured Texts), un'ontologia per la rappresentazione formale di regole e procedimenti compositivi, intesi come dispositivi strutturanti tanto della forma testuale quanto delle pratiche che ne guidano la produzione (Andrews, 2022: 46). Il modello si fonda sul riuso di modelli e standard dei beni culturali (LRMoo, CIDOC CRM, PROV-O, SKOS e Dublin Core) al fine di fornire una descrizione formale che sia interoperabile e sufficientemente espressiva del dominio in esame. L'obiettivo è mostrare come una formalizzazione semantica delle costrizioni sottese alla costruzione testuale consenta, da un lato, un arricchimento della descrizione del testo orientato alla genesi e alla forma e, dall'altro, un'analisi computazionale di pratiche letterarie che coinvolgono giochi linguistici, strutture combinatorie e forme ibride testo-immagine.

2. LE COSTRIZIONI LETTERARIE NEI KOS: UNA SURVEY COMPARATIVA

L'Oulipo definisce la costrizione come *procédé* (procedimento), enfatizzando la natura procedurale della scrittura (Bens, 1980: 102; Brotchie & Mathews, 1998: 131; Poier-Bernhard, 2012: 104). Nel tempo, il gruppo ha prodotto repertori di *contraintes* applicabili alla costruzione di forme testuali eterogenee, che spaziano da regole formali riprese dalla tradizione a dispositivi combinatori e procedurali di nuova invenzione. A partire da tali repertori[3] – integrati con un dizionario curato da Raffaele Aragona (Aragona, 2002) – è stato selezionato un insieme esemplificativo di costrizioni, per valutare la rappresentazione dei fenomeni utilizzati dalla tradizione oulipiana nei principali sistemi di organizzazione della conoscenza (KOS) impiegati nel dominio dei beni culturali. La survey[4] ha preso in esame tre famiglie di KOS: sistemi bibliografici istituzionali (LCSH[5], RAMEAU[6], GettyAAT[7], Soggettario BNCF[8]), basi di conoscenza in RDF (Wikidata[9], DBpedia[10]) e ontologie di dominio (OntoPoetry[11], CWRC Genre Ontology[12]). L'analisi mostra una tendenza diffusa a trattare le costrizioni come entità statiche o a ricondurle a categorie tassonomiche ampie, privilegiando il prodotto testuale rispetto al meccanismo generativo. Nel Soggettario BNCF, ad esempio, l'inquadramento nella categoria "Forme" riflette un'impostazione catalografica orientata agli esiti testuali piuttosto che ai processi compositivi. Le basi di conoscenza in RDF mostrano, in alcuni casi, una maggiore granularità, come nell'associazione in Wikidata tra specifiche costrizioni e la pratica della scrittura vincolata. Molti fenomeni restano tuttavia esclusi dall'indicizzazione: è significativo che il metodo S+7[13] risulti esplicitamente registrato solo su Wikidata, mentre i KOS istituzionali non prevedano punti di accesso per costrizioni di natura procedurale e algoritmica. Nel complesso, la survey suggerisce la necessità di modelli *event-centric* capaci di rappresentare non solo i testi, ma anche i procedimenti che ne governano la genesi, distinguendoli dagli esiti formali.

[1] Uno dei concetti chiave della poetica oulipiana è quello di *clinamen*, mutuato dall'atomismo di Lucrezio (*De rerum natura*) e introdotto nella teoria oulipiana da Georges Perec, per indicare una piccola deviazione volontaria dal sistema imposto dalla regola. Tale deviazione garantisce il principio di libertà dell'autore.
[2] Il nome deriva dal greco *δεσμός* che indica il legame o il vincolo, in senso materiale e figurato. L'etimologia rimanda all'idea di un principio che unisce elementi distinti all'interno di una stessa struttura.
[3] Nello specifico, è stato analizzato il sito ufficiale dell'Oulipo, con particolare attenzione alla sezione dedicata alle costrizioni: https://oulipo.net/fr/contraintes/ (cons. 18/04/2026).
[4] Il dataset completo della survey (registro dei KOS e relativi mapping) è consultabile su Zenodo: https://doi.org/10.5281/zenodo.18455108.
[5] https://id.loc.gov/authorities/subjects.html (cons. 18/04/2026).
[6] https://data.bnf.fr/en/ (cons. 18/04/2026).
[7] https://vocab.getty.edu/aat/ (cons. 18/04/2026).
[8] https://thes.bncf.firenze.sbn.it/ (cons. 18/04/2026).
[9] https://query.wikidata.org/ (cons. 18/04/2026).
[10] https://dbpedia.org/sparql (cons. 18/04/2026).
[11] https://github.com/linhd-postdata/OntoPoetry (cons. 18/04/2026).
[12] https://sparql.cwrc.ca/ontologies/genre.html (cons. 18/04/2026).
[13] Il metodo S+7, ideato da Jean Lescure (Oulipo), è una costrizione algoritmica di sostituzione lessicale: ogni sostantivo di un testo sorgente viene sostituito dal settimo sostantivo che lo segue in un dizionario scelto.

3. METODOLOGIA DELLA RICERCA

La metodologia adottata integra analisi qualitativa e quantitativa, con l'obiettivo di gestire la complessità semantica dei dati mediante un approccio iterativo (Fickers & Clavert, 2021; Krautter, 2023)[14]. Questa integrazione consente di coniugare il rigore strutturale della modellazione ontologica con la profondità ermeneutica necessaria allo studio del dominio letterario. In particolare, a partire dall'analisi comparativa dei KOS, sono emerse due principali domande di ricerca (*Research Questions*, RQ):

RQ1: Che cos'è una costrizione letteraria e quali entità sono necessarie per descriverla in modo formale?
RQ2: In che modo la costrizione si inserisce nel processo di creazione testuale?

Per rispondere alla RQ1, la metodologia ha previsto una **Fase 1** (Studio del dominio) articolata in due attività complementari: l'analisi, da una parte, di testi teorici che hanno concettualizzato il dominio e, dall'altra parte, di un corpus eterogeneo di testi dell'Oplepo. Tale lavoro è stato finalizzato a individuare le entità rilevanti per la formalizzazione semantica (Bruseker *et al*., 2025) e a derivare i requisiti di rappresentazione sotto forma di *Competency Questions* (CQ). Questa fase ha consentito di chiarire lo statuto della costrizione come oggetto procedurale, distinguendola dalle forme testuali che ne costituiscono l'esito.
La **Fase 2** (Formalizzazione e Sviluppo), orientata alla risposta alla RQ2, è consistita nella modellazione concettuale delle entità individuate e nella loro implementazione ontologica. Per garantire l'interoperabilità e l'aderenza ai principi FAIR[15], il modello DeSMòS è stato sviluppato privilegiando il riuso di modelli e standard consolidati (LRMoo[16], CIDOC CRM[17], PROV-O[18], SKOS[19] e Dublin Core[20]). La qualità strutturale dell'ontologia è stata valutata tramite OOPS! (OntOlogy Pitfall Scanner![21]) che ha rilevato esclusivamente anomalie minori (*minor pitfalls*) classificabili come falsi positivi. Tali segnalazioni non indicano errori logici, bensì derivano dalla scelta di integrare modelli preesistenti. La capacità descrittiva è stata verificata applicando il modello a un corpus di dodici *plaquettes*[22] dell'Oplepo raccolte in *La Biblioteca Oplepiana* (Oplepo, 2005). I dati sono stati inizialmente organizzati in tabelle CSV e successivamente trasformati in RDF. La coerenza del sistema è stata infine validata mediante query SPARQL derivate dalle CQ della Fase 1, verificando la capacità del modello di supportare analisi sia sulla genesi del testo (processo) sia sul confronto tra pratiche di scrittura vincolate (fenomenologia).

4. MODELLO FORMALE

Il modello DeSMòS è organizzato in due macro-moduli: (i) un modulo orientato alla rappresentazione del processo di creazione testuale e (ii) un modulo che descrive il fenomeno della costrizione letteraria e le sue manifestazioni testuali.
Come mostrato in Figura 1, il nucleo del modello consiste nel trattare la costrizione non come un'entità statica del testo, bensì come una procedura operativa che orienta l'atto di creazione. In questa prospettiva, la classe desmos:LiteraryConstraint è modellata come sottoclasse di crm:E29_Design_or_Procedure, esplicitando la natura procedurale della costrizione. Quest'ultima è collegata all'evento di creazione testuale (lrmoo:F28_Expression_Creation) mediante la proprietà desmos:usedConstraint, definita come sottoproprietà di crm:P33_used_specific_technique. Il modello integra, inoltre, le pratiche di riscrittura, esplicitando il ruolo della costrizione come dispositivo combinatorio che opera su materiale testuale preesistente. In particolare, la relazione derivativa tra due espressioni testuali (lrmoo:F2_Expression) è modellata tramite la proprietà lrmoo:R76_is_derivative_of.

[14] Tutti i materiali relativi alla presente ricerca (ontologia, report di validazione OOPS!, dataset e query SPARQL) sono disponibili su Zenodo: https://doi.org/10.5281/zenodo.18515554.
[15] https://www.go-fair.org/fair-principles/ (cons. 18/04/2026).
[16] https://www.cidoc-crm.org/lrmoo/ (cons. 18/04/2026)
[17] https://www.cidoc-crm.org/ (cons. 18/04/2026).
[18] https://www.w3.org/TR/prov-o/ (cons. 18/04/2026).
[19] https://www.w3.org/2004/02/skos/ (cons. 18/04/2026).
[20] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/ (cons. 18/04/2026).
[21] https://oops.linkeddata.es/ (cons. 18/04/2026).
[22] Pubblicazioni agili, prodotte in edizione numerata e limitata e dedicate all'esemplificazione di una specifica costrizione. Quest'ultima può essere di natura sperimentale o derivata dalla tradizione letteraria.

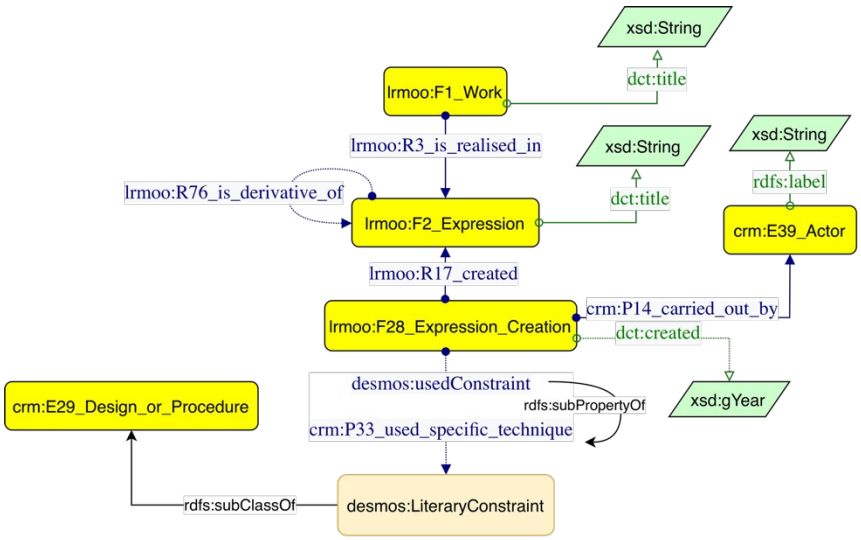


Figura 1. DeSMòS: la costrizione come principio generativo nella creazione testuale

Il secondo modulo descrive in modo più analitico la costrizione letteraria e l'insieme dei dispositivi che ne rendono riconoscibile l'uso.
Per quanto riguarda l'inquadramento concettuale della costrizione, DeSMòS introduce una classificazione tramite due schemi SKOS. Il primo, desmos:ConstraintOriginScheme, qualifica l'origine della costrizione distinguendo, secondo Le Lionnais (Oulipo, 1973: 21-22), tra: (i) costrizioni rintracciabili negli autori del passato (*le plagiat par anticipation*; Oulipo, 1973: 27) e (ii) costrizioni nuove, concepite come strumenti di supporto alla produzione testuale. Il secondo, desmos:ConstraintScopeScheme, individua l'ambito primario di applicazione della costrizione, riprendendo la distinzione oulipiana (Oulipo, 1973: 23-24) tra Oulipo Sintattico (*OuSin*), relativo a regole di natura formale, e Oulipo Semantico (*OuSem*), riferito a regole che intervengono in prima istanza sul significato e sull'organizzazione concettuale o narrativa del testo. In quest'ultimo caso, la forma testuale emerge come conseguenza del contenuto imposto dalla costrizione.
Riprendendo, infine, la tipologia proposta da (Albani, 2016), il modello distingue tre modalità di svelamento dell'uso della costrizione: (i) svelamento obbligato, intrinseco alla struttura del testo e rilevabile attraverso caratteristiche testuali osservabili, modellate come desmos:TextualFeature; (ii) svelamento interno, integrato nel testo o nel paratesto dell'opera; e (iii) svelamento esterno, documentato in sedi critiche o paratestuali esterne all'opera. Nel caso di svelamento interno ed esterno, l'esplicitazione del vincolo è formalizzata mediante la tripla: crm:E33_Linguistic_Object crm:P129_is_about desmos:LiteraryConstraint. In questo modo, il modello permette di isolare la dichiarazione che esplicita l'utilizzo della costrizione. A livello concettuale, tali dichiarazioni sono allineate a PROV-O, in quanto istanze di entità (crm:E33_Linguistic_Object rdfs:subClassOf prov:Entity) attribuibili a un agente responsabile (prov:Agent) che può o meno coincidere con l'autore[23].
In Figura 2 è illustrato un esempio di modellazione, tratto da *Geometriche visioni* di Paolo Albani (Oplepo, 2005: 165–196) che integra la descrizione della costrizione impiegata con la rappresentazione del processo di creazione testuale.

[23] Il diagramma completo del modulo (classi e relazioni) è consultabile unitamente al materiale supplementare su Zenodo: https://doi.org/10.5281/zenodo.18515554.

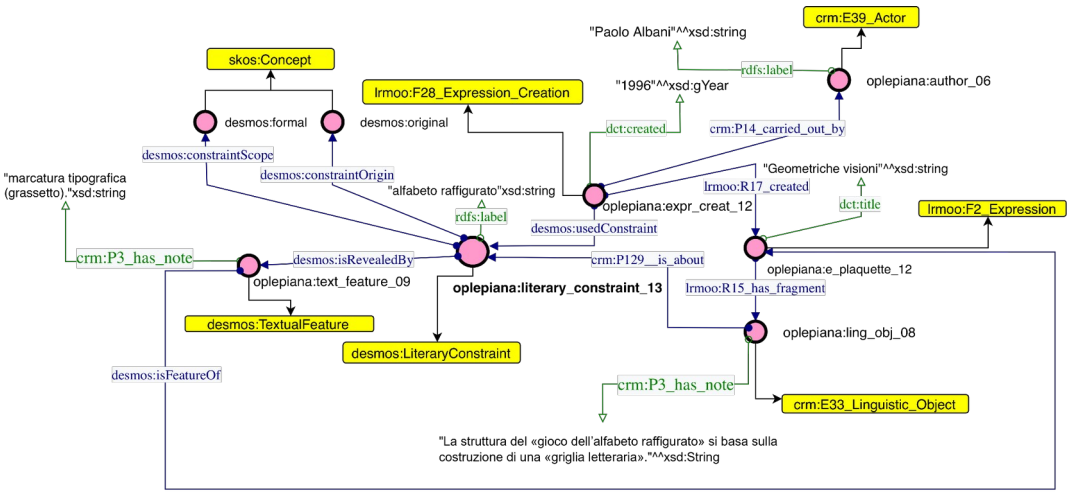
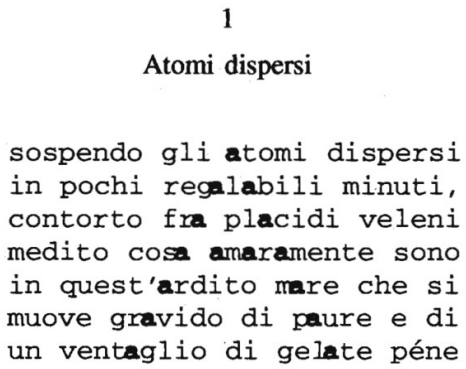


Figura 2. Esempio di modellazione di *Geometriche visioni*, Paolo Albani

A fini esemplificativi, una costrizione (desmos:LiteraryConstraint) inventata dall'autore e a prevalente ambito formale è collegata a un evento di creazione testuale (lrmoo:F28_Expression_Creation) e resa riconoscibile sia tramite una desmos:TextualFeature intratestuale (marcatura tipografica) sia mediante una dichiarazione autoriale paratestuale in prefazione (crm:E33_Linguistic_Object crm:P3_has_note). In questo contesto, l'alfabeto raffigurato è inteso come il procedimento compositivo che vincola la disposizione grafico-spaziale del testo (Figura 3), distinguendo la regola generativa dall'esito tipologico (poesia figurata o *technopaegnion*).



1

Atomi dispersi

sospendo gli atomi dispersi
in pochi regalabili minuti,
contorto fra placidi veleni
medito cosa amaramente sono
in quest'ardito mare che si
muove gravido di paure e di
un ventaglio di gelate péne

Figura 3. Estratto da *Geometriche visioni*, Paolo Albani

Una parte dello scenario presentato è formalizzata in sintassi Turtle come segue:

```
<https://w3id.org/desmos/oplepiana/expression_creation/expr_creat_12>
   a lrmoo:F28_Expression_Creation ;
   lrmoo:R17_created <https://w3id.org/desmos/oplepiana/expression/e_plaquette_12> ;
   desmos:usedConstraint <https://w3id.org/desmos/oplepiana/constraint/literary_constraint_13> .

<https://w3id.org/desmos/oplepiana/constraint/literary_constraint_13>
   a desmos:LiteraryConstraint ;
   rdfs:label "Alfabeto raffigurato"@it , "Figured Alphabet"@en ;
   desmos:constraintOrigin desmos:original ;
   desmos:constraintScope desmos:formal  .

<https://w3id.org/desmos/oplepiana/ling_obj/ling_obj_08> a crm:E33_Linguistic_Object ;
        crm:P3_has_note "La struttura del «gioco dell'alfabeto raffigurato» si basa sulla
```

costruzione di una «griglia letteraria»."^^xsd:string ;
crm:P129_is_about <https://w3id.org/desmos/oplepiana/constraint/literary_constraint_13> ;
prov:wasAttributedTo <https://w3id.org/desmos/oplepiana/author/author_06> .

5. RISULTATI, CONCLUSIONI E SVILUPPI FUTURI

La ricerca ha proposto un modello formale basato sulle tecnologie del Web Semantico per rappresentare e interrogare le regole sottese alla produzione testuale. DeSMòS si configura come strumento critico e computazionale per l'esplicitazione degli ingranaggi sottesi alla "macchina del testo" (Calvino, 2023: 201-221), offrendo una base empirica per trasformare la *contrainte* da fenomeno testuale implicito in conoscenza esplicita, strutturata e interrogabile.
L'applicazione del modello al corpus oplepiano offre riflessioni critiche di rilievo. In primo luogo, emerge una propensione allo svelamento della costrizione, un dato che inserisce l'Oplepo nel dibattito nato in seno all'Oulipo (d'Abele, 2014: 125-127), in cui coesistono posizioni divergenti tra chi rifiuta di rendere note le proprie *contraintes* e chi, al contrario, ne rivendica l'esplicitazione (Albani, 2016: 27). In secondo luogo, l'analisi delle relazioni di derivazione (lrmoo:R76_is_derivative_of) mostra che circa il 40% della produzione considerata si fonda su ipotesti riconducibili soprattutto al canone (Dante Alighieri; Francesco Petrarca). Nei casi di riscrittura individuati, tali ipotesti risultano rielaborati mediante costrizioni formali, con una preferenza per la manipolazione strutturale rispetto alla variazione semantica.
I passi successivi della ricerca si concentreranno sulla formalizzazione di una tassonomia delle operazioni logico-combinatorie che presiedono alla genesi delle diverse costrizioni letterarie impiegate nel testo (es. permutazione, addizione, sottrazione, moltiplicazione). L'ampliamento del corpus consentirà di analizzare ulteriori fenomeni costrittivi, valutando la capacità espressiva del modello su architetture testuali non lineari, come strutture combinatorie e forme ibride testo-immagine. Tale evoluzione porrà le basi per lo sviluppo di una piattaforma digitale di esplorazione semantica, orientata alla navigazione e all'analisi di fenomeni testuali che tendono a rimanere impliciti oppure non rappresentati.

BIBLIOGRAFIA

Albani, P. (2016). Le regole segrete: La contrainte nella letteratura potenziale. In A. Pozzo (Ed.), Svelare il segreto. Le strategie della dissimulazione. Atti della Giornata di studio internazionale, Bologna, 26 ottobre 2016. Scuola Superiore di Studi Umanistici.

Andrews, C. (2022). How to Do Things with Forms. The Oulipo and its inventions. Montreal & Kingston: McGill-Queen's University Press.

Aragona, R. (2002). Oplepiana. Dizionario di Letteratura Potenziale. Bologna: Zanichelli.

Bens, J. (1980). Oulipo 1960-1963. Paris: Christian Bourgois.

Brotchie, A., & Mathews, H. (1998). Oulipo Compendium. London: Atlas Press.

Bruseker, G., Carboni, N., Fielding, M., Nenova, D., & Hänsli, T. (2025). The Semantic Reference Data Modelling Method: Creating Understandable, Reusable and Sustainable Semantic Data Models. Journal of Open Humanities Data, 11(1). https://doi.org/10.5334/johd.282.

Calvino, I. (2023). Cibernetica e fantasmi. In Una pietra sopra (pp. 201-221). Milano: Mondadori.

d'Abele, M. C. (2014). L'Oulipo e Italo Calvino. Bern: Peter Lang.

Fickers, A., & Clavert, F. (2021). On pyramids, prisms, and scalable reading. Journal of Digital History, 1, 1–13. https://doi.org/10.1515/jdh-2021-1008.

Krautter, B. (2023). The Scales of (Computational) Literary Studies: Martin Mueller's Concept of Scalable Reading in Theory and Practice. In F. Armaselu & A. Fickers (Eds.), Zoomland: Exploring Scale in Digital History and Humanities, 261–286. Berlin: De Gruyter. https://doi.org/10.1515/9783111317779.

Oplepo. (2005). La Biblioteca Oplepiana. Bologna: in riga edizioni.

Oulipo. (1973). La littérature potentielle: (Créations, Re-créations, Récréations). Paris: Gallimard.

Poier-Bernhard, A. (2012) Texte nach Bauplan. Studien zur zeitgenössischen ludisch-methodischen Literatur in Frankreich und Italien. Heidelberg: Universitätsverlag Winter.

Queneau, R. (1965). Bâtons, chiffres et lettres. Paris: Gallimard.

[4/A] Studi letterari digitali e distant reading

Tracing Intertextuality at Scale. A Network Study of Poetic Influence

GABRIELE VEZZANI
RWTH AACHEN UNIVERSITY – GABRIELE.VEZZANI@IFAAR.RWTH-AACHEN.DE

ABSTRACT (ENGLISH)

In this paper, I argue that intertextuality constitutes a particularly promising domain for distant reading research. The use of Text Reuse Detection (TRD) algorithms, in fact, makes it possible to map intertextual relations at scale and to identify regularities that can shed new light on the dynamics underlying literary influence. To test this approach on a broad and heterogeneous sample of literary production, I draw on the PoeTree collection and construct corpora of poetry from the period 1750–1900 in seven languages: Italian, English, Russian, French, Czech, German, and Hungarian. Applying TRD, I extract intertextual networks and compare selected structural features, with particular attention to the temporal reach of intertextual links and their distribution across authors. The results reveal significant corpus-specific differences in the temporal dimension of intertextuality. By contrast, the distribution of intertextual references among authors appears stable across corpora, converging on a heavy-tailed form that is consistent with self-reinforcing, prestige-driven dynamics.
Keywords: Intertextuality, Distant Reading, Prestige, Networks

ABSTRACT (ITALIANO)

Intertestualità su vasta scala e network di influenza fra poeti.
In questo contributo sostengo che l'intertestualità rappresenti un ambito di ricerca particolarmente promettente per il distant reading. L'impiego di algoritmi di Text Reuse Detection (TRD) consente infatti di mappare su larga scala le relazioni intertestuali e di individuarvi possibili regolarità, le quali potrebbero approfondire la nostra conoscenza delle dinamiche alla base dei processi di influenza letteraria. Al fine di ottenere un campione ampio e diversificato su cui testare un simile approccio, mi avvalgo della collezione PoeTree e costruisco corpora di poesie relative al periodo 1750–1900 in sette lingue: italiano, inglese, russo, francese, ceco, tedesco e ungherese. L'applicazione di TRD mi permette di ricavare network intertestuali, di cui confronto alcune caratteristiche strutturali, in particolare l'estensione temporale dei legami intertestuali e la distribuzione degli stessi tra gli autori. I risultati evidenziano differenze significative tra i corpora per quanto riguarda la prima delle caratteristiche menzionate. La distribuzione di rimandi intertestuali tra autori sembra invece rimanere costante, convergendo su una forma 'heavy tailed' probabilmente causata dalla natura auto-rinforzantesi del prestigio letterario.
Parole chiave: Intertestualità, Distant Reading, Prestigio, Networks

1. INTRODUCTION

As noted by Ciotti , much research in the field of distant reading suffers from a mismatch between theory and methods, as the adoption of computational techniques often relies on literary-theoretical frameworks that, being grounded in interpretation and hermeneutics, lend themselves only partially to data-driven analysis. Intertextuality has been proposed as a research area wherein such divide might be attenuated. From a methodological perspective, the detection of textual exchange (more or less explicit and exact) between literary works constitutes a straightforward computational problem, thus allowing for robust, reproducible analyses. From a theoretical perspective, the study of intertextual relations at scale may serve as the bedrock for developing a conception of literature as a complex evolving system. Such evolutionary perspective, already present in the seminal works by Moretti , has been recognized as a viable theoretical framework to ground distant reading reaserch , and is currently receiving increasing attention by scholars who combine computational approaches to the theories and models of cultural evolution .
In this paper, I explore the use of Text Reuse Detection (TRD) as a computational framework to map intertextual relations in six different corpora of Italian, English, Russian, French, Czech, German and

Hungarian poems. The scope of my reaserch allows me to probe possible general dynamics of intertextual exchange, testing whether cross-corpora regularities can be found in: a) the average timespan between the publication of a poem and its subsequent intertextual reuses, and b) the distribution of instances of intertextual reuse across poets.

Intertextuality

The origin of the concept of intertextuality can be traced back to French post-structuralism, and particularly to the work of Julia Kristeva who, drawing inspiration from Saussurre and Bachtin, interpreted the literary text as "a permutation of texts, an intertextuality: in the space of a given text, several utterances, taken from other texts, intersect and neutralize one another" Famously aiming at downplaying the role of the author to foreground the text's own "productivity," post-structuralism proposed a radical view of intertextuality, conceptualizing it as the very condition for the emergence of literature. Radical, and, at the same time, somewhat vague: in Barthes, for instance, the intertextual field becomes a "general field of anonymous formulas, whose origin is rarely identifiable, of unconscious or automatic citations, given without quotation marks" .

Such a loose conception of intertextuality hinders an operational take on the subject – indeed, how can one write an algorithm to identify formulas whose origin is, by definition, "rarely identifiable?" Luckly, the vagueness of post-structuralist intertextuality has already been noted by several scholars, such as Genette , who famously re-organized the concept under the label of "transtextuality." A more restrictive, yet also more easily operationalizable concept is what Juvan calls "citationality," namely a correspondence between literary works based on "certain 'objective' textual features" . Based on this, in my work I limit the scope of intertextuality to cases in which the relationship between hypotext and hypertext entails a clear and demonstrable exchange of textual material.

Text Reuse Detection

By conceptualizing intertextuality as "citationality" it becomes possible to operationaize it through TRD. TRD refers to a computational technique aimed at identifying "meaningful reiteration[s] of text, usually beyond the simple repetition of common language" in a corpus. Such reiterations needn't necessarily be verbatim quotes. In fact, TRD algorithms often capture even allusive or paraphrastic correspondences, similar to those "unconscious or automatic citations" mentioned by Barthes. Roe used this technique to uncover the influences behind the *Encyclopedie* by Diderot e D'Alembert, as Düring and colleagues did to map the intertextual relations within a corpus of historical journals.

As already evident in these examples, the automatic detection of text reuse is not usually an end in itself, but rather aims at mapping patterns of influence within a corpus. With regard to this, however, the fact that TRD remains limited to the literal level of the text constitutes a limitation, since not every case of literary influence results in the (more or less explicit) exchange of textual material. Nonetheless, it can be assumed that these 'invisible' cases occur at a steady rate across the corpus, so even though TRD gives a partial view of influence patterns, the degree of information lost should be consistent for every work. In other words, I am considering text reuse patterns as a *proxy* for influence dynamics.

Cultural Evolution and Literary Influence

Cultural evolution has been proposed as a framework to explain how cultural systems change over time. According to this perspective, change is the macro-level, emergent result of micro-level interactions between individuals, whereby cultural information is exchanged through social learning processes that tend to favour certain cultural traits over others. A similar evolutionary framework has been applied to the study of literature, by conceptualizing literary forms as cultural traits transmitted through intertextual processes between authors .

Tracing the network of intertextual relations in several distinct corpora may help uncover general patterns, possibly shedding light over the long-studied phenomenon of literary evolution . To this end, I opted for an explorative approach and decided to tackle the following reaserch questions:

1. Whether the average timespan between the publication of a poem and its subsequent intertextual reuses varies significantly across corpora. This research question is inspired by the work of Huges and colleagues ,who demonstrate that stylistic influence in literary history is strongly temporally localized, with most significant similarities occurring among authors active within a limited time window.

2. Whether the distribution of instances of intertextual reuse across poets varies significantly across corpora. This research question is inspired by extensive research on citation networks and literary fame , which has shown that, across many domains, prestige dynamics tend to give rise to predictable distributions, whereby a small number of agents attract a disproportionate share of references.

2. MATERIALS AND METHODS

Data

My corpus is built on the basis of PoeTree , a multilingual collection of high-quality, metadata-rich poetic corpora accessible through a dedicated API. Given the rarity of cases of citationality, I anticipated corpus size to be a key parameter in my study. I therefore decided exclude the corpora with less than 10.000 poems, thus focusing only on the Hungarian, Russian, German, Italian, English, French, and Czech corpora.

Language	Version	# Poems	# Authors	# Tokens
Italian	original	39,749	243	11,073,882
	deduplicated	31,817	177	8,754,717
French	original	18,241	195	6,488,685
	deduplicated	18,067	194	6,421,757
Hungarian	original	12,953	48	2,560,232
	deduplicated	12,936	48	2,557,163
Russian	original	45,574	370	6,997,456
	deduplicated	45,488	368	6,987,947
English	original	41,806	447	15,039,289
	deduplicated	37,439	431	13,398,702
Czech	original	80,229	606	14,932,621
	deduplicated	70,798	606	13,241,243
German	original	74,351	258	13,458,630
	deduplicated	66,297	258	12,462,280

Table 1: Information on the corpora before and after deduplication

From each of these, I first excluded all poems lacking a publication date. I then applied an additional filter for duplicates. Although PoeTree authors report that each corpus is deduplicated, this appears to apply only within the work of individual authors (that is, duplicates as identical poems from the same author). However, I identified numerous unflagged duplicates *between* authors. For example, the Italian corpus includes poems from both Malipiero's *Petrarca spirituale* and Leopardi's *Crestomazia italiana poetica*, which anthologize overlapping selections of Petrarch's work. Because such cases may confound automatic intertextuality detection, I personally deduplicated each corpus, treating as duplicates any poems with a Levenshtein similarity above 0.8 to an earlier poem, regardless of authorship.

Intertextual Networks

I performed TRD on each corpus individually using Passim . Manual inspection of text reuse clusters[1] confirmed the versatility of the tool, capable of detecting not only verbatim quotes but also instances of parody or allusion.[2]

The inspection phase further revealed many clusters of intertextual relations between poems by the same pairs of authors. To prevent such tight reciprocal relationships from disproportionately shaping the overall network structure, I decided to retain only a single cluster per author group.

At this point I created, for each corpus, a directed network of intertextual relations by considering each poem as a node, with attributes encoding for its author and date of publication. I then used text reuse clusters to derive edges. For clusters with two elements, the latest poem was considered the source and the earliest the target of the edge. For clusters of size $n > 2$, n-1 edges were created, each connecting a 'later' poem (source) to the earliest (target).[3]

Poems in PoeTree corpora cover a vast period of time, ranging from the XIII to the XX century, with the distribution of poems per period varying significantly between corpora.[4] To allow for meaningful comparisons, I decided to limit my research to the period with the highest density of poems, namely

[1] In the context of TRD, a cluster is the set of documents that share the same passage.

[2] A telling example is the clearly parodical link identified by passim between Poe's *Annabel Lee* ("It was many and many a year ago, In a kingdom by the sea, That a maiden there lived whom you may know By the name of ANNABEL LEE; And this maiden she lived with no other thought Than to love and be loved by me. I was a child and she was a child, In this kingdom by the sea") and a poem by Robert F. Murray ("It was many and many a year ago, In a city by the sea, That a man there lived whom I happened to know By the name of Andrew M'Crie; And this man he slept in another room, But ground and had meals with me. I was an ass and he was an ass, In this city by the sea").

[3] In other words, for every passage shared by more than 2 works, I considered the earliest work as the original and all the others as 'imitators', discounting the possibility of mediated influence (work B quotes work A and work C quotes the same passage from work B). It should go without saying that this is a simplyfying, though necessary assumption. Note, however, that clusters of size > 2 turned out to be very rare, amounting to only 4% of the total.

[4] This is already made evident by the graph on the PoeTree web page: https://versologie.cz/poetree/general_info

1750-1900, and sample 20.000 poems within this period from each corpus, discarding those that didn't have enough poems. I then pruned the network of each of the remaining corpora by keeping only the edges whose source and target nodes were included in the downsampled corpus. Networks that contained fewer than 10 edges after pruning were excluded, as such small samples would not support reliable statistical inference. As a result, only the English, German, and Russian corpora were retained for analysis.

Language	Clusters	Edges Before Downsampling	Edges After Downsampling
Italian	265	50	NA
French	7	4	NA
Hungarian	9	8	NA
Russian	41	32	22
English	60	42	21
Czech	54	40	9
German	319	120	18

Table 2. Text reuse clusters and number of edges in each corpus before and after downsampling.

3. ANALYSIS

To test whether the average timespan between the publication of a poem and its subsequent intertextual reuses varies significantly across corpora, I computed, for each node in each network, the difference between the source's and target's publication dates. I then ran a Kruskal–Wallis test (since the data were non-normally distributed), with difference scores as the dependent variable (DV) and corpus as the independent variable (IV).

To address my second research question, I first aggregated each network's in-degree distribution by author; that is, I counted how many times each author of a poem in the network had been the target of an intertextual link. Authors in the corpus who never appear as the target of an intertextual link were assigned a value of 0.

I then tested whether the overall likelihood of being the target of an intertextual link varies across corpora. To do this, I created a binary variable encoding each author's quotation status (quoted vs. unquoted) and used a chi-square test of independence to compare quotation status across corpora.

Finally, I focused only on the intertextual networks (ignoring all non-quoted authors) and used a k-sample Anderson–Darling test to determine whether the nodes' in-degree distribution (aggregated by author) varies significantly across corpora.

4. RESULTS

The Kruskal–Wallis test revealed a significant effect of corpus on timespan, $H(2) = 10.5$, $p = .005$, indicating that the temporal scope of intertextual practices varies across corpora. The average timespan in the English (41 years, median = 39, SD = 25) and German (40 years, median = 35, SD = 33) corpora are considerably larger than the Russian corpus (18 years, median = 6, SD = 22). This result partially countradicts those of Hughes and colleagues , who found that literary influence tends to be localized within relatively small time windows and decreases as these get wider. It must be noted, though, that their measure of literary influence was based on sylometric symilarity, which is *bound* to decrease monotonicaly with time, as a result of simple linguistic drift. By contrast, the prestige of certain poems or authors, and therefore the likelyhood that they will be the target of an intertextual link, can at times grow with time, according to a process that Moretti aptly described as a "low of increasing return" (Moretti 2000b).

Proof of the presence of a similar general dynamic regulating intertextual patterns emerged from the investigation of my second research question, which found no significant variation in the distribution of intertextual links across corpora. The chi-square ($\chi^2(2) = 1.3$, $p = .5$) indicates that the likelihood of an author being quoted is very small (5% in the English and Russian corpora, 8% in the German corpus), and does not vary significantly across corpora. The k-sample Anderson–Darling test also did not reach conventional significance ($A^2 = -0.8$, $p = .6$) suggesting that the distributions of authors' in-degree counts are not significantly different across corpora. It is important to consider that the relatively small sample sizes[5] may have limited the power of this test. Nonetheless, the distributions clearly display a common

[5] 15 authors in the English corpus, 17 in the Russian one and 12 in the German one.

heavy-tailed behavior, whereby most authors receive only one or two intertextual links and only a few receive more, with a maximum of 4 in the English (Wordsworth) and German (Schiller and Herder) corpora, and of 5 in the Russian corpus (Zhukovskij).

5. DISCUSSION

Intertexruality is a promising research area to bridge literary theory and computational approaches. By conceptualizing it as citationality , in fact, it becomes possible to operationalize it as a TRD problem and study it at scale. This approach has been already proposed in recent research (Roe, 2024; Rebora, in press), to which my work adds further proof of its feasibility. However, my study also raises possible concerns. As data show, explicit text reuse between literary works is very rare. Conceptually, this doesn't necessarily pose an issue, as it can be assumed that, though text reuse networks are only a partial sample taken from larger intertextual networks, they still constitute a valid proxy for assessing their structural properties. Pragmatically, though, the rarity of text reuse limits the applicability of this method, as it requires very large corpora, from which a sufficient number of text reuse links can be extracted to allow for statistical inference.

The investigation of cross-corpora dynamics gave mixed results. Corpus-specific patterns were observed with regard to average timespan between intertextually linked poems. This finding indicates the absence of rigid time constraints on intertextuality, that is, the relative influence that old and contemporary authors may exert on a poem is contingent and context-specific.

By contrast, the distribution of intertextual links amongst authors follows a constant pattern: on the one hand, the likelyhood of being the target of an intertextual link is very small, but constant across corpora; on the other hand, all distributions of intertextual links across authors converge onto a 'heavy tail' shape, with many rarely quoted authors and a low number of highly quoted ones. A similar behavior has been observed in other areas, such as citantions in scientific papers , sales figures across books , and crtical mentions of poets . What these phenomena have in common is their reliance on the self-feeding loop of prestige dynamics, whereby highly successful works or authors are more likely than others to acqiure additional success. More than historically-determined factors, this dynamic seems to rely on the gregarious nature of human attention-attribution strategies (we tend to pay more attention to what others are paying attention to) and practical matters of visibility (highly successful authors have more chances to be read, and therefore quoted). My data can therefore be interpreted as suggesting the presence of a universal prestige-based dynamic regulating intertextual exchange and influence dynamics.

REFERENCES

Barabàsi, A.L, & Albert, R. (1999). Emergence of Scaling in Random Networks. Science, 286(5439), 509–512. https://doi.org/10.1126/science.286.5439.509

Barthes, R. (1990). Texte. In Encyclopaedia Universalis (Vol. 22, pp. 370–374). Encyclopaedia Universalis.

Boyd, R., & Richerson, P. J. (1985). Culture and the evolutionary process. University of Chicago Press.

Cavalli-Sforza, L. L., & Feldman, M. W. (1981). Cultural Transmission and Evolution: A Quantitative Approach. Princeton University Press.

Ciotti, F. (2021). Distant Reading in Literary Studies: A Methodology in Quest of Theory. Testo e Senso, (23), 195–213.

Düring, M., Romanello, M., Ehrmann, M., Beelen, K., Guido, D., Deseure, B., Bunout, E., Keck, J., & Apostolopoulos, P. (2023). impresso Text Reuse at Scale. An interface for the exploration of text reuse data in semantically enriched historical newspapers. Frontiers in Big Data, Volume 6-2023. https://doi.org/10.3389/fdata.2023.1249469

Fenner, T., Levene, M., & Loizou, G. (2010). Predicting the long tail of book sales: Unearthing the power-law exponent. Physica A: Statistical Mechanics and Its Applications, 389(12), 2416–2421. https://doi.org/10.1016/j.physa.2010.02.021

Genette, G. (1982). Palimpsestes: La littérature au second degré. Éditions du Seuil.

Henrich, J., & Gil-White, F. J. (2001). The evolution of prestige: Freely conferred deference as a mechanism for enhancing the benefits of cultural transmission. Evolution and Human Behavior, 22(3), 165–196. https://doi.org/10.1016/S1090-5138(00)00071-4

Hughes, J. M., Foti, N. J., Krakauer, D. C., & Rockmore, D. N. (2012). Quantitative patterns of stylistic influence in the evolution of literature. Proceedings of the National Academy of Sciences, 109(20), 7682–7686. https://doi.org/10.1073/pnas.1115407109

Jiménez, Á. V., & Mesoudi, A. (2019). Prestige-biased social learning: Current evidence and outstanding questions. Palgrave Communications, 5(1), 20. https://doi.org/10.1057/s41599-019-0228-7

Juvan, M. (2008). History and Poetics of Intertextuality. Purdue University Press.

Konle, L., Kröncke, M., Jannidis, F., & Winko, S. (2024). On the Unity of Literary Change. The Development of Emotions in German Poetry, Prose, and Drama between 1850 and 1920 as a Test Case. Proceedings of the Computational Humanities Research Conference (CHR 2024), Aarhus, December 4-6, 2024, CEUR Workshop Proceedings, 3834, 282–300. https://ceur-ws.org/Vol-3834/paper137.pdf

Kristeva, J. (1980). Desire in Language: A Semiotic Approach to Literature and Art. Columbia University Press.

Martindale, C. (1995). Fame more fickle than fortune: On the distribution of literary eminence. Poetics, 23(3), 219–234. https://doi.org/10.1016/0304-422X(94)00026-3

Moretti, F. (2000a). Conjectures on World Literature. New Left Review, 54–68. https://doi.org/10.64590/hxj

Moretti, F. (2000b). The Slaughterhouse of Literature. MLQ: Modern Language Quarterly, 61, 207–227.

Moretti, F. (2003). Graphs, Maps, Trees. New Left Review, 24(Nov–Dec), 67–93. https://doi.org/10.64590/06r

Plecháč, P., Cinková, S., Kolár, R., Šeļa, A., Sisto, M. D., Nugues, L., Haider, T., & Kočnik, N. (2024). PoeTree: Poetry Treebanks in Czech, English, French, German, Hungarian, Italian, Portuguese, Russian, Slovenian and Spanish. Research Data Journal for the Humanities and Social Sciences, 1–17. https://doi.org/10.1163/24523666-bja10044

Rebora, S. (in press). Computational literary studies: Theory and methods. Routledge.

Roe, G. (2024). Text reuse as cultural practice: Intertextuality in the 18th-century digital archive. Digital Enlightenment Studies, 2(1). https://doi.org/10.61147/des.23

Romanello, M., & Hengchen, S. (2021). Detecting Text Reuse with Passim. The Programming Historian, 10. https://doi.org/10.46430/phen0092

Sheridan, P., & Onodera, T. (2018). A Preferential Attachment Paradox: How Preferential Attachment Combines with Growth to Produce Networks with Log-normal In-degree Distributions. Scientific Reports, 8(1), 2811. https://doi.org/10.1038/s41598-018-21133-2

Sobchuk, O. (2023). Evolution of Modern Literature and Film. In J. J. Tehrani, J. Kendal, & R. L. Kendal (Eds.), Oxford Handbook of Cultural Evolution (1st ed., pp. 554–564). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198869252.013.45

Tynianov, Y. (2019). On Literary Evolution. In A. Morse & P. Redko (Eds.), Selected Essays on Literature, Theory and Film (pp. 267–282). Academic Studies Press. https://doi.org/10.1515/9781644690635-015

The Landscape of Italianistica: Italian Literary Studies and Digital Humanities in Semantic Space

MARIA LEVCHENKO
UNIVERSITY OF BOLOGNA – MARIIA.LEVCHENKO@UNIBO.IT

ABSTRACT

This paper presents an empirical characterization of Italian literary studies (italianistica) and its relationship to Digital Humanities research in Italy. We analyze three corpora: ADI conference proceedings (2,345 papers, 2008–2023), Italinemo journal abstracts (39,077 articles, 1996–2025), and AIUCD conference papers (710 papers, 2013–2025). Using LLM-based metadata extraction and semantic clustering with OpenAI embeddings, we map the thematic and methodological landscape of each community. Our findings show that ADI and Italinemo are semantically nearly identical (centroid distance 0.70), organized around canonical authors and dominated by intertextual and stylistic analysis. AIUCD occupies a distinct semantic region (distance 6.36 from ADI), organized by method rather than object of study, with only 1.1% of papers falling within the traditional scholarship space. We identify specific opportunities for methodological dissemination: intertextuality and stylistic analysis — central to italianistica but underrepresented in DH — are well-suited to computational approaches. The success of digital philology projects suggests models for bridging the two communities.
Keywords: digital humanities; Italian literary studies; corpus analysis; semantic clustering; computational text analysis

ABSTRACT

Il paesaggio dell'italianistica: tra studi letterari e Digital Humanities.
Questo contributo presenta una caratterizzazione empirica degli studi letterari italiani (italianistica) e della loro relazione con la ricerca nelle Digital Humanities in Italia. Analizziamo tre corpora: gli atti dei convegni ADI (2.345 contributi, 2008–2023), gli abstract delle riviste indicizzate in Italinemo (39.077 articoli, 1996–2025) e gli atti dei convegni AIUCD (710 contributi, 2013–2025). Attraverso l'estrazione di metadati tramite LLM e il clustering semantico con embedding OpenAI, mappiamo il panorama tematico e metodologico di ciascuna comunità. I risultati mostrano che ADI e Italinemo sono semanticamente quasi identici (distanza tra centroidi 0,70), organizzati attorno ad autori canonici e dominati dall'analisi intertestuale e stilistica. AIUCD occupa una regione semantica distinta (distanza 6,36 da ADI), organizzata per metodo piuttosto che per oggetto di studio. Identifichiamo opportunità concrete per la disseminazione metodologica: l'intertestualità e l'analisi stilistica — centrali per l'italianistica ma sottorappresentate nelle DH — si prestano ad approcci computazionali.
Parole chiave: digital humanities; studi letterari italiani; analisi di corpora; clustering semantico; analisi computazionale del testo

1. INTRODUCTION

The relationship between digital humanities and traditional literary scholarship in Italy has long been a subject of theoretical debate. Since the founding of AIUCD in 2011, the Italian DH community has navigated a productive tension between two visions of the field: as a methodological space with its own epistemological claims, or as a set of computational approaches that serve the interpretive goals of established disciplines (AIUCD Board, 2024). Fabio Ciotti (2017) has acknowledged that "i metodi computazionali non hanno avuto una influenza soddisfacente negli studi letterari 'tradizionali,'" attributing this in part to "la (asserita) irriducibilità del letterario a ogni approccio formale e quantitativo."
Previous work has characterized the Italian DH community's internal structure. Sprugnoli et al. (2019) compared AIUCD and CLiC-it conferences (2014–2017) using co-authorship networks, terminology extraction, and citation analysis, finding that AIUCD exhibits a fragmented "plural-world" structure with limited collaboration across research groups. Hall (2019) surveyed scholars at the intersection of Italian Studies and Digital Humanities, documenting professional identity tensions and limited institutional support. Our study complements these works by examining the relationship between Italian DH and

traditional literary scholarship, asking what *italianistica* actually does and where — if anywhere — computational methods might contribute.

This gap matters because the terms of engagement have shifted. The past decade has seen significant advances in digital philology and scholarly editing: Paola Italia's *Editing Duemila* (2020) theorizes how digital transmission transforms textual production and editorial responsibility, while Tiziana Mancinelli and Elena Pierazzo's *Che cos'è un'edizione scientifica digitale* (2020) offers a conceptual and methodological introduction to digital scholarly editions, surveying key projects and helping readers understand their possibilities and limits. Major projects like the Petrarchive, the Naples Dante Project with its Illuminated Dante archive, the BibliotecaItaliana, Edizione Nazionale Pirandello, PhiloEditor, LeggoManzoni, and the Hypermedia Dante Network demonstrate that Italian DH can serve the textual-genetic and editorial concerns of *italianistica* with sophistication and rigor. But philological infrastructure — however valuable — addresses only a portion of what literary scholars do. What about the broader interpretive enterprise: the analysis of intertextuality, style, themes, and meaning that is typically understood as the discipline's core activity? Has computational research engaged with these questions, or has it primarily developed tools and methodologies for its own ends? Answering them requires moving from theoretical debate to empirical characterization of both fields.

This paper presents a preliminary large-scale empirical characterization of Italian literary studies and its relationship to Digital Humanities. Analyzing over 42,000 documents from ADI conference proceedings, Italinemo journal abstracts, and AIUCD papers, we ask three interrelated questions: (1) How is the field of Italian literary studies organized thematically and methodologically? (2) What is the semantic and topical relationship between traditional scholarship and Italian DH research? (3) Where do the analytical priorities of italianistica—intertextuality, stylistic analysis, thematic interpretation—intersect with the capabilities of computational methods, and where do significant gaps remain? Our goal is not to adjudicate the theoretical debate over DH's disciplinary status, but to provide an empirical foundation for a more productive conversation: one that identifies specific opportunities for computational methods to serve the interpretive questions that Italian literary scholars actually ask.

2. MAPPING THE FIELD

2.1. Corpora and Approach. This study draws on three complementary corpora representing Italian literary scholarship and Digital Humanities research: *ADI* (Associazione degli Italianisti) conference proceedings (2,345 papers, 2008–2023), *Italinemo* journal abstracts (39,077 articles from 147 journals, 1996–2025), and *AIUCD* conference proceedings (710 papers, 2013–2025). All sources were scraped from their respective websites; PDFs were converted to markdown using *pymupdf4llm* and *docling*. For the full-text corpora (ADI and AIUCD), we used GPT-5-mini with a structured output schema to extract metadata including authors discussed, historical periods, analytical focus, and research questions. To map the semantic landscape, we embedded all documents using OpenAI's *text-embedding-3-small* and applied UMAP with HDBSCAN clustering, yielding 16 clusters for ADI, 41 for Italinemo, and 14 for AIUCD. For cross-corpus comparison, we projected all 42,132 documents into a shared semantic space.

2.2. Landscape of Italianistica. What does traditional Italian literary scholarship study, and how? We applied LLM-based metadata extraction to 2,345 ADI conference papers to characterize the field's methods, objects, and temporal scope. Close reading remains the dominant method, characterizing half of all papers (50.1%). Comparative analysis (23.6%) and archival research (22.9%) follow, while theoretical approaches account for only 2.4% and quantitative methods for just 1.0%. This methodological profile shows no significant temporal trend: computational approaches remain as rare in 2023 as in 2008. Figure 1 summarizes the field's temporal and object scope. The Secondo Novecento and Ottocento dominate, followed by the Renaissance; contemporary literature receives less attention. The object scope reveals a discipline focused on individual authors and single works.

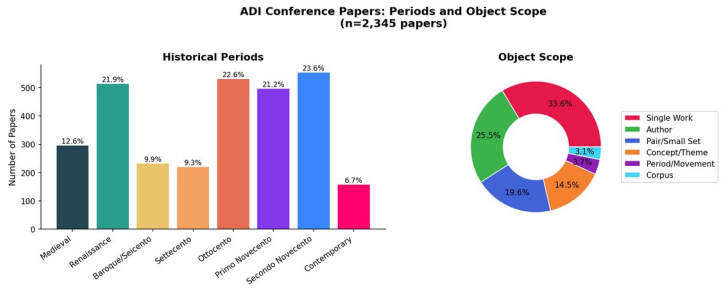


Figure 1. Temporal and object scope of ADI conference papers

The semantic clustering confirms this author-centric organization (Figure 2). ADI's 16 clusters are predominantly organized around individual authors — Dante Alighieri, Francesco Petrarca, Ludovico Ariosto, Luigi Pirandello — with additional clusters for period-based groupings like "Secondo Novecento: Form/Style/Narration" and "Ottocento: Historical Context." Italinemo's 41 clusters reproduce this pattern at larger scale: dedicated clusters for Dante Commedia Exegesis (3,130 papers), Verga and Verismo (1,170), Leonardo Sciascia Studies (1,213). Specialized journals cluster by topic rather than venue — *Studi danteschi* papers cluster with Dante scholarship from general journals — indicating genuine disciplinary organization. In both corpora, the field organizes around *objects of study* rather than methods or theoretical frameworks.

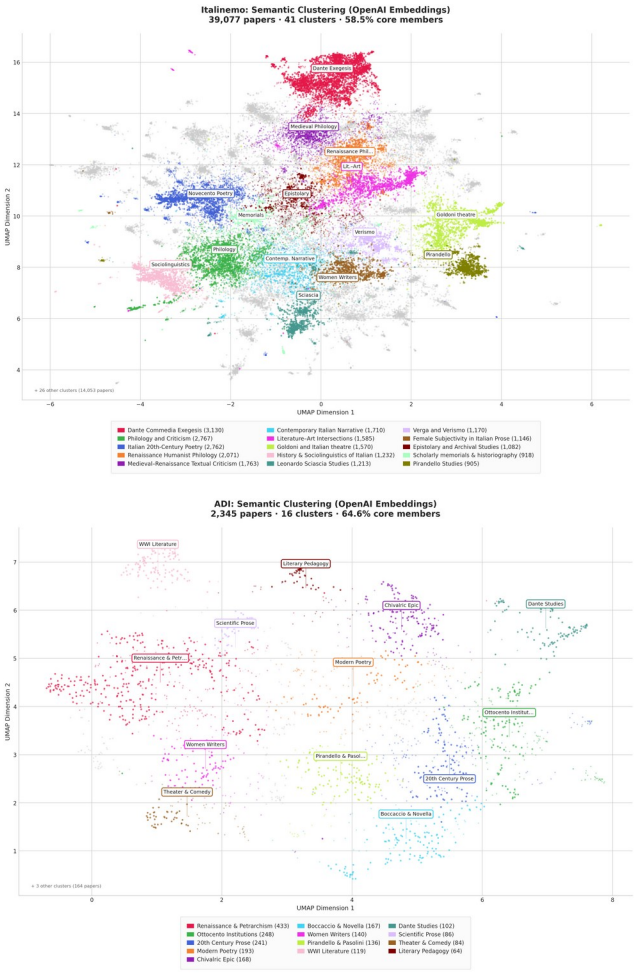


Figure 2. Semantic clusters for Italinemo (41 clusters) and ADI (16 clusters)

2.3 Digital Humanities in the Semantic Space. AIUCD's semantic structure reveals a fundamentally different organizing principle. Figure 3 shows the 14 clusters: Digital Editions (153 papers), Libraries & Archives (88), Linked Data Infrastructure (83), Digital Philology (70), DH Theory (56), Computational Linguistics (44), Stylometry (19). Where *italianistica* clusters by author, DH clusters by *method*. The

largest cluster, "Digital Editions – Italian," contains projects on Verga, Pirandello, and Manzoni, but these sit alongside classical and medieval editions in a methodologically-defined space.

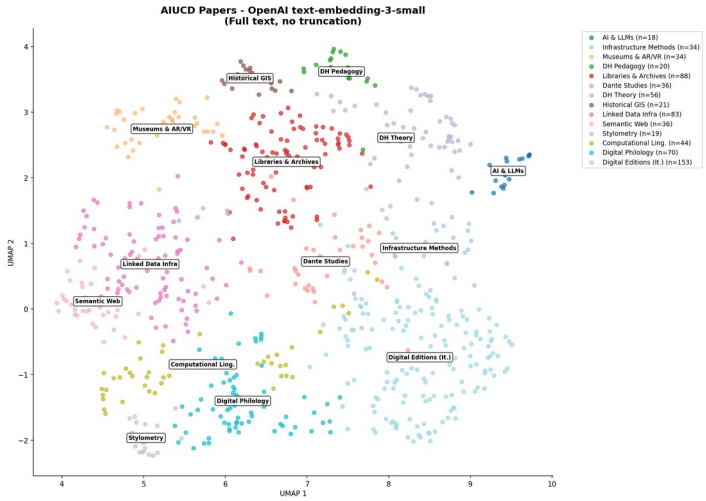


Figure 3. AIUCD clusters

Only 140 AIUCD papers (19.7%) concern Italian literature. How do these compare to traditional scholarship? Figures 4 reveal systematic gaps. In analytical focus (Figure 4), intertextuality dominates ADI (45.2%) but appears in only 10% of AIUCD Italian literature papers. Similar gaps appear for thematic analysis and form/style. The one area where AIUCD exceeds ADI is textual genesis (+18pp) — precisely the philological infrastructure work that digital methods have successfully served. The author coverage gap is equally stark. Even when Digital Humanities engages Italian literature, it prioritizes editorial infrastructure over the interpretive questions that define the discipline.

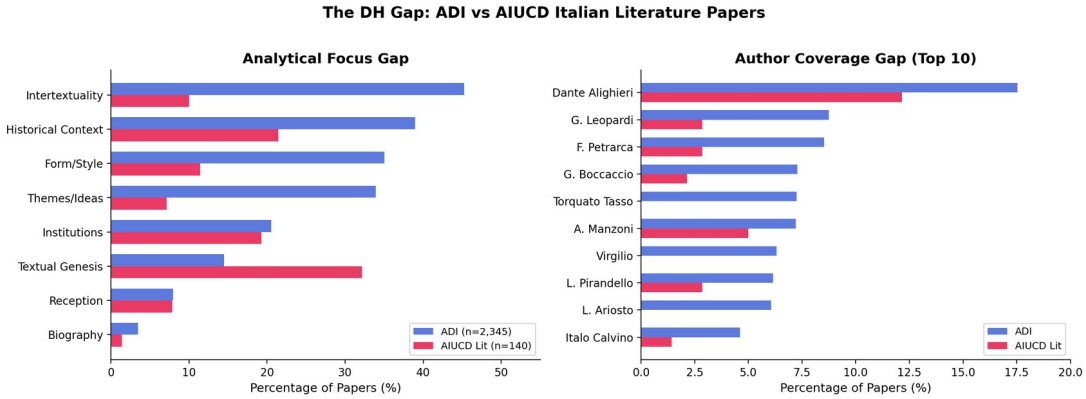


Figure 4. The DH gap: analytical focus and author coverage in ADI (n=2,345) versus AIUCD Italian literature papers (n=140)

2.4. The Separation. Figure 5 projects all 42,134 documents into a shared semantic space. The result confirms the patterns above: AIUCD (red) forms a compact cluster at the edge, while ADI (blue) and Italinemo (green) overlap almost completely. The centroid distance between ADI and Italinemo is just 0.70 — they are semantically indistinguishable, confirming that conferences and journals represent the same field. The distance from AIUCD to ADI is 6.36, nearly ten times greater. This is not gradual divergence but categorical separation.

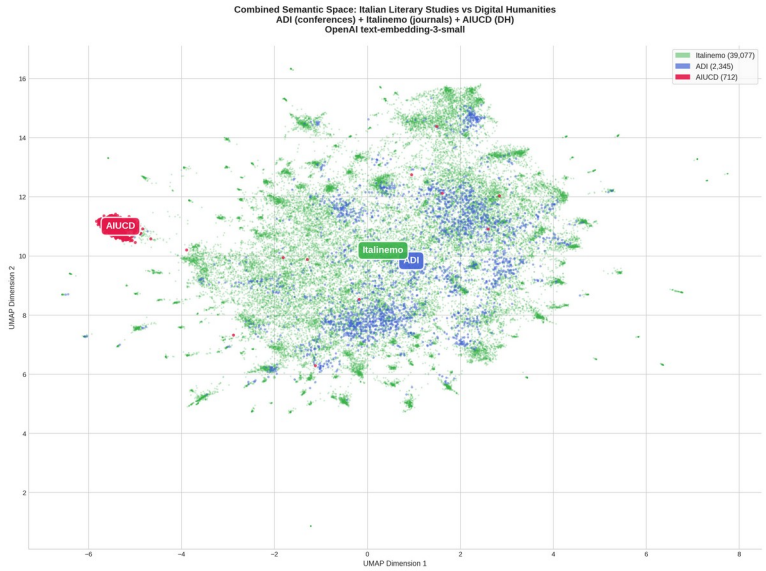


Figure 5. Shared semantic space

Only 8 AIUCD papers (1.1%) fall within the ADI/Italinemo region. These "bridge papers" share common features: they engage canonical authors, take philological approaches, and frame contributions in terms legible to traditional scholarship — a digital edition of *Se questo è un uomo* (Tagliacozzo, 2024), NLP analysis of Renaissance correspondence (Vetrugno, 2016; Bersano, 2020), a poetry analysis tool for Italian verse (Delmonte, 2023). The communities are also socially distinct: of 1,059 AIUCD authors and 1,394 ADI authors, only 35 (1.5%) have published in both venues. These bridge scholars are predominantly DH-primary researchers who occasionally present at literary conferences, not literary scholars adopting computational methods. The flow is asymmetrical.

2.5. Opportunities. The gap analysis suggests where computational methods could serve traditional research questions that italianistica actually asks. Two areas emerge as highest priority: intertextuality and stylistic analysis. Intertextuality — the study of sources, allusions, and influences — dominates traditional scholarship (45% of ADI papers) but appears in only 10% of AIUCD work, a gap of 35 percentage points. This is computationally tractable: text reuse detection, embedding-based similarity search, and network analysis of citation patterns are mature techniques that could directly serve scholars tracing how Ariosto reworks Boiardo or how Montale echoes Dante. Similarly, form and style analysis (35% ADI, 11% AIUCD, −24pp) aligns well with computational stylistics, authorship attribution methods, and automated meter analysis — tools already developed but rarely applied to the questions Italian literary scholars ask.

Thematic interpretation (34% ADI, 7% AIUCD) and historical contextualization (39% ADI, 21% AIUCD) present moderate opportunities: topic modeling and named entity recognition can support but not replace humanistic judgment in these domains.

The one area where DH has demonstrably succeeded is textual genesis: AIUCD actually exceeds ADI here (32% vs 15%), reflecting the community's strength in digital critical editions and collation tools. Projects like the Pirandello Nazionale hyperedition and LeggoManzoni show what is possible when computational methods engage directly with philological questions. The challenge is extending this model beyond editorial infrastructure to the interpretive work — the analysis of meaning, influence, and form — that defines the discipline.

3. CONCLUSION

This paper has offered a preliminary empirical characterization of Italian literary studies and its relationship to Digital Humanities research. Three findings emerge from the analysis. First, italianistica as represented in ADI and Italinemo exhibits consistent organization around canonical authors, with intertextuality, historical contextualization, and stylistic analysis as dominant analytical concerns, and close reading as the prevailing method. Second, the ADI and Italinemo corpora are semantically nearly identical (centroid distance 0.70), confirming that conferences and journals represent the same

disciplinary space. Third, AIUCD occupies a distinct semantic region (distance 6.36 from ADI), organized primarily by *method* rather than *object of study*, with limited overlap in authors, analytical focus, and research questions.

A citation analysis confirms this divide: only 5.2% of ADI papers reference any digital resource. When they do, they cite content databases (Biblioteca Italiana, Dartmouth Dante Project) as sources to consult — not as tools or methods to engage with.

These patterns suggest two concrete opportunities. First, computational methods like text reuse detection and stylometry could directly serve the interpretive concerns — intertextuality, stylistic analysis — that remain central to traditional scholarship. Second, existing digital infrastructure for philological work remains underutilized beyond specialist communities. Bridge scholars and institutional hubs offer models for closing both gaps.

Our analysis has limitations. LLM-based extraction reduces complex arguments to structured metadata, inevitably flattening interpretive nuance; semantic clustering represents scholarly work as vectors, privileging lexical patterns over conceptual depth. Conference papers may also not fully represent either field's research output. The corpus examined cannot claim exhaustiveness: ADI primarily represents one sector of Italian literary studies, while philology, comparative literature and contemporary studies have separate associations and publication venues. However, by combining ADI proceedings with Italinemo's 147 journals — including philological and comparative publications — we aimed to approximate a survey spanning diverse sectors of Italian literary scholarship, while acknowledging that other potential corpora exist.

Data and code are available on GitHub: https://github.com/mary-lev/italian_studies

REFERENCES

AIUCD Board. (2024, April 15). Come stanno le Digital Humanities in Italia? Informatica Umanistica e Cultura Digitale. https://infouma.hypotheses.org/3083

Bersano, O. (2020). Il trattamento automatico del linguaggio applicato all'italiano volgare. In AIUCD 2020: La svolta inevitabile. AIUCD.

Ciotti, F. (2017). Modelli e metodi computazionali per la critica letteraria. In Letteratura e scienze: Atti del XIX Congresso dell'ADI. ADI Editore.

Delmonte, R. (2023). Verso la creazione della versione italiana di SPARSAR. In AIUCD 2023: La memoria digitale. AIUCD.

Hall, C. (2019). Digital Humanities and Italian Studies: Intersections and oppositions. Italian Culture, 37(2), 97–115. https://doi.org/10.1080/01614622.2019.1717754

Italia, P. (2020). Editing Duemila: Per una filologia dei testi digitali. Salerno Editrice.

Mancinelli, T., & Pierazzo, E. (2020). Che cos'è un'edizione scientifica digitale. Carocci.

Sprugnoli, R., Pardelli, G., Boschetti, F., & Del Gratta, R. (2019). Un'analisi multidimensionale della ricerca italiana nel campo delle Digital Humanities e della Linguistica Computazionale. Umanistica Digitale, 3(5). https://doi.org/10.6092/issn.2532-8816/8581

Tagliacozzo, D. (2024). Per un'edizione digitale di *Se questo è un uomo*. In AIUCD 2024: Filologie e Digitalità. AIUCD.

Vetrugno, R. (2016). Mapping Castiglione's Letters. In AIUCD 2016: Edizioni digitali. AIUCD.

L’Edicola della Storia di Trieste: NLP e *distant reading* Per il confronto tra periodici storici italiani e sloveni nella Trieste del 1902

CRISTINA FENU[1], GIOVANNI PINNA[2]

[1]BIBLIOTECA CIVICA “ATTILIO HORTIS” TRIESTE, ITALIA - CRISTINA.FENU@COMUNE.TRIESTE.IT
[2]UNIVERSITÀ DEGLI STUDI DI TRIESTE, ITALIA - GIOVANNI.PINNA@PHD.UNITS.IT

ABSTRACT (ITALIANO)

Il contributo presenta un primo studio esplorativo che applica tecniche di NLP all’analisi comparata di due periodici storici triestini: "Il Piccolo" (in lingua italiana) ed "Edinost" (in lingua slovena). Attraverso un approccio di *distant reading*, il lavoro analizza un *subset* delle due testate, 50 fascicoli pubblicati nel febbraio e nel settembre 1902, periodo segnato, in particolare, dallo sciopero generale dei fuochisti del Lloyd Austriaco e dai conseguenti scontri di piazza. La *pipeline* metodologica prevede correzione OCR assistita da LLM, traduzione automatica in inglese come *medium* linguistico comparativo, *Topic Analysis* (BERTopic) e *Sentiment Analysis* con modelli specifici per italiano, sloveno e inglese. L’analisi *intra-corpus* rivela strategie discorsive strutturalmente divergenti: "Il Piccolo" adotta un registro cronachistico-neutro (87-88% *Sentiment* neutro), mentre "Edinost" privilegia un orientamento valutativo e identitario (*Sentiment* negativo 40-49%). L’analisi *inter-corpora* condotta sulle versioni in inglese quantifica le divergenze: il *gap* di *Sentiment* negativo raggiunge 7,5 punti percentuali in febbraio, con agende tematiche largamente autonome e assenza di convergenza sistematica sui *Topic*. I risultati confermano che i due quotidiani operano come attori discorsivi distinti, portatori di funzioni politico-identitarie diverse nello spazio pubblico triestino. Il progetto si colloca nella prospettiva della *digital public history*, proponendo il *distant reading* come strumento di mediazione per rendere accessibili narrazioni storiche parallele in una città plurilingue.
Parole chiave: NLP; periodici storici; *distant reading*; analisi comparativa multilingue; Trieste

ABSTRACT (ENGLISH)

The Trieste History Newsstand: NLP and distant reading for comparing Italian and Slovenian historical periodicals in Trieste in 1902.
This paper presents an initial exploratory study that applies NLP techniques to the comparative study of two historical Trieste periodicals: “Il Piccolo” (an Italian-language daily newspaper) and “Edinost” (a Slovenian-language newspaper). Using a distant reading approach, the work analyses a limited subset of the two newspapers, 50 issues published in February and September 1902, a period marked, in particular, by the general strike of the Austrian Lloyd stokers and the ensuing street clashes. The methodological pipeline includes LLM-assisted OCR correction, machine translation into English as a comparative linguistic medium, Topic Analysis (BERTopic) and Sentiment Analysis with specific models for Italian, Slovenian and English. The intra-corpus analysis reveals structurally divergent discursive strategies: “Il Piccolo” adopts a neutral, chronicle-style register (87-88% neutral sentiment), while “Edinost” favours an evaluative and identity-based approach (40-49% negative sentiment). The *inter-corpora* analysis conducted on the English versions quantifies the divergences: the negative sentiment gap reaches 7.5 percentage points in February, with largely autonomous thematic agendas and no systematic convergence on topics. The results confirm that the two newspapers operate as distinct discursive actors, performing different political-identity functions in the Trieste public sphere. The project is part of the digital public history perspective, proposing distant reading as a mediation tool to make parallel historical narratives accessible in a multilingual city.
Keywords: NLP; historical periodicals; distant reading; multilingual comparative analysis; Trieste

1. INTRODUZIONE

Trieste, porto dell’Impero asburgico fino al 1918 e città italiana di confine, rappresenta un contesto urbano ed etnoculturale di straordinaria complessità, in cui la coesistenza - talvolta polarizzata - tra comunità italiana e slovena emerge con particolare evidenza anche attraverso l’analisi della stampa

periodica. L'assunto di partenza del presente lavoro[1] è che i quotidiani triestini di lingua italiana e slovena non restituiscano una narrazione unitaria della storia cittadina, ma costruiscano rappresentazioni divergenti degli eventi, riconducibili a differenti sensibilità e contesti socioculturali e politici. In questa prospettiva, la stampa periodica è considerata come luogo di produzione discorsiva, in cui si costruiscono cornici interpretative e strategie di legittimazione del discorso pubblico.
Il progetto propone l'applicazione di tecniche di *Natural Language Processing* (NLP) allo studio comparato di un *subset* circoscritto di due periodici storici triestini: "Il Piccolo" (1881–), ad oggi il principale quotidiano di lingua italiana della città il cui archivio digitalizzato, consultabile *online*[2], è conservato presso la Biblioteca civica di Trieste, ed "Edinost" (1876–1928), giornale di lingua slovena, organo della omonima Società Politica, digitalizzato e reso disponibile in pubblico dominio dalla *Slovenian Digital Library*[3]. L'analisi adotta l'approccio del *distant reading* (Moretti, 2000; 2013) per analizzare simultaneamente centinaia di articoli, estrarre *Topic* ricorrenti, distribuzioni di *Sentiment* e frequenze lessicali, al fine di mettere in luce le differenze narrative tra le due testate. Tale approccio consente di rendere accessibile al pubblico più ampio il patrimonio informativo custodito nei periodici storici multilingui, trasformando fonti altrimenti riservate alla ricerca specialistica in narrazioni interrogabili e comparabili.
Le domande di ricerca che guidano il lavoro sono le seguenti:

- È possibile individuare divergenze e convergenze tra narrazioni giornalistiche operando su *corpora* in lingue diverse?
- Emergono differenze di *framing* nel *Sentiment* associato agli stessi eventi?
- In che misura la traduzione automatica in una lingua veicolare contemporanea consente di rendere più comparabili i risultati?

L'originalità del contributo risiede nel confronto sistematico, attraverso l'integrazione di tecniche di NLP e interpretazione storica, di narrazioni giornalistiche coeve prodotte nello stesso contesto urbano da comunità linguistiche differenti.

2. STATO DELL'ARTE

L'analisi computazionale dei periodici storici si inserisce in un filone di ricerca ormai consolidato, ma nel panorama italiano l'applicazione sistematica di tecniche di *distant reading* a giornali storici plurilingui rappresenta un ambito ancora in fase di sviluppo. Il settore di studi ha conosciuto una crescita significativa grazie a campagne di digitalizzazione su larga scala ed alla progressiva trasformazione dei giornali d'epoca in *Big Data del passato*. Il progetto *impresso*[4] (Ehrmann et al., 2020) ha affrontato la natura non canonica dei testi giornalistici storici attraverso normalizzazione testuale, *Topic modeling* multilingue e risorse annotate, configurando le lingue storiche come scenari *low-resource* che richiedono *pipeline* calibrate e costante validazione interpretativa. Studi recenti hanno evidenziato, poi, il potenziale dei LLM nell'elaborazione dei testi storici. Se il *dataset ChroniclingAmericaQA*[5] (Piryani et al., 2024) documenta come il rumore OCR riduce drasticamente le prestazioni dei modelli NLP, Thomas et al. (2024) propongono l'uso di modelli generativi guidati da istruzioni proprio per la correzione automatica dell'OCR di giornali del XIX secolo. In contesti multilingui e *low-resource*, Borenstein et al. (2023) mostrano, infine, come la traduzione automatica verso lingue ad alta disponibilità di risorse possa migliorare sensibilmente le *performance* dei modelli NLP. Sul versante dell'analisi discorsiva, Marjanen et al. (2020) applicano modelli LDA e *Dynamic Topic Models* a quotidiani finlandesi del XIX-XX secolo, mostrando come tali strumenti intercettino discorsi di ampia scala ma evidenziando limiti legati alla rappresentazione del tempo.

---

[1] Il presente contributo prende origine dal *project work* realizzato da Cristina Fenu nell'ambito del Master in *Digital Humanities* dell'Università degli Studi di Milano (a.a. 2024–2025), sviluppato durante un tirocinio presso il Machine Learning Lab del DIA – Università degli Studi di Trieste, e successivamente rielaborato in forma di articolo con Giovanni Pinna, dottore di ricerca in *Applied Data Science and Artificial Intelligence* presso l'Università di Trieste.
[2] https://archivio.ilpiccolo.it/ (cons. 08/02/2026).
[3] La testata in formato digitale è consultabile alla URN:NBN:SI:spr-BJPVOFS8 (cons. 08/02/2026).
[4] https://impresso-project.ch/ (cons. 08/02/2026).
[5] https://github.com/DataScienceUIBK/ChroniclingAmericaQA (cons. 08/02/2026). Il portale del *National Digital Newspaper Program* (*Library of Congress* e *National Endowment for the Humanities*) rende consultabili oltre 6 milioni di pagine di quotidiani storici pubblicati in 30 Stati: https://www.loc.gov/collections/chronicling-america/ (cons. 08/02/2026).

Un riferimento diretto per il presente lavoro è costituito dalle ricerche sui periodici storici sloveni confluite nella costruzione del *corpus sPeriodika* (Dobranić et al., 2023) reso interrogabile tramite l'infrastruttura CLARIN.SI[6]. In diretta continuità, Gorjanc et al. (2025) hanno presentato alla conferenza ADHO DH2025 i primi risultati di applicazioni di analisi linguistica automatica e *distant reading* applicati al discorso giornalistico storico sloveno.
Nel panorama italiano, l'applicazione sistematica di tecniche NLP ai periodici storici resta limitata (Cafagna et al., 2020: 9–22; Ludovisi, 2025: 222–227).

3. *CORPUS* E *PIPELINE* METODOLOGICA

Il *corpus* analizzato comprende 50 fascicoli[7]: 23 numeri di "Edinost"[8] e 27 de "Il Piccolo", pubblicati dal 17 al 28 febbraio e dal 1° al 15 settembre 1902. La selezione è stata guidata da eventi di cronaca locale valutati da esperti di dominio come rilevanti per entrambe le comunità linguistiche: nel febbraio 1902, lo sciopero dei fuochisti del Lloyd Austriaco[9]; nel settembre, la morte del *leader* sloveno Ivan Nabergoj[10] e l'inaugurazione della tranvia elettrica che collega il centro urbano al Carso triestino[11].
La *pipeline* metodologica[12] si articola in fasi progressive (vd. Fig. 1). L'OCR di base, fornito come *layer* testuale nei PDF disponibili negli archivi digitali delle due testate, è stato prodotto in fase di digitalizzazione de "Il Piccolo" con Tesseract 4.1.1 (release candidate 1, build gb36c, basata su motore LSTM)[13]. L'*output* presenta caratteristiche compatibili con un riconoscimento a livello di parola senza analisi del layout di pagina: ordine di lettura compromesso, colonne concatenate, frasi spezzate. È stato pertanto migliorato tramite correzione assistita da LLM (Claude Opus 4.5) mediante *prompting* iterativo e adattivo calibrato sulle specificità dei documenti digitalizzati. La valutazione dell'accuratezza su un campione rappresentativo ha restituito un WER aggregato del 20,44% per "Il Piccolo" (12% escludendo gli inserti pubblicitari) e del 10,61% per "Edinost". Tutti i fascicoli sono stati tradotti in inglese tramite DeepL e sottoposti a revisione assistita da AI per garantire coerenza terminologica.
Il *preprocessing* ha incluso: pulizia del *corpus* (eliminazione di annunci, inserti pubblicitari, romanzi d'appendice); segmentazione strutturale *rule-based* implementata in *Python*; tokenizzazione e lemmatizzazione con *spaCy* (v. ≥ 3.8.0), utilizzando i modelli *it_core_news_lg* per l'italiano, *sl_core_news_sm* per lo sloveno ed *en_core_web_lg* per l'inglese, integrati con liste di *stop words* personalizzate calibrate sui testi storici.
La *Topic Analysis* è stata condotta con BERTopic (*embeddings*: multilingual *MiniLM*[14] per l'italiano, *SloBERTa*[15] per lo sloveno) assumendo come unità documentale il singolo segmento prodotto dalla segmentazione *rule-based*, per un totale di documenti significativamente superiore ai 50 fascicoli iniziali. La coerenza dei *topic* estratti è stata validata qualitativamente tramite rilettura esperta dei documenti originali, verificando la corrispondenza semantica tra i *cluster* prodotti e i contenuti effettivi dei segmenti.
La *Sentiment Analysis* ha impiegato modelli specifici: *neuraly/bert-base-italian-cased-sentiment*[16] per l'italiano, *CLASSLA xlm-r-parlasent*[17] per lo sloveno, e *cardiffnlp/twitter-roberta-base-sentiment-latest* per le versioni in inglese.

---

[6] https://www.clarin.si/repository/xmlui/handle/11356/1881 (cons. 08/02/2026). Le annate della testata "Edinost", oggetto del presente studio, sono incluse nel *corpus sPeriodika*.
[7] La selezione è stata condizionata dalle risorse disponibili e dal significativo dispendio di tempo-uomo richiesto dalla fase di trascrizione assistita tramite LLM.
[8] "Edinost", come dichiarato dalla testata, "esce una volta al giorno, tranne le domeniche e i festivi", quindi non è pubblicato di domenica (13 febbraio; 7 e 14 settembre) e nella festività religiosa cattolica della Natività della Beata Vergine Maria che cade l'8 settembre.
[9] Lo sciopero dei fuochisti dell'Österreichischer Lloyd scoppiò tra il 14 e il 15 febbraio 1902 e si estese a mobilitazione generale: una grande manifestazione operaia degenerò in scontri con le forze armate austroungariche provocando vittime e feriti (N.C.C.d.L. – C.G.I.L., 2002).Per la biografia di Ivan Nabergoj si rimanda alla voce biografica sul portale *Slovenska biografija* https://www.slovenska-biografija.si/oseba/sbi383433/.
[10] Per la biografia di Ivan Nabergoj si rimanda alla voce biografica sul portale *Slovenska biografija* https://www.slovenska-biografija.si/oseba/sbi383433/.
[11] https://it.wikipedia.org/wiki/Tranvia_di_Opicina (cons. 08/02/2026).
[12] Il codice sorgente della *pipeline* è disponibile in un *repository* GitHub per garantire la riproducibilità della ricerca (https://github.com/giovannipinna96/ita-slo).
[13] Non è stato possibile recuperare questo dato in relazione alla campagna di digitalizzazione della testata "Edinost" condotta dalla Digitalna knjižnica Slovenije: i metadati dei PDF non contengono informazioni sul software
[14] https://huggingface.co/microsoft/Multilingual-MiniLM-L12-H384 (cons. 08/02/2026).
[15] https://huggingface.co/EMBEDDIA/sloberta (cons. 08/02/2026).
[16] https://huggingface.co/neuraly/bert-base-italiancased-sentiment (cons. 08/02/2026).
[17] https://huggingface.co/classla/xlm-r-parlasent (cons. 08/02/2026).

La distinzione tra modelli per lingua nell'analisi *intra-corpus* e il ricorso a un modello unico sui testi tradotti in inglese nell'analisi *inter-corpora* è stata adottata deliberatamente per evitare comparazioni dirette tra scale di *Sentiment* non calibrate: nell'analisi *intra-corpus* ciascuna testata è confrontata con sé stessa nel tempo mediante un unico modello, mentre la traduzione in inglese funge da *tertium comparationis* nel confronto *inter-corpora*.

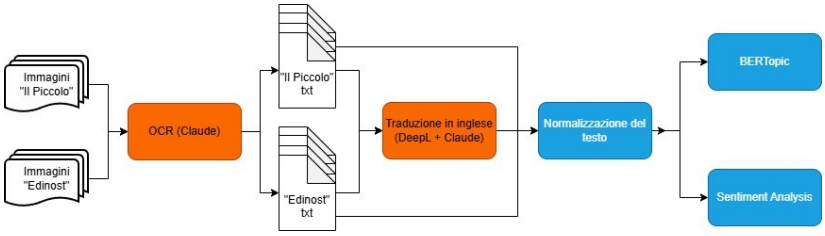


Figura 1. Schema della *pipeline* metodologica: correzione OCR → preprocessing → traduzione EN → *Topic* e *Sentiment Analysis* → analisi *intra/inter-corpora* → sintesi interpretativa.

4. RISULTATI

L'analisi *intra-corpus*, condotta separatamente sulle due testate nelle rispettive lingue originali, rivela strategie discorsive strutturalmente distinte. "Il Piccolo" presenta una dominanza del registro neutro sia in febbraio (media 87,6%) sia in settembre (media 88,3%), con oscillazioni contenute anche in presenza di eventi drammatici. Il 17 febbraio, primo numero uscito dopo la sospensione delle pubblicazioni per lo sciopero, il quotidiano mantiene un tono descrittivo nella ricostruzione degli scontri, con *keyword* istituzionali ("corona", "guardia", "autorità") che riflettono una strategia di autocontrollo linguistico funzionale alla censura preventiva imposta dalla Direzione di Polizia di Vienna.
"Edinost" adotta, invece, una strategia diametralmente opposta. Il *Sentiment* negativo risulta strutturalmente dominante, con una media del 48,8% in febbraio (*range* 36,2–68,9%) e del 40,2% in settembre. Il 19 febbraio, il picco massimo di negatività (68,9%) coincide con una narrazione fortemente empatica degli scontri e verso le vittime. Le *keyword* politico-identitarie ("vlada" governo, "zakon" legge, "slovenski" sloveno) inquadrano fin da subito gli eventi in una cornice di conflitto nazionale. A differenza de "Il Piccolo", "Edinost" non mostra alcuna traiettoria di normalizzazione editoriale nel corso del periodo analizzato.
Il confronto sistematico del *Sentiment* sui testi tradotti in inglese (analisi *inter-corpora*) quantifica le divergenze (vd. Fig. 2). In febbraio, il *Sentiment* negativo medio raggiunge il 28,1% per "Edinost" contro il 20,6% de "Il Piccolo" (+7,5 punti percentuali). La divergenza massima si registra il 28 febbraio (18,8 pp). Le strategie si polarizzano: "Il Piccolo" adotta una linea di normalizzazione attraverso la cronaca mondana e culturale, dando spazio alle celebrazioni del centenario della nascita dello scrittore Victor Hugo ed alla trasposizione del Discorso della Corona italiana[18], mentre "Edinost" mantiene un registro di denuncia senza attenuare la critica alle autorità austriache. In settembre, il differenziale si riduce a 2,7 pp (22,5% vs 19,8%), con una convergenza di circa 5 punti rispetto a febbraio. Questa riduzione suggerisce che, pur mantenendo posizionamenti radicati, le strategie discorsive mostrano una maggiore prossimità in un contesto di tensione sociale meno acuta. I picchi di *Sentiment* positivo rispondono a funzioni narrative diverse: per "Il Piccolo" segnalano notizie di risoluzione istituzionale, per "Edinost" si concentrano sulla commemorazione di figure della comunità slovena (necrologio di Ivan Nabergoj, 10 settembre).

[18] Il 21 febbraio 1902 ricorreva il centenario della nascita di Victor Hugo, celebrato dalla stampa europea. Nello stesso giorno, Vittorio Emanuele III re d'Italia pronunciava il Discorso della Corona all'apertura della legislatura, annunciando riforme sociali: un orizzonte riformista italiano che contrastava implicitamente con la repressione austriaca.

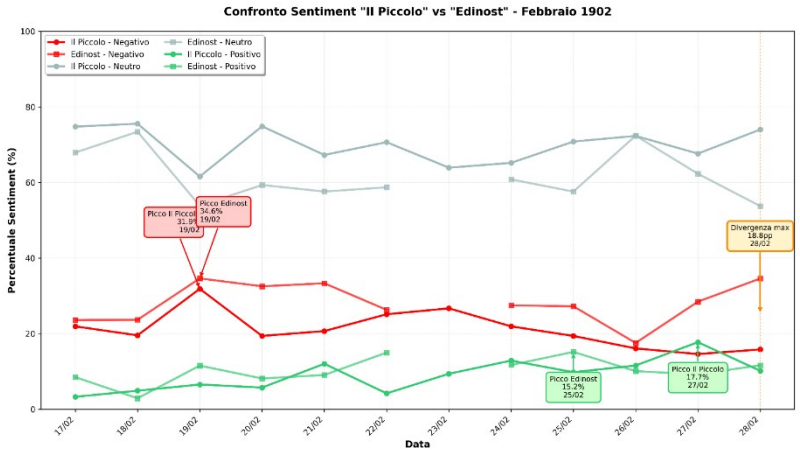


Figura 2. Confronto *Sentiment* "Il Piccolo" vs "Edinost" – Febbraio 1902 (testi in inglese, modello: cardiffnlp/twitter-roberta-base-sentiment-latest). Le linee mostrano l'andamento del *Sentiment* negativo, positivo e neutro per entrambe le testate nel periodo 17–28 febbraio. La divergenza massima si registra il 28 febbraio (18,8 pp)

L'analisi comparativa dei *Topic* estratti mediante *BERTopic* sui *corpora* tradotti in inglese rivela agende tematiche largamente autonome (vd. Fig. 3). In febbraio, l'assenza di *Topic* con convergenza costante (presenti in entrambe le testate per oltre il 70% delle date analizzate) evidenzia come i due giornali operino con logiche di selezione e gerarchizzazione profondamente diverse. "Il Piccolo" domina sui *Topic* istituzionali e cronachistici ("strike", "old", "general", "minister"), mentre "Edinost" prevale sui Topic identitari ("trieste", "italian", "society", "deputy"). La prevalenza di "Edinost" sul Topic "trieste" è particolarmente significativa: pur condividendo il contesto geografico, il giornale sloveno ne parla con frequenza sistematicamente superiore (8 date vs 3), suggerendo una diversa costruzione dell'identità locale.

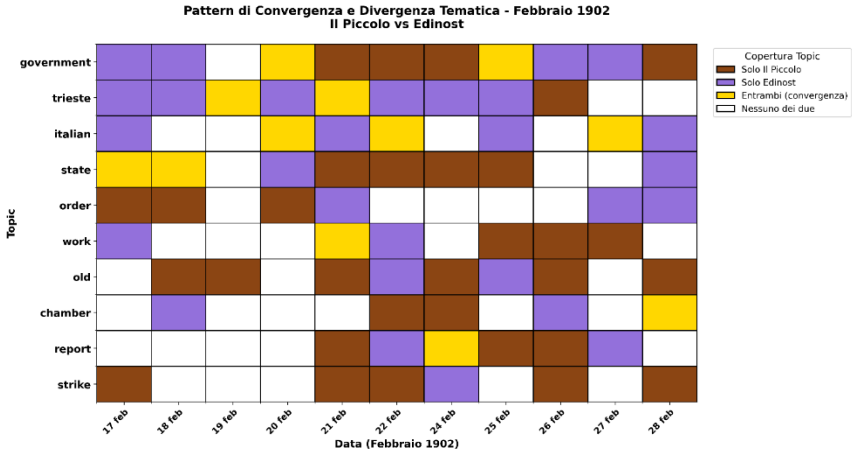


Figura 3. *Heatmap* dei *pattern* di convergenza e divergenza tematica – Febbraio 1902. I colori indicano la copertura esclusiva de "Il Piccolo" (marrone), di "Edinost" (viola) o di entrambi (giallo). L'assenza di convergenza sistematica conferma agende tematiche autonome.

Il *Topic* "emperor" emerge esclusivamente in "Edinost" in corrispondenza della visita dell'imperatore Francesco Giuseppe I nel Litorale austriaco, evento assente ne "Il Piccolo".
Il confronto tra risultati ottenuti nelle lingue originali e sui testi tradotti evidenzia una sostanziale coerenza dei macro-temi, ma con riorganizzazioni semantiche significative. Nei testi sloveni di "Edinost", i *Topic* sono fortemente connotati in senso nazionale e identitario ("narod" nazione, "jezik" lingua, "šola" scuola); nelle versioni inglesi, acquistano visibilità termini più istituzionali e procedurali ("government", "law", "office"). La traduzione preserva i contenuti informativi ma ne attenua la carica ideologica, favorendo una lettura più distaccata. Analogamente, il *Sentiment* negativo di "Edinost" si riduce sensibilmente nei testi inglesi, per effetto sia della normalizzazione lessicale operata dalla traduzione, sia dell'addestramento dei modelli di *Sentiment* su *corpora* contemporanei. Ciò conferma che la traduzione in

una lingua veicolare contemporanea non è un sostituto dell'interpretazione, ma uno strumento metodologico che rende possibile il confronto sistematico evidenziando al contempo i propri limiti.

5. DISCUSSIONI E CONCLUSIONI

L'analisi comparativa condotta dimostra come l'impiego di tecniche di NLP consenta di produrre nuova conoscenza storica attraverso la riorganizzazione semantica delle fonti. "Il Piccolo" ed "Edinost" non restituiscono semplicemente prospettive diverse sugli stessi eventi, ma operano come attori discorsivi distinti: il primo tende a costruire consenso attraverso la selezione, la frammentazione e la normalizzazione delle notizie; il secondo struttura il discorso giornalistico come spazio di argomentazione politica e costruzione identitaria. Le differenze nel *Sentiment*, nei *Topic* e nel *framing* riflettono strategie editoriali strutturali, non reazioni contingenti.
Sul piano metodologico, la *pipeline* proposta – basata su correzione OCR assistita da LLM, traduzione automatica come *medium* linguistico, *Topic* e *Sentiment Analysis* con modelli specifici per lingua – si è rivelata efficace per l'analisi di *corpora* storici multilingui caratterizzati da forte rumore OCR e scarsità di risorse annotate. Il lavoro, che si configura come primo studio esplorativo, è orientato alla validazione metodologica di una *pipeline* applicata a un caso di studio circoscritto. La dimensione del *corpus* (50 fascicoli, corrispondenti a un numero di unità documentali significativamente maggiore dopo la segmentazione) richiede cautela nell'interpretazione dei risultati su scala generale e, nello specifico, sulla stabilità statistica dei *topic* estratti, la cui coerenza è stata validata per via qualitativa tramite rilettura esperta. I modelli di *Sentiment*, addestrati su testi contemporanei, possono sottostimare la carica emotiva di registri giornalistici storici; la non piena calibrazione delle scale tra modelli diversi nell'analisi *intra-corpus* è stata mitigata, nel confronto *inter-corpora*, dal ricorso a un modello unico applicato ai soli testi tradotti in inglese, che funge da *tertium comparationis*, introducendo tuttavia una parziale attenuazione delle specificità culturali e retoriche originali.
In prospettiva, il progetto si colloca nella cornice della *digital public history*, proponendo il *distant reading* dei periodici storici triestini come strumento di mediazione capace di rendere osservabili, in forma comparabile, anche mediante tecniche di *data visualization* (che corroborano di senso, anticipandone con immediatezza i contenuti informativi) narrazioni altrimenti difficilmente accessibili. In una città plurilingue come Trieste, dove comunità linguistiche diverse hanno storicamente mantenuto narrazioni parallele, un progetto di *Edicola digitale della Storia* mira a ricomporre un mosaico narrativo frammentato, le cui tessere sono state finora integrate solo parzialmente in contesti storiografici o museali – pur articolati e rigorosi – inevitabilmente orientati da specifiche prospettive culturali e linguistiche[19]. Questo lavoro pone, dunque, le premesse per un'interfaccia web di interrogazione semantica dei periodici storici triestini, orientata tanto alla ricerca quanto alla divulgazione culturale, capace di raggiungere un pubblico più ampio rispetto alla sola comunità accademica. Parallelamente, la sistematizzazione dei testi de "Il Piccolo" apre alla possibilità di estrarre un *corpus* in lingua italiana non letteraria, di sicuro interesse per studi linguistici e applicazioni avanzate di NLP.

RINGRAZIAMENTI

Si ringrazia la Digitalna knjižnica Slovenije – dLib.si per aver messo a disposizione i file PDF con *layer* OCR dell'intera collezione del periodico "Edinost", risorsa indispensabile per la realizzazione di questo studio. Si ringraziano inoltre il prof. Eric Medvet e il prof. Andrea De Lorenzo del DIA – Dipartimento di Ingegneria e Architettura dell'Università degli Studi di Trieste per il supporto e la supervisione nelle fasi di sviluppo del progetto, nonché il prof. Alfio Ferrara, docente di *Natural Language Processing* e *tutor* del *project work* nell'ambito del Master in Digital Humanities dell'Università degli Studi di Milano.

BIBLIOGRAFIA

Borenstein, N., da Silva Perez, N., & Augenstein, I. (2023). Multilingual Event Extraction from Historical Newspaper Advertisements. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 10430–10444.

[19] https://projet.biblissima.fr/ (cons. 05/02/2026)

Cafagna, M., De Mattei, L., & Nissim, M. (2020). Embeddings-based detection of word use variation in Italian newspapers. Italian Journal of Computational Linguistics, 6(2), 9–22. https://doi.org/10.4000/ijcol.703.

Dobranić, F., Evkoski, B., Ljubešić, N. (2023, December 20). Corpus of Slovenian periodicals (1771–1914) sPeriodika 1.0 [Corpus]. Slovenian language resource repository – CLARIN.SI. http://hdl.handle.net/11356/1881

Ehrmann, M., Romanello, M., Flückiger, A., Clematide, S., Clivaz, P., & Kaplan, F. (2020). Impresso: A Large-Scale Digital Humanities Project for the Processing of Historical Newspapers. Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020), 958–968.

Gorjanc, V., Pretnar Žagar, A., Dobranić, F., & Fišer, D. (2025). Accessing Historical Periodicals: Newspaper Discourse on Slovene Language [Presentazione]. ADHO Digital Humanities Conference 2025. https://doi.org/10.5281/zenodo.16087978.

Ludovisi, M. (2025). Il progetto ArPeR. Per un Archivio dei periodici romaneschi. In S. Rebora, M. Rospocher, & S. Bazzaco (Eds.), Diversità, Equità e Inclusione: Sfide e Opportunità per l'Informatica Umanistica nell'Era dell'Intelligenza Artificiale. Proceedings del XIV Convegno Annuale AIUCD2025, 222–227. AlmaDL, University of Bologna Digital Library. https://doi.org/10.6092/unibo/amsacta/8380.

Marjanen, J., Zosa, E., Hengchen, S., Pivovarova, L., & Tolonen, M. (2020). Topic Modelling Discourse Dynamics in Historical Newspapers. In S. Reinsone, I. Skadiņa, J. Daugavietis, & A. Baklāne (Eds.), DHN 2020, Digital Humanities in the Nordic Countries 2020. Post-Proceedings of the 5th Conference Digital Humanities in the Nordic Countries, 63–77. http://nbn-resolving.de/urn:nbn:de:0074-2865-8.

Moretti, F. (2000). Conjectures on world literature. New Left Review, 1, 54–68.

Moretti, F. (2013). Distant Reading. Verso.

N.C.C.d.L. – C.G.I.L. (2002). 1902-2002: La lotta dei fuochisti del Lloyd Austriaco. Trieste: Istituto "Livio Saranz".

Piryani, B., Mozafari, J., & Jatowt, A. (2024). ChroniclingAmericaQA: A Large-Scale Question Answering Dataset based on Historical American Newspaper Pages. In SIGIR '24: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2038–2048. https://doi.org/10.1145/3626772.3657891.

Thomas, A., Gaizauskas, R., & Lu, H. (2024). Leveraging LLMs for Post-OCR Correction of Historical Newspapers. In Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA), 116–121. European Language Resources Association.

Literary Tropes as Prescriptive Devices: A Case Study from BookTok

SILVIA CAPPA[1], MARIANGELA GIGLIO[2], AND ANNA SOFIA LIPPOLIS[3]

[1] CNR - INSTITUTE OF COGNITIVE SCIENCES AND TECHNOLOGIES, BOLOGNA, ITALY - SILVIACAPPA@CNR.IT
[2] UNIVERSITY OF BOLOGNA, ITALY - MARIANGELA.GIGLIO2@UNIBO.IT
[3] UNIVERSITY OF BOLOGNA, ITALY - ANNASOFIA.LIPPOLIS2@UNIBO.IT

ABSTRACT (ENGLISH)

This paper examines the changing function of literary tropes within platform-based social reading environments, suggesting that in communities such as BookTok (a subcommunity of TikTok), they operate primarily as prescriptive devices that orient reading practices, rather than as descriptive labels applied retrospectively. The discussion is grounded in an observational case study centred on the *enemies to lovers* trope, based on a curated sample of TikTok content. The analysis does not aim to produce experimentally generalisable results; rather, it uses empirical observation as a heuristic lens to explore how trope attribution structures access to texts, shapes readers' expectations, and contributes to the establishment of normative frameworks of evaluation prior to the act of reading.
By situating trope usage within the platform's algorithmic, commercial, and communal dynamics, the article highlights a possible shift toward a more prescriptive organisation of literary categorisation: tropes tend to function less as interpretive tools and more as promises of experience that guide selection and anticipation, while also influencing modes of evaluation. Rather than constituting a fully demonstrable transformation, this shift suggests a broader tendency that invites further reflection on contemporary literary circulation, the role of the reader, and the negotiation of interpretive authority within digital social reading cultures.
Keywords: literary theory; media and cultural studies; New Media; platform studies; digital social reading

ABSTRACT (ITALIANO)

I tropi letterari come dispositivi prescrittivi: un caso di studio da BookTok
Il contributo indaga il mutamento della funzione dei tropi letterari negli ambienti di lettura sociale basati su piattaforme, suggerendo che in comunità come BookTok (sottocomunità di TikTok) essi operino prevalentemente come dispositivi prescrittivi che orientano le pratiche di lettura, più che come etichette descrittive applicate a posteriori. La discussione si avvale di un caso di studio osservativo incentrato sul trope *enemies to lovers,* costruito a partire da un campione curato di contenuti di TikTok. L'analisi non mira a produrre risultati sperimentali generalizzabili, ma utilizza l'osservazione empirica come lente euristica per esaminare in che modo l'attribuzione dei tropi strutturi l'accesso ai testi, modelli le aspettative dei lettori e contribuisca a istituire quadri normativi di valutazione già prima dell'atto di lettura.
Collocando l'uso dei tropi all'interno delle dinamiche algoritmiche, commerciali e comunitarie della piattaforma, il contributo evidenzia una possibile riorganizzazione in senso prescrittivo della categorizzazione letteraria: i tropi tendono a funzionare meno come strumenti interpretativi e più come promesse di esperienza che orientano la selezione e l'anticipazione, incidendo anche sulle modalità di valutazione. Questo spostamento suggerisce una tendenza che invita a ripensare la circolazione letteraria contemporanea, il ruolo del lettore e le forme di negoziazione dell'autorità interpretativa nelle culture digitali della lettura sociale.
Parole chiave: teoria letteraria; media and cultural studies; new media; platform studies; digital social reading

1. INTRODUCTION

In recent years, TikTok has emerged as a significant space for book discovery and reader interaction, characterised by the intensive use of literary tropes in recommendation and discussion practices (Rozaki, 2023; Guehring, 2023). Within this platform-based reading environment, tropes appear to function not only as descriptive labels applied after reading, but also as devices that orient readers toward specific narrative experiences before the act of reading takes place.

This paper explores how, on TikTok, trope attribution may operate as a prescriptive device that precedes reading and structures how texts are anticipated and accessed. Rather than simply naming what a book is, tropes increasingly communicate the kind of experience a book is expected to deliver, suggesting a possible reconfiguration in the temporal logic of literary categorisation, from post-reading interpretation toward forms of pre-reading orientation.
Focusing on the *enemies to lovers* trope as a case study inside the algorithmically organised niche of TikTok, BookTok[1], the paper examines how trope attribution pre-organizes reading experiences and establishes normative frameworks of evaluation within algorithmic reading communities. In this context, the trope can be understood as an access filter and an expectation-setting mechanism, as well as a point of reference against which reading experiences are negotiated, so that trope attribution does not merely accompany literary circulation but participates in shaping its conditions.
By conceptualizing this phenomenon as a prescriptive turn in literary categorization, the study introduces the notion of tropification to describe the process through which tropes become predictive and normative devices embedded in platform logics.
By situating BookTok within broader processes of social reading and algorithmic mediation, the paper also argues for greater attention to platform-based reading communities within literary studies, insofar as they offer a valuable site for observing how categorisation practices, reader engagement, and forms of literary value are negotiated and transformed in contemporary digital environments.

2. THEORETICAL FRAMEWORK: TROPES AS DEVICES

In this study, literary tropes are understood as verbalized literary devices articulated through tags and labels in social platform environments concerned with books. Drawing on contemporary trope taxonomies such as *The Trope Thesaurus*, the collaborative wiki *TVTropes*, and *TropeTrove*, the analysis observes a semantic expansion of the term *trope* beyond its rhetorical origins: as Börzsei argues in her analysis of *Television Tropes and Idioms*, these taxonomies constitute a technological continuation of archetypal literary criticism, translating recurring narrative elements into accessible and often ironic classificatory forms. Unlike classical archetypes, such as the Proppian hero or antagonist, contemporary tropes focus on highly specific narrative dynamics between characters, including granular patterns such as in the *forced proximity* trope or the *who-did-this-to-you*.
Tropes can be understood not only as semantic descriptors but as operational categories that mediate visibility, discoverability, and reader orientation within platform-based reading cultures. More specifically, the analysis focuses on what tropes do in circulation: how they classify texts, frame expectations, and guide acts of recommendation, selection, and evaluation.
Within this framework, tropes operate along two interrelated modalities, descriptive and prescriptive: in their descriptive function, they serve as tools of literary categorization and interpretation, as when readers describe *Pride and Prejudice* as an instance of *enemies to lovers*, performing an interpretive act through a trope-based framework on which subsequent evaluations often hinge. In their prescriptive function, however, tropes orient access to texts, reader expectations, and circulation patterns according to the narrative experience they promise, increasingly signaling not what a book is but how it should be read, why it should be read, and for whom it is intended. These modalities are not strictly separable but function as loose analytical categories that may coexist within the same usage; what matters, however, is that they can be meaningfully distinguished at the level of intention, even when they co-occur in practice.
While tropes resemble genres in their classificatory role, they operate with significantly greater granularity, extending to micro-tropes that refer to specific scenes, phrases, or narrative moments, and thus function not only as categories but as devices that actively structure reading trajectories.
This prescriptive capacity emerges historically from trope practices developed in fanfiction platforms, where tags functioned as organizational and navigational tools facilitating findability, orienting reader access, and providing content warnings within largely non-commercial, non-algorithmic environments. Rather than enacting a Barthesian (1967) "death of the author", such platforms privileged authorial self-curation, as writers described their own works to guide reader expectations and circulation, retaining

[1] BookTok is a hashtag-driven subcommunity on TikTok (#BookTok) centred on books and reading, where users create short-form video content ranging from reviews and recommendations to emotional reactions, "book hauls", memes, and scene recreations; in recent years, it has grown into a significant force shaping both publishing dynamics and contemporary reading culture (Rascati & Grantham, 2025).

interpretive authority within shared narrative worlds, as also suggested by Palmer's (2021) analysis of tropes in popular romance fiction.
When these practices migrate from fanfiction platforms to proprietary social networks such as TikTok, their orientation toward consumption is preserved but their structural conditions change. Within algorithmically driven environments structured by personalisation and viral promotion, trope attribution becomes entangled with influencer practices, platform visibility, and publishing interests, intensifying what online communities refer to as *tropification*. In this process, tropes increasingly operate as predictive and normative devices, shaping not only how books are marketed but, in some cases, how they are conceived and produced, as the granularity of trope specifications enables unprecedented customization of reading experiences that aligns seamlessly with publishing industry strategies of market segmentation. Tropification thus represents the full realization of tropes as commercial-prescriptive devices, inverting the traditional relationship between text and category: rather than categories emerging from texts, texts increasingly emerge from categories. This shift raises critical concerns about the algorithmic-commercial restructuring of literary production itself, as recent commentary suggests that the colonization of fanfiction worlds by the publishing industry risks simplifying literature through formulaic production optimised for trope-based visibility.
At the same time, the trajectory from fanfiction to BookTok entails a reconfiguration of interpretive authority: whereas fanfiction authors retained privileged interpretive control through self-curation, BookTok operates within a diffuse field in which influencers, publishers, and community consensus collectively negotiate trope attribution. Yet this does not amount to a return to reader-centered interpretation; rather, it constitutes a form of algorithmic-commercial interpretation, where authority is distributed across a network of users, platforms, and market actors, aligned toward prescriptive rather than descriptive functions. For this reason, trope attribution on BookTok should be read not only as an interpretive practice but as a socio-technical one, emerging from the interaction between human judgment, platform affordances, and commercial incentives.
The trope becomes a privileged site where interpretation, recommendation, and marketability converge. As Low and colleagues (2025) describe through the notion of "algorithmic imaginings", interpretive authority on BookTok emerges at the intersection of user practices, platform affordances, and commercial imperatives, positioning trope attribution as a central mechanism through which literary categorization, circulation, and agency are being reshaped in contemporary platform-based environments.

3. METHODOLOGICAL APPROACH

To test whether the hypothesis that tropes function prescriptively holds in observable practice, we conduct a qualitative analysis supported by empirical data drawn from TikTok. This platform provides a valuable site for examining tropes *in situ*, since trope-based discourse emerges within a socio-technical environment shaped by recommendation systems and genre-centred communities. The central research question is how does attributing a trope operate as a prescriptive act that mediates access to and interpretation of a literary text within BookTok discourse. In addressing this question, we prioritise the trope's pragmatic function, its role in organising description (whether central or ancillary) and, above all, in prescribing a readerly experience. Accordingly, rather than defining what a trope is in semantic or narratological terms, or analysing the specific content it is taken to promise, the study focuses on what tropes do when they are applied to literary works: how they orient access, structure expectations, and guide interpretation. Methodologically, the study combines a documented and methodologically explicit guided observation with qualitative analysis informed by minimal quantitative contextualisation. The resulting design can be framed as an empirically grounded qualitative case study.
For the empirical analysis, we selected a single trope as a case study: *enemies to lovers*, which describes a narrative dynamic in which two characters begin as adversaries and subsequently develop a romantic relationship. This trope was chosen because it is consistently identified in editorial sources and meta-discussions on BookTok as one of the most popular and widely circulated tropes within the community. It is repeatedly cited in compilations of most loved tropes and in trend reports, and it occupies a prominent place in both online and physical book retail environments, where it often appears as a dedicated

category. Its use is particularly pervasive in recommendation practices and discussions related to the romance and romantasy genres.[2] (Tristanty & Khusyairi, 2025)

The unit of analysis consists of BookTok content—videos, images, and slideshows—that explicitly discuss or employ the trope, with particular attention to their expressive and discursive functions. The analysis focuses primarily on recommendation and discussion videos, in which creators use the trope to suggest a book to potential readers who have not yet read it, with the practical aim of guiding choice and framing expectations. It is in these specific contexts where the trope functions as a shorthand for interpretation and access. Additionally, comment sections are treated as a complementary site of analysis, as they often host discussions about the applicability of the trope to specific books and, by extension, debates over interpretation.

Data were collected manually, without automated scraping. This choice reflects both the qualitative focus of the study, centred on observing a specific mechanism through which tropes operate as interpretive devices, and legal and ethical constraints. TikTok prohibits unauthorised automated scraping; by contrast, the manual observation of publicly available content for academic research is generally consistent with EU fair-use exceptions for research and criticism, provided that creators are not profiled. Accordingly, creators were anonymised and treated not as individuals but as discursive instances, following established practice in EU-based digital humanities and social media studies.

Methodologically, the study acknowledges a structural challenge in social media research: TikTok's algorithm is opaque, adaptive, and strongly personalised, making fully controlled datasets unattainable. This is not a limitation to be solved but a condition of the field (Flores, 2023; Martens et al., 2022). As a result, the study does not aim for representativeness or algorithmic replicability; instead, it ensures procedural replicability by documenting a controlled data collection protocol, in line with standards in digital humanities and digital social reading research (Pianzola et al., 2020).

To capture a sample that is meaningful within TikTok's platform logic, we adopted a mixed strategy combining search-based visibility and recommendation-based discovery. Data collection was conducted through a newly created "clean" TikTok account on the mobile app, with no personal information provided beyond a birth year (2000). First, we performed a search query ("booktok enemies to lovers") using the default relevance ranking and saved the first 50 videos. To trigger controlled algorithmic training, we then liked the first 10 videos and followed their creators. Second, we conducted a 90-minute scrolling session on the *For You Page*—the algorithmically personalised feed that curates content for each user based on inferred interests and engagement patterns—and saved the first 50 recommended videos explicitly related to BookTok and the enemies to lovers trope, including content not easily retrievable through hashtags or captions. The platform's For You Page algorithm does not merely surface popular content but actively shapes what counts as a "real" instance of a trope through amplification patterns that reward certain definitional framings over others.

The final corpus (N = 100) was compiled from these two streams. For each item, we recorded minimal metadata including upload date, engagement metrics (likes, comments, shares), hashtags, caption text, and relevant text displayed within the video or slideshow.

We developed an analytical framework designed to trace whether trope attribution precedes or follows reading, and whether it actively orients the reading experience. This approach responds to concerns about algorithmic imaginaries and critical digital literacy on BookTok (Low et al., 2025). Building on recent BookTok scholarship (Rozaki, 2023; Guehring, 2023), we implemented a typological coding scheme based on primary creator intent. Videos were classified into two macro-categories: "recommendation", where creators explicitly recommend books framed as enemies to lovers, and "interpretation", encompassing all other content in which the trope is used primarily to describe, evaluate, or discuss a work, with recommendation as secondary or absent. In practice, however, many videos exhibit elements of both categories, as recommendation and interpretation frequently coexist within the same content. Nonetheless, the format of recommendation videos tends to make their primary intent more legible, for instance through explicit book stacks, ranking structures, or self-promotional framing, whereas interpretive videos foreground discussion and evaluation. Despite this overlap, each

[2] To give a sense of scale and contextualize the trope visibility and relevance within the platform, as of January 29, 2026, the hashtag #BookTok accounts for approximately 72.9 million posts. The enemies-to-lovers trope, considering multiple spellings and related tags (e.g. enemiestolovers, enemiestolover, enemiestoloversbooks, enemiestoloverstrope), accounts for nearly 6.12 million posts.

video can still be analysed in terms of a dominant pragmatic orientation, understood as its primary communicative purpose.
Within these categories, we identified recurring subtypes aligned with platform trends observed during collection, including: "normative evaluation" (defining what the trope is or is not), text-based excerpts, emotional reaction videos, POV formats, TBR lists, structure-focused discussions (how the trope is implemented narratively), wrap-ups, self-promotion.
Content distribution shows that more than half of the corpus (N = 53) falls into the recommendation category, suggesting that the trope's prescriptive function is central in actual BookTok usage. Normative evaluation emerges as the most frequent subtype, especially within recommendations, often through formulaic expressions such as "this is a real/true enemies-to-lovers". In interpretive videos, normative evaluation tends to shift toward defining the trope itself and is frequently paired with structure-focused discussions. These discussions consistently converge on the claim that the "enemy" phase must occupy a substantial portion of the narrative to qualify as a "true" instance of the trope. Comment sections further reinforce this normative negotiation, whereas in more conventional book reviews the trope appears less central, plausibly because post-reading analysis relies less on shorthand categorisation.
Our observations suggest that, within the BookTok corpus analysed, the enemies to lovers trope often functions as more than a descriptive label or a retrospective category. In many recommendation-oriented videos, it operates as a prospective frame: it helps users decide whether a book is worth reading, and it pre-structures the expectations through which the narrative is likely to be approached.
While this does not imply that reading becomes fully determined by trope discourse, it indicates that trope attribution can precede reading and shape access to the text, orienting the reading act.
A second recurring pattern concerns the normative dimension of trope discourse. The prevalence of "normative evaluation" formats, particularly the repeated distinction between "real/true" and insufficient instances, indicates that trope attribution on BookTok frequently involves boundary-setting. These negotiations are visible both in videos and in comment sections, where users collectively refine what should count as a legitimate instance of the trope. Notably, such discussions often unfold in contexts where the book is being considered rather than reviewed, suggesting that normativity operates as a prescriptive criterion for selection, not only as an interpretive conclusion.
Finally, our data provide limited but consistent evidence that trope attribution functions as a shorthand for narrative structure. Discussions of what counts as "true" enemies to lovers repeatedly converge on temporal and structural expectations (e.g., the "enemy" phase must last long enough). This suggests that the trope is frequently understood not as a thematic content descriptor but as a structural promise about pacing and plot dynamics. In this sense, BookTok trope discourse is best understood through the interplay of description and prescription: tropes describe texts in compressed form, but they also prescribe a mode of access and a horizon of interpretation that is negotiated collectively and circulated through platform-specific formats.

4. CONCLUSION AND FUTURE WORK

This study examined how trope attribution operates within TikTok's BookTok community through an empirically grounded qualitative case study of the enemies to lovers trope. Across the analysed corpus, trope usage emerges most prominently in recommendation-oriented content, where it functions not merely as descriptive shorthand but as a practical device for guiding selection and framing expectations. In such contexts, tropes are frequently mobilised prior to reading, less to account for interpretive experience than to organise it in advance, shaping how prospective readers approach the text. Rather than serving as retrospective interpretive vocabulary, trope attribution thus operates as a criterion for inclusion, exclusion, and expectation-setting, with broader implications for understanding contemporary literary circulation and digital social reading practices (Pianzola et al., 2020).
Taken together, these patterns suggest a shift in how books are encountered in digital social reading environments: from exploratory discovery toward a more expectation-driven mode of selection, in which anticipated affective and narrative payoffs are foregrounded. Within this configuration, trope attribution appears to stabilise shared interpretive horizons in advance, subtly constraining the range of possible readings while facilitating rapid alignment within the community. This dynamic also points to the often underexamined role of algorithmic infrastructures in shaping platform-based reading cultures, insofar as such pre-structured signals of preference and expectation become readily legible to recommendation systems, reinforcing their circulation and visibility.

Methodologically, the study contributes a documented protocol for manual data collection and a typological coding framework for videos and comments. This framework supports procedural replicability and can be adapted to further research on trope-based social reading, complementing existing approaches to TikTok content and BookTok subcommunities (Troncone, 2025; Flores, 2023). At the same time, the scope of the findings remains limited, as the analysis focuses on a single trope, a single platform, and a dataset shaped by algorithmic personalisation that cannot be fully controlled, and therefore does not claim statistical representativeness or broad generalisability.

We think that rather than diminishing the significance of the observed dynamics, these constraints help to situate them, indicating that tropification extends beyond a surface feature of social media discourse, shaping both literary categorisation and reader engagement. As tropes move from fanfiction's reader-driven, non-commercial spaces into TikTok's algorithmically optimised and commercially aligned environment, they shift from tools of self-curation toward devices that guide visibility, selection, and expectation.

Several lines of inquiry emerge from these findings, positioning tropes as a productive lens through which to examine not only changing reading practices but also the platform-driven reorganisation of literary value and visibility.

A first direction concerns a broader account of tropes as both descriptive and prescriptive devices. Extending the analysis beyond a single case would make it possible to map which tropes gain prominence within specific digital literary environments such as BookTok, and to assess how their functions vary across contexts. In particular, this would clarify the extent to which platform-specific affordances shape not just the circulation of tropes, but their relative stability and uptake.

A second line of research would develop the notion of tropification more systematically, examining its effects not only on reader engagement but also on publishing dynamics. This includes assessing whether and how trope-based framing contributes to commercial performance, as well as tracing its impact on editorial strategies. At stake here is not only the scale of production but also its orientation, insofar as the collective negotiation of tropes may reconfigure assumptions about authorship and encourage forms of writing and dissemination that depart from more traditional literary models.

Finally, further work is needed to understand how these dynamics reshape reading practices over time. In particular, analysing the implications of expectation-driven or promise-based reading could shed light on its effects on interpretive variability and on the conditions under which literary meaning is produced and stabilised within digital environments. More broadly, situating tropification within literary-critical inquiry would contribute to a more comprehensive account of how reading is structured, mediated, and transformed in platform-based cultures.

Taken together, these directions suggest that tropification offers a useful perspective on how reading emerges through the ongoing negotiation of expectations within digitally mediated environments.

REFERENCES

Barthes, R. (1977). The death of the author. In Image–Music–Text: Essays selected and translated by Stephen Heath (pp. 142–148). New York: Hill and Wang. (Original work published 1967)

Flores, Z. (2023). Investigating subcommunities in BookTok: An analysis using hashtags and videos. https://doi.org/10.17615/dsab-e438

Guehring, K. (2023). Algorithmic recommendation and book discovery on TikTok (Master's thesis). Universitetet i Bergen.

Low, B., Ehret, C., & Hagh, A. (2025). Algorithmic imaginings and critical digital literacy on #BookTok. New Media & Society, 27(4), 2336–2353. https://doi.org/10.1177/14614448231206466

Matthews, A., Ashman, J., Heffernan, M., Hogan, P., Guy, A., Stewart, H., Stanley, K., Cothren, A., & Duffield, E. (2025). Editorial: Degrees of Love and Trope Actually. TEXT. https://doi.org/10.52086/001c.142608.

Palmer, A. (2021). Deconstructing tropes in popular romance fiction (MAIS final paper). Athabasca University. https://alyssalinnpalmer.com/wp-content/uploads/2023/05/Deconstructing-Tropes-in-Popular-Romance-Fiction-Alyssa-PalmerID3513585-MAIS601Final.pdf

Pianzola, F. (2021). Digital Social Reading: Sharing Fiction in the 21st Century. MIT Press. https://doi.org/10.1162/ba67f642.a0d97dee

Pianzola, F., Rebora, S., & Lauer, G. (2020). Wattpad as a resource for literary studies: Quantitative and qualitative examples of the importance of digital social reading and readers' comments in the margins. PLoS ONE, 15(1), e0226708. https://doi.org/10.1371/journal.pone.0226708

Rascati, D., & Grantham, S. (2025). Bestseller best practices: A digital ethnography of #BookTok content on TikTok. Journal of Social Media Research, 2, 139–154. https://doi.org/10.29329/jsomer.26

Rozaki, E. (2023). BookTok and the transformation of reading cultures: Literary recommendation and reader interaction on TikTok (Master's thesis). Utrecht University.

Stanko, D. V. (2023). METHODOLOGY OF INVESTIGATION OF THE ENGLISH FANFICTION. Transcarpathian Philological Studies, 2(29), 147–151. https://doi.org/10.32782/tps2663-4880/2023.29.2.28

Tristanty, A., & Khusyairi, J. (2025). Mass-produced romance: BookTok society and the homogenisation of literary culture. Jurnal Studi Komunikasi (Indonesian Journal of Communications Studies). https://doi.org/10.25139/jsk.v9i1.9576.

Woodall, T. (2025). Romancing the Reader: An Analysis of Tropes in the Romance Genre [Master of Arts, Chapman University]. https://doi.org/10.36837/chapman.000668

[4/B] Musei e accesso digitale

Rigenerare un patrimonio desueto con il phygital: la Casa Museo Bartolo Longo

MAURO DE BARI
DIPARTIMENTO DI RICERCA E INNOVAZIONE UMANISTICA DIRIUM – UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO, ITALIA – MAURO.DEBARI@UNIBA.IT

ABSTRACT (ITALIANO)

Il contributo presenta un caso di studio sulla Casa Museo Bartolo Longo (Latiano), dove uno spazio in disuso viene trasformato in percorso museale tramite un impianto phygital che combina pannellistica autoportante e reversibile, adottata per vincoli conservativi, contenuti digitali a livelli, un elemento immersivo e un avatar basato su IA che guida il visitatore in prima persona. La tesi è che, in contesti di patrimonio desueto/minor, il passaggio a una mediazione ibrida non introduca solo nuovi canali, ma riorganizzi gerarchie e autorità del racconto: in particolare, l'avatar può aumentare continuità e accessibilità, ma anche produrre un "surplus di autenticità", spostando la fiducia dal piano delle fonti al piano della presenza digitale. Nel caso di Bartolo Longo, tale rischio è accentuato dalla densità simbolica della figura, insieme locale, religiosa, civica e celebrativa: l'avatar, parlando in prima persona, può rendere meno percepibile la distanza tra documento, interpretazione curatoriale e messa in scena. Il paper discute questo nodo in chiave critico-antropologica e propone contromisure progettuali per mantenere la mediazione riconoscibile e verificabile: disclosure dell'uso di IA, segnali di provenienza e rimandi alle fonti, regole di scrittura per limitare la prima persona a contenuti documentabili, e scelte di design che evitino effetti testimoniali. In chiusura si propone una verifica leggera in loco, basata su osservazione e, opzionalmente, brevi riscontri dei visitatori e/o log d'uso, per valutare profondità di fruizione e uso dei rimandi alle fonti, e per capire se l'avatar è percepito come guida o come testimone.
Parole chiave: patrimonio desueto; museo phygital; avatar IA; autenticità; provenienza delle fonti

ABSTRACT (ENGLISH)

Revitalising an uncoventional heritage through phygital approaches: the Casa Museo Bartolo Longo. This paper presents a case study on the Casa Museo Bartolo Longo (Latiano, Italy), where a disused space is turned into a museum route through a phygital setup combining a reversible freestanding display system, adopted due to conservation constraints, multi-level digital content, an immersive element, and an AI-based avatar that guides visitors in the first person. The claim is that, in contexts of unconventional/minor heritage, the shift to hybrid mediation does not merely introduce new channels, but reorganises the hierarchies and authority of the narrative: in particular, the avatar can increase continuity and accessibility, but it can also produce a "surplus of authenticity", shifting trust from the level of sources to that of digital presence. In the case of Bartolo Longo, this risk is heightened by the symbolic density of the figure, at once local, religious, civic and celebratory: by speaking in the first person, the avatar may make the distance between document, curatorial interpretation and staging less perceptible. The paper discusses this issue from a critical anthropological perspective and proposes design countermeasures to keep mediation recognisable and verifiable: AI disclosure, provenance cues and links to sources, writing rules that limit first-person narration to documentable content, and design choices that avoid testimonial effects. Finally, it proposes a lightweight on-site evaluation, based on observation and, optionally, brief visitor feedback and/or usage logs, to assess depth of engagement and use of source references, and to understand whether the avatar is perceived as a guide or as a witness.
Keywords: unconventional heritage; phygital museum; AI avatar; authenticity; source provenance

1. INTRODUZIONE

Molti contesti culturali a forte densità di significati locali restano fuori dai circuiti della valorizzazione "maggiore": luoghi e beni spesso poco noti, talvolta poco accessibili o non pienamente documentati, esposti all'incuria e alla perdita di attrattività rappresentano questa fattispecie. Tale pratica, presente in una classificazione "tradizionale", è stata trasferita anche nel digitale, privilegiando l'estetica al di sopra

del contesto. Ne consegue la generazione di una costellazione di patrimonio meno "apprezzato", descritta come *minor heritage*: patrimoni "minori" per l'appunto, non per valore intrinseco, ma per visibilità e riconoscimento, con un divario tra la percezione pubblica e il valore effettivo (Torre & Christodoulou, 2021).
In questa prospettiva, l'autore usa, per continuità con precedenti lavori sul patrimonio culturale digitale, sulla fruizione phygital e sull'interazione con beni culturali non convenzionali, il termine "patrimonio desueto" come sinonimo operativo: un patrimonio composto da beni ricchi di significato e di valore intrinseco, ma indeboliti dal disuso, dalla marginalità e dalla scarsa capacità di attivare pratiche continuative di cura (De Bari, 2023; De Bari, 2024). In altre parole, la "desuetudine" non è una qualità del bene, ma una condizione relazionale: riguarda i legami che si allentano tra luogo, comunità e pratiche di fruizione, fino a rendere il patrimonio culturalmente "silenzioso", benché ancora densissimo di senso.
Le strategie phygital, intese come integrazione tra le dimensioni fisica e digitale in un *continuum* esperienziale, sono state proposte come leva per riattivare patrimoni di prossimità (Nofal et al., 2017; Lo Turco & Giovannini, 2020; Lupo, 2021). In questa direzione si colloca anche l'investimento europeo nella transizione digitale del patrimonio e delle industrie culturali e creative (ICC), che sostiene esplicitamente esperienze "*physical–digital*" attraverso progetti di Horizon Europe, come *AMPLIFY – PHYGITAL Solutions for the Cultural and Creative Industries* (Grant Agreement 101177413) (European Commission, 2025).
Più nello specifico, questo approccio trova particolare riscontro nel settore delle *house museums*. Si ritiene che la tecnologia ibrida possa amplificare l'esperienza utente, ma anche spostare la percezione dell'autenticità sull'interfaccia e sul racconto (Miano & Borsotti, 2023). Tale direzionamento è particolarmente rilevante quando il digitale non si limita a "mostrare" o "spiegare", ma introduce un soggetto enunciante: una figura che prende la parola, scandendo tempi e soglie della visita, producendo un effetto di prossimità e orientando implicitamente ciò che merita attenzione. In termini antropologici, la visita non è più soltanto una consultazione di contenuti, ma un incontro con una presenza mediata che può attivare processi di identificazione, riconoscimento o rifiuto, soprattutto in contesti in cui il patrimonio è legato a memorie locali e appartenenze (Falk & Dierking, 2016).
Il presente contributo propone, come caso di studio, una recente esperienza realizzata da una startup barese, CRHACK LAB BARI – Officine di Cultura Digitale s.r.l.[1], volta a ridefinire, in una nuova prospettiva comunitaria, gli spazi in disuso di quella che è stata inaugurata come Casa Museo Bartolo Longo a Latiano[2]. Il caso riportato è particolarmente degno di nota perché la figura al centro del percorso non è un semplice referente biografico locale. Nato a Latiano nel 1841 e morto a Pompei nel 1926, Bartolo Longo è ricordato per la fondazione del Santuario del Rosario di Pompei, per la devozione mariana e per le opere caritative, sfociate nella canonizzazione del 19 ottobre 2025 (Dicastero delle Cause dei Santi, s.d.). Tale biografia, tuttavia, non è lineare né priva di tensioni narrative: durante gli studi a Napoli, Longo si avvicinò allo spiritismo, allontanandosi dalla fede cattolica, prima della successiva conversione e dell'impegno religioso e caritativo (Dicastero delle Cause dei Santi, s.d.; Leone XIV, 2025). Questa traiettoria — dal contesto spiritista alla devozione mariana, dalla crisi personale alla costruzione di un'opera pubblica e religiosa — rende la Casa Museo un luogo potenzialmente celebrativo. Il progetto ha definito un impianto phygital sostenibile che comprende un avatar basato sull'IA, un Ermes algoritmico dell'ecosistema, impersonato dalla figura di Bartolo Longo: un mediatore dell'esperienza che regola l'accesso, il tono e la forma dell'interazione tra lo spazio fisico e il livello digitale.
Il taglio che l'autore propone è problematizzante: quando volto e voce sintetici parlano in prima persona, l'avatar può generare un surplus di autenticità e redistribuire l'autorità del racconto (Huang & Jung, 2022), con effetti sulla fiducia e sull'interpretazione tipici dei media sintetici (Vaccari & Chadwick, 2020). In questo caso, l'avatar non è solo uno strumento di accompagnamento: mette in scena una forma di "presenza" che può incidere sulla costruzione sociale del patrimonio, perché attribuisce un timbro di voce, un ritmo e un volto a una figura storica su cui convergono memorie, aspettative e valori. La conseguenza è ambivalente: da un lato l'avatar può intensificare l'engagement e l'accessibilità; dall'altro può produrre effetti di appropriazione (chi parla davvero?), di semplificazione (quale Longo viene reso dicibile?) e di normalizzazione (quale immagine "coerente" viene restituita al visitatore e, indirettamente, alla comunità). Nel caso di Longo, tale ambivalenza è resa più evidente dalla compresenza di memoria civica, religiosa e celebrativa: l'avatar, parlando in prima persona, può rendere meno percepibile la distanza tra

[1] Home Page, Crhack Lab Bari, https://www.crhacklabbari.com (cons. 17/04/2026).
[2] Comune di Latiano, Inaugurazione Casa Museo – 13 febbraio 2026 https://www.comune.latiano.br.it/eventi/INAUGURAZIONE-CASA-MUSEO.html (cons. 17/04/2026).

il documento, l'interpretazione curatoriale e la messa in scena. Quindi, l'obiettivo è discutere le frizioni della trasmissione digitale di un patrimonio complesso e proporre mitigazioni progettuali ancorate alla trasparenza e alle fonti (Feng et al., 2023).

2. CASO DI STUDIO: LA CASA MUSEO BARTOLO LONGO COME DISPOSITIVO PHYGITAL SOSTENIBILE

Il caso di studio riguarda la realizzazione di un percorso espositivo permanente presso la Casa Museo Bartolo Longo (Latiano), concepito come intervento di rigenerazione e valorizzazione in chiave phygital. Il progetto si sviluppa in due fasi: una prima azione propedeutica di comunicazione e di ingaggio territoriale, mediante pannelli autoportanti interattivi installati in uno spazio di transito (aeroporto di Brindisi), e una seconda fase *in situ*, centrata sull'allestimento interno della Casa Museo, recentemente inaugurata. In questa seconda fase, la soluzione integra una pannellistica tematica reversibile e non invasiva con apparati digitali (interfacce *touch-free* fruibili tramite smartphone, un avatar-guida in prima persona e un dispositivo immersivo basato su contenuti olografici tramite un ventilatore 3D), secondo un modello che concepisce il phygital come un *continuum* esperienziale tra supporto fisico e layer digitale (Nofal et al., 2017; Lo Turco & Giovannini, 2020; Lupo, 2021). La scelta di una pannellistica autoportante ha rappresentato un compromesso progettuale necessario: ha consentito di garantire leggibilità e continuità narrativa senza intaccare le murature, che, per vincoli conservativi, non potevano essere toccate.
La selezione dei nuclei tematici è stata costruita a partire dallo studio e dall'analisi della documentazione testuale e iconografica pervenuta direttamente da Latiano, con l'intento di delineare il contesto di appartenenza della figura di Longo nel quadro latianese e non esclusivamente nella più consueta associazione con Pompei, luogo in cui egli ha trascorso una parte rilevante della propria vita. In questa scelta si riconosce una postura non neutra: l'allestimento non "trasferisce" semplicemente contenuti, ma contribuisce a rinegoziare l'ancoraggio territoriale e simbolico della figura, trasformando lo spazio museale in un dispositivo di riconoscimento comunitario. Ne derivano nuclei fondanti su cui strutturare la narrazione phygital, intesi non solo come categorie informative, ma come cornici interpretative attraverso cui il visitatore è invitato a leggere (e dunque a "situare") la biografia del personaggio:

- Corrispondenza
- Profano
- Sacro
- Solidarietà
- Musica

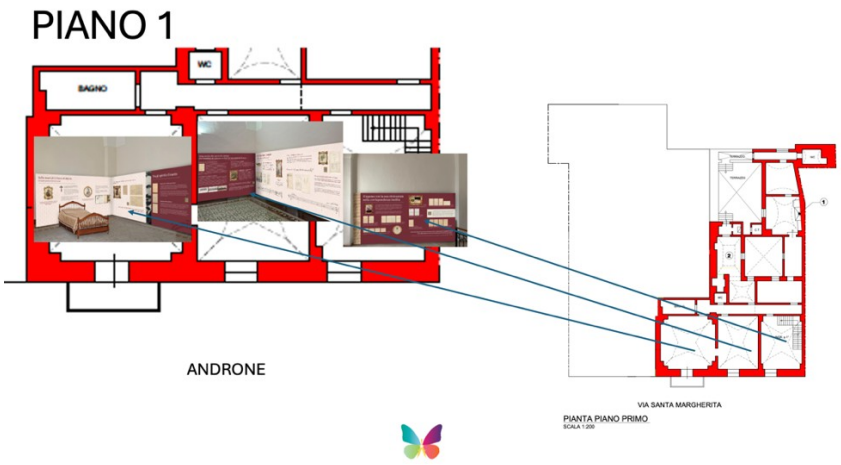


Figura 1. Piano 1: localizzazione della pannellistica tematica e corrispondenza con l'impianto planimetrico della Casa Museo

La relazione tra i nuclei tematici, in particolare tra "profano", "sacro" e "solidarietà", è delicata, perché favorisce la riflessione sulla transizione di un percorso di vita complesso, fondato sull'inquietudine di un uomo perso in pratiche occulte, per poi ritrovarsi redento e totalmente convertito. Longo può essere raccontato come uomo del territorio, figura devozionale, fondatore di opere sociali, santo canonizzato o

protagonista di una conversione. In questo quadro di pluralità interpretativa, l'avatar può stabilizzare una di queste letture come voce dominante.
Per evitare che tale complessità venga immediatamente assorbita da una sola chiave interpretativa, il percorso prevede anche un dispositivo preliminare di orientamento biografico.
Accanto a tali nuclei tematici, una timeline è stata concepita come dispositivo di facilitazione e orientamento: una soglia che introduce e rende accessibile la traiettoria biografica di Longo — collocata al piano terra — prima che l'utente attraversi i pannelli tematici dislocati al primo piano. In questo senso, la timeline funziona come "grammatica d'ingresso": ordina, semplifica e rende condivisibile una prima sequenza di eventi, predisponendo il visitatore all'incontro con la complessità tematica successiva.
L'elemento più critico, dal punto di vista antropologico, è però il ruolo dell'avatar-guida. L'avatar accoglie, introduce e spiega all'utente la mostra, fornendo una propria definizione all'inizio del percorso.
L'accompagnamento "in prima persona" non si limita a migliorare l'usabilità o a rendere più fluida la visita: introduce una forma di presenza mediata che può attivare empatia, prossimità e identificazione, ridefinendo l'asimmetria tra chi enuncia e chi ascolta. È proprio questa dinamica a rendere l'impianto potenzialmente efficace sul piano del coinvolgimento, ma anche delicato: l'avatar può essere percepito come "soggetto" più che come interfaccia e può quindi orientare non solo l'attenzione, ma anche il modo in cui la comunità e i visitatori attribuiscono autenticità e autorità al racconto. Il rischio diventa particolarmente evidente quando la narrazione riguarda conversione, missione religiosa o opere caritative: se formulate in prima persona, queste informazioni possono essere recepite non come sintesi museale, ma come la parola diretta del personaggio. D'altro canto, l'empatia prodotta dall'incontro con il personaggio può essere progettualmente utilizzata per promuovere una partecipazione attiva: il visitatore, riconoscendosi nella figura e nel contesto locale, può essere incoraggiato a conferire contenuti e documenti (memorie, fotografie, materiali d'archivio familiare), contribuendo all'aggiornamento dell'archivio digitale e, dunque, alla sostenibilità dei contenuti nel tempo. Questo introduce una logica circolare e potenzialmente autosostentante: il museo non solo comunica il patrimonio desueto, ma attiva pratiche di cura e di resilienza del patrimonio stesso attraverso la comunità.
Infine, l'intervento richiede sostenibilità manutentiva: l'allestimento è pensato per essere smontabile e aggiornabile (montaggi a secco), mentre i contenuti digitali sono progettati per essere sostituibili e versionabili. In questo modo correzioni e integrazioni possono avvenire senza interventi strutturali sull'immobile, mantenendo l'impianto coerente e durevole nel tempo (Nofal et al., 2017; Lo Turco & Giovannini, 2020; Lupo, 2021).

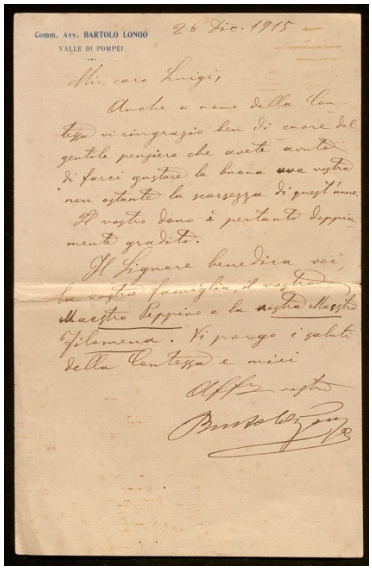
3. CATENA ANALITICA E IMPLICAZIONI SOCIO-CULTURALI DELL'AVATAR IA NEL PERCORSO PHYGITAL

Il contributo adotta un approccio di studio di caso qualitativo (Yin, 2017) e analizza la Casa Museo come un sistema ibrido in cui le scelte materiali e digitali incidono direttamente su ciò che viene reso "centrale" o "secondario" nella visita (Petrelli et al., 2013). Il punto non è dimostrare che il phygital "funziona", ma chiarire che cosa cambia quando un patrimonio desueto/minor viene organizzato lungo due layer (fisico e digitale): cambia la gerarchia dei contenuti e, con essa, la cornice interpretativa entro cui il visitatore attribuisce valore e autenticità (Nofal et al., 2017; Lo Turco & Giovannini, 2020). Nel caso della Casa Museo Bartolo Longo, questa redistribuzione non riguarda soltanto l'accesso ai contenuti, ma anche la gerarchia tra le fonti documentarie, la tradizione religiosa, la memoria locale e il racconto curatoriale. *Nelle house museums*, questa redistribuzione è particolarmente sensibile perché l'esperienza è già fondata su prossimità, atmosfera e aspettative di autenticità; l'aggiunta del digitale può quindi diventare o un supporto marginale (Bolter & Grusin, 1999) o, al contrario, il luogo in cui l'autenticità viene "spostata" sull'interfaccia e sul racconto (Miano & Borsotti, 2023).
È su questo crinale che si colloca l'avatar dell'IA. Quando la narrazione passa alla prima persona e si appoggia a volto e voce sintetici, l'avatar non "spiega" soltanto: assume la posizione di soggetto enunciante e può produrre un effetto-testimonianza (Huang & Jung, 2022) che riassegna l'autorità al medium più che alle fonti. In un contesto di media sintetici, tale slittamento incide sulla fiducia e sulla lettura del visitatore, rendendo necessario esplicitare le condizioni di produzione e i confini tra il documento, la ricostruzione e la messa in scena (Vaccari & Chadwick, 2020). Nel caso di Longo, l'effetto-testimonianza non è generico: riguarda una figura già investita di autorità religiosa e celebrativa. La prima persona può rendere più persuasiva la narrazione, ma anche spostare la fiducia dal documento alla performance digitale. Il problema non è solo l'eventuale errore, ma il fatto che la presenza digitale può

apparire più "autentica" della mediazione curatoriale che la produce. Per questo l'analisi assume come requisito operativo la presenza di segnali di provenienza e di un ancoraggio sistematico alle fonti: non un dettaglio tecnico, ma la condizione per mantenere verificabile la mediazione e ridurre l'ambiguità di attribuzione (Feng et al., 2023).
Su tale base, il contributo propone mitigazioni progettuali finalizzate a trasformare l'avatar da "testimone" implicito a interfaccia dichiarata di mediazione (trasparenza sull'uso di IA, protocollo di scrittura, distanziamento semiotico, versioning e reversibilità), coerenti con indicazioni etiche sull'IA nei beni culturali (Tiribelli et al., 2024). Infine, si propone una micro-valutazione leggera (osservazione *in situ* e, opzionalmente, brevi interviste post-visita e/o log d'uso) per verificare i trade-off più rilevanti: profondità di fruizione tra layer fisico e digitale, comprensione della provenienza e percezione dell'avatar come guida vs testimone (Falk & Dierking, 2016; Ceccarelli et al., 2024; Malik et al., 2023; Xia et al., 2025).

4. DALLA MEDIAZIONE ALLA PRESENZA: IMPLICAZIONI SOCIO-CULTURALI DELL'AVATAR IA

L'allestimento phygital viene letto qui come pratica culturale che non si limita a "comunicare" il patrimonio, ma contribuisce a produrre appartenenza e a rinegoziare pubblicamente ciò che conta come memoria condivisa. In questa prospettiva, la scelta di far "parlare" una figura storica attraverso un avatar IA sposta il baricentro dall'oggetto al rapporto: la visita diventa un incontro con una presenza mediata che orienta identificazione, riconoscimento o rifiuto (Falk & Dierking, 2016), e quindi incide sul modo in cui la comunità può riconoscersi nel sito come "proprio". Questo passaggio è cruciale perché l'avatar tende a trasformare il patrimonio in una forma di *heritage-as-discourse*: ciò che viene detto (e come viene detto) diventa un dispositivo di legittimazione, di selezione e di normalizzazione del passato (Smith, 2006). In contesti europei segnati da una forte "memorializzazione" dei luoghi, tali dispositivi contribuiscono alla costruzione dell'identità e alla governance della memoria (Macdonald, 2013). Ne deriva un'esigenza specifica: se l'empatia attivata dall'avatar può sostenere un ciclo partecipato di conferimento di materiali e memorie, tale apertura va accompagnata da regole di curatela e da segnali espliciti di provenienza e mediazione, per evitare che la "presenza" sintetica sostituisca la verificabilità delle fonti.


Comm. Avv. BARTOLO LONGO
VALLE DI POMPEI

26 Dic. 1915

Fig.2 Esempio di fonte primaria proveniente dalla Casa Museo Ribezzi Petrosillo, Latiano, utilizzate come base documentaria per la costruzione dei contenuti del percorso phygital.

In questo senso, l'impianto phygital può diventare non solo una soluzione di accesso, ma un dispositivo di partecipazione responsabile, coerente con l'idea di museo come istituzione "al servizio della società" e con il coinvolgimento delle comunità. La posta antropologica, quindi, si sostanzia nella governance della memoria che esso abilita: rendere leggibile il confine tra documento, interpretazione e messa in scena, e mantenere oggettivo il messaggio. In questo quadro, l'apertura a contributi e materiali dei visitatori può diventare una risorsa per la sostenibilità dei contenuti solo se inserita in un perimetro curatoriale chiaro

(selezione, contestualizzazione, tracciabilità), in modo che la partecipazione sia chiaramente trasparente e veritiera nella sua essenza.



Figura 3. Avatar IA di Bartolo Longo: ricostruzione in prima persona, con volto e voce sintetici, per la guida phygital.

5. CONCLUSIONI

Il caso della Casa Museo Bartolo Longo mostra che la valorizzazione phygital di un patrimonio desueto può riattivare il luogo non solo come spazio espositivo, ma anche come dispositivo di riconoscimento: attraverso la selezione dei nuclei tematici e la costruzione di una soglia biografica (timeline). La visita reinquadra la figura di Longo in un contesto latianense, con ripercussioni sulla memoria pubblica e sull'appartenenza locale. In questo quadro, la tecnologia non è neutrale: modifica le gerarchie di attenzione tra i layer fisico e digitale e, soprattutto, introduce nuovi regimi di autorità quando la narrazione assume una voce "in prima persona".
In questo approccio, l'avatar IA è il punto più delicato. Se da un lato aumenta la continuità narrativa e l'accessibilità, dall'altro può produrre effetto-testimonianza e surplus di autenticità, spostando l'autorità dal piano delle fonti a quello della "presenza" percepita. Ne consegue che la questione centrale non è "usare o non usare" l'avatar, ma a quali condizioni: la mediazione deve restare riconoscibile e verificabile, e il confine tra documento, interpretazione e messa in scena deve essere reso leggibile. Lo storytelling aumentato non deve prevalere sulla realtà storica (Claisse et al., 2018).
Dal caso emergono quattro indicazioni trasferibili:

1. Progettare soglie e gerarchie: garantire un minimo comune di orientamento (prima della complessità tematica), evitando che ciò che sta nel digitale diventi invisibile o meramente ornamentale.
2. Rendere l'avatar un'interfaccia alle fonti: disclosure dell'uso dell'IA, segnali di provenienza, rimandi alle fonti e un protocollo di scrittura che limiti l'interiorità non attestabile; l'avatar non deve "sostituire" la fonte, ma accompagnare verso la fonte.
3. Gestire le figure ad alta densità simbolica: quando l'avatar rappresenta una figura religiosamente, civicamente o celebrativamente connotata, il rischio di surplus di autenticità aumenta. La progettazione deve quindi rendere leggibile il confine tra fonte, interpretazione curatoriale e messa in scena, evitando che la prima persona sposti la fiducia dal documento alla sola presenza digitale.
4. Sostenibilità come governance: versioning e reversibilità non sono solo manutenzione tecnica; sono condizioni per correggere, aggiornare e mantenere la responsabilità curatoriale. Se si attiva un ciclo partecipato di conferimento di materiali da parte dei visitatori, si richiede un perimetro di curatela chiaro (selezione, contestualizzazione, tracciabilità).

Infine, la micro-valutazione proposta (osservazione, brevi interviste, log d'uso) è necessaria per verificare due aspetti: se la stratificazione offre davvero accesso ai livelli profondi e se l'avatar viene percepito come guida in veste di mediatore o come testimone di sé stesso. È su questa evidenza che si misura la tenuta culturale dell'impianto: non l'effetto spettacolare, ma la capacità di attivare la partecipazione senza opacizzare le fonti.

RINGRAZIAMENTI

L'autore ringrazia il Comune di Latiano per la collaborazione e la disponibilità dimostrate, nonché per aver autorizzato l'analisi della Casa Museo Bartolo Longo ai fini del presente contributo. Un ringraziamento va inoltre a tutti i soggetti coinvolti nelle attività di valorizzazione e di messa in opera dell'allestimento, in particolare alla dott.ssa Alessia Cioce.

BIBLIOGRAFIA

Bolter, J. D., & Grusin, R. (1999). Remediation: Understanding new media. MIT Press.

Ceccarelli, S., Cesta, A., Cortellessa, G., De Benedictis, R., Fracasso, F., Leopardi, L., Ligios, L., Lombardi, E., Malatesta, S. G., Oddi, A., Pagano, A., Palombini, A., Romagna, G., Sanzari, M., & Schaerf, M. (2024). Evaluating visitors' experience in museum: Comparing artificial intelligence and multi-partitioned analysis. Digital Applications in Archaeology and Cultural Heritage, 33, e00340. doi:10.1016/j.daach.2024.e00340.

Claisse, C., Petrelli, D., Dulake, N., Marshall, M., & Ciolfi, L. (2018). Multisensory interactive storytelling to augment the visit of a historical house museum. In Proceedings of the Digital Heritage International Congress (DigitalHERITAGE). IEEE. doi:10.1109/DigitalHeritage.2018.8810099.

De Bari, M. (2023). Verso una fruizione phygital del Patrimonio Culturale in termini di Open Culture. In G. Fioretti & C. Campobasso (a cura di), Atti del II Congresso Beni Culturali in Puglia. Il patrimonio culturale pugliese. Ricerche, applicazioni e best practices (pp. 182–186). Edizioni Fondazione Pasquale Battista.

De Bari, M. (2024). A study for innovation of methodologies and processes in digitisation and digital creation for improving the interaction with the Cultural Heritage. Procedia Computer Science, 237, 187–194. https://doi.org/10.1016/j.procs.2024.05.095.

Dicastero delle Cause dei Santi. (s.d.). Bartolo Longo. https://www.causesanti.va/it/santi-e-beati/bartolo-longo.html.

European Commission. (2025). PHYGITAL solutions for the cultural and creative industries (AMPLIFY) (Grant Agreement No. 101177413). CORDIS – EU research results. https://doi.org/10.3030/101177413.

European Commission. (2025). The Cultural Heritage Cloud (European Collaborative Cloud for Cultural Heritage). Research and Innovation.

Falk, J. H., & Dierking, L. D. (2016). The Museum Experience Revisited. Routledge.

Feng, K. J. K., Ritchie, N., Blumenthal, P., Parsons, A., & Zhang, A. X. (2023). Examining the impact of provenance-enabled media on trust and accuracy perceptions. Proceedings of the ACM on Human-Computer Interaction, 7(CSCW2), 1–42. doi:10.1145/3610061.

Huang, J., & Jung, Y. (2022). Perceived authenticity of virtual characters makes the difference. Frontiers in Virtual Reality, 3, 1033709. doi:10.3389/frvir.2022.1033709.

Leone XIV. (2025, 19 ottobre). Cappella Papale per la Canonizzazione dei Beati: Ignazio Choukrallah Maloyan, Peter To Rot, Vincenza Maria Poloni, Maria del Monte Carmelo Rendiles Martínez, Maria Troncatti, José Gregorio Hernández Cisneros e Bartolo Longo. Libreria Editrice Vaticana. https://www.vatican.va/content/leo-xiv/it/homilies/2025/documents/20251019-canonizzazioni.html.

Lo Turco, M., & Giovannini, E. C. (2020). Towards a phygital heritage approach for museum collection. Journal of Archaeological Science: Reports, 34, 102639. doi:10.1016/j.jasrep.2020.102639.

Lupo, E. (2021). Design and innovation for the Cultural Heritage. Phygital connections for a heritage of proximity. Agathón | International Journal of Architecture, Art and Design, 10, 186–199. doi:10.19229/2464-9309/10172021.

Macdonald, S. (2013). Memorylands: Heritage and identity in Europe today. Routledge.

Malik, U. S., Micoli, L. L., Caruso, G., & Guidi, G. (2023). Integrating quantitative and qualitative analysis to evaluate digital applications in museums. Journal of Cultural Heritage, 62, 65–71. doi:10.1016/j.culher.2023.06.002.

Miano, A., & Borsotti, M. (2023). Experiencing authenticity of the house museums in hybrid environments. Multimodal Technologies and Interaction, 7(7), 72. doi:10.3390/mti7070072.

Nofal, E., Reffat, R. M., & Vande Moere, A. (2017). Phygital heritage: An approach for heritage communication. In Proceedings of the 3rd Annual Conference of the Immersive Learning Research Network (iLRN 2017) (pp. 220–229). doi:10.3217/978-3-85125-530-0-36.

Petrelli, D., Ciolfi, L., van Dijk, D., Hornecker, E., Not, E., & Schmidt, A. (2013). Integrating material and digital: A new way for cultural heritage. interactions, 20(4), 58–63. doi:10.1145/2486227.2486239.

Smith, L. (2006). Uses of heritage. Routledge https://doi.org/10.4324/9780203602263.

Tiribelli, S., Pansoni, S., Frontoni, E., & Giovanola, B. (2024). Ethics of artificial intelligence for cultural heritage: Opportunities and challenges. IEEE Transactions on Technology and Society, 5(3), 293–305. doi:10.1109/TTS.2024.3432407.

Torre, A., & Christodoulou, C. (2021). Co-creation of narratives for "minor" sites of cultural heritage in Euro-Mediterranean peri-urban areas: Conditions of a small temple on the east coast of Attica, Greece. Heritage, 4(4), 2918–2941. doi:10.3390/heritage4040163.

Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. Social Media + Society, 6(1), 1–13. doi:10.1177/2056305120903408.

Xia, T., Wu, Y., Qiu, A., Liu, Z., & Fan, M. (2025). The Impact of AI Guide Language Strategies on Museum Visitor Experience: The Mediating Role of Psychological Distance in the Arousal–Topic Fit Effect. Behavioral Sciences, 15(11), 1569. doi:10.3390/bs15111569.

Yin, R. K. (2017). Case study research and applications: Design and methods (6th ed.). SAGE.

Musei senza confini: il digitale come infrastruttura di memoria partecipata nei musei delle aree interne. Il GeoMuseo Monte Arci Stefano Incani di Masullas e il Museo Paleontologico PARC di Genoni

PAOLA PALMAS

UNIVERSITÀ DEGLI STUDI DI CAGLIARI, FACOLTÀ DI BENI CULTURALI,
P.PALMAS9@STUDENTI.UNICA.IT

ABSTRACT

I musei delle aree interne svolgono una funzione strategica nella conservazione e trasmissione del patrimonio culturale materiale e immateriale, ma sono spesso esposti a processi di marginalizzazione legati a barriere geografiche, infrastrutturali e di accessibilità. Nel quadro delle Digital Humanities, il digitale può essere interpretato non solo come insieme di strumenti, ma come infrastruttura epistemica capace di ridefinire il ruolo di queste istituzioni e i loro rapporti con le comunità di riferimento. Il contributo analizza due casi studio situati nella Sardegna centrale — il Geomuseo Monte Arci "Stefano Incani" di Masullas e il Museo Paleontologico PARC di Genoni — con l'obiettivo di valutare in che misura pratiche digitali orientate alla partecipazione, anche a bassa soglia tecnologica, possano favorire accessibilità, co-generazione della memoria e sostenibilità culturale nei contesti periferici. L'analisi adotta un approccio qualitativo basato su casi studio, osservazione dei dispositivi museali e valutazione critica delle potenzialità di digitalizzazione e archiviazione in relazione ai principi FAIR e dell'Open Science. Il contributo propone una riflessione sul ruolo dei musei delle aree interne come laboratori privilegiati per le Digital Humanities applicate, evidenziando implicazioni teoriche e operative in termini di equità culturale, giustizia territoriale e public engagement, e superando una concezione meramente strumentale della digitalizzazione.
Parole chiave: Digital Humanities; public engagement; public history; citizen humanities; memoria digitale

ABSTRACT (ENGLISH)

Museums Without Borders. Digital Technologies as an Infrastructure of Participatory Memory in Inner-Area Museums. The Geomuseo Monte Arci "Stefano Incani" in Masullas and the PARC Paleontological Museum in Genoni.

Museums located in inner and marginal areas play a crucial role in preserving and transmitting both tangible and intangible cultural heritage, yet they often face marginalization due to geographical, infrastructural, and accessibility barriers. Within the framework of Digital Humanities, digital technologies can be understood not merely as tools, but as epistemic infrastructures capable of reshaping the role of museums and their relationship with local communities.This paper examines two case studies from central Sardinia — the GeoMuseo Monte Arci "Stefano Incani" in Masullas and the PARC Paleontological Museum in Genoni — in order to investigate how participatory digital practices,

including low-threshold technologies, can enhance accessibility, memory co-generation, and cultural sustainability in peripheral contexts. The analysis adopts a qualitative approach based on case studies, observation of museum mediation devices, and critical assessment of digitalization and archiving strategies in light of FAIR principles and Open Science.The paper argues that inner-area museums can function as privileged laboratories for applied Digital Humanities, highlighting both theoretical and operational implications in terms of cultural equity, territorial justice, and public engagement, and moving beyond a purely instrumental view of digital heritage.
Keywords: Digital Humanities; public engagement; public history; citizen humanities; digital memory

1. INTRODUZIONE

Negli ultimi due decenni, le Digital Humanities hanno profondamente trasformato il modo in cui il patrimonio culturale viene prodotto, interpretato e condiviso. Il digitale non si limita più alla mera digitalizzazione delle collezioni, ma interviene sui processi epistemologici della ricerca, favorendo modelli aperti, collaborativi e orientati alla partecipazione. In questa prospettiva, il digitale può essere inteso come un'infrastruttura critica di produzione della conoscenza, capace di mettere in discussione approcci meramente strumentali e di ridefinire le pratiche di ricerca e mediazione culturale (Berry & Fagerjord, 2017).In ambito museale, tali trasformazioni si intrecciano con le pratiche della public history, che enfatizzano la dimensione dialogica della costruzione del sapere storico e la condivisione dell'autorità interpretativa tra studiosi, istituzioni e comunità (Cauvin, 2016). Il public engagement assume dunque una funzione strutturale, in particolare nei contesti in cui il pubblico è riconosciuto come attore attivo nella produzione di significato e nell'esperienza del patrimonio (Bonacchi, 2014).I musei delle aree interne rappresentano un osservatorio privilegiato per analizzare queste dinamiche. Collocati in territori spesso esclusi dai principali flussi culturali e turistici, tali musei svolgono un ruolo essenziale di presidio identitario e di mediazione culturale, ma si confrontano con fragilità infrastrutturali e limitate risorse. In questi contesti, il digitale può contribuire a ridefinire il rapporto tra patrimonio, territorio e comunità, intervenendo sulle dimensioni spaziali e narrative della conoscenza culturale (Bodenhamer, Corrigan & Harris, 2010). Il presente contributo intende dunque valutare in che misura il digitale possa configurarsi come infrastruttura di memoria partecipata nei musei delle aree interne, superando una concezione meramente strumentale della digitalizzazione. In questa prospettiva, i musei delle aree interne non sono considerati esclusivamente come contesti di applicazione delle Digital Humanities, ma come ambienti capaci di generare modelli teorici situati e trasferibili.

2. QUADRO TEORICO

Il contributo si colloca all'intersezione di tre ambiti teorici delle Digital Humanities: la public history, le citizen humanities e i digital memory studies. Questi approcci condividono una visione del digitale come spazio di mediazione culturale e di produzione collaborativa della conoscenza, piuttosto che come semplice supporto tecnologico.La public history enfatizza la dimensione partecipativa della produzione storica, promuovendo pratiche di condivisione dell'autorità interpretativa e di co-curatela. In ambito museale, ciò implica il superamento di narrazioni univoche a favore di racconti plurali, costruiti attraverso il dialogo con le comunità di riferimento e con i pubblici del patrimonio (Cauvin, 2016).Le citizen humanities, derivate dal paradigma della citizen science, estendono tali principi alle discipline umanistiche, riconoscendo il valore epistemico delle conoscenze locali e informali. Le piattaforme digitali diventano spazi di collaborazione nei quali cittadini, ricercatori e istituzioni contribuiscono congiuntamente alla produzione, descrizione e interpretazione dei dati culturali, secondo modelli di partecipazione distribuita (Ridge, 2014).Infine, i digital memory studies concepiscono la memoria come un processo dinamico, sociale e tecnologicamente mediato. La digitalizzazione non è neutra, ma influisce sulle modalità di selezione, organizzazione e visualizzazione delle memorie, rendendo necessario un approccio critico ai dispositivi narrativi e visuali che strutturano gli archivi digitali (Drucker, 2014). In questo senso, il patrimonio digitale si configura come un processo relazionale e partecipativo, piuttosto che come un insieme statico di oggetti (Giaccardi, 2012).

3. METODOLOGIA

Il contributo adotta un approccio qualitativo basato sull'analisi comparativa di due casi studio museali. La metodologia integra osservazione dei dispositivi espositivi e narrativi, analisi delle pratiche di mediazione

culturale e valutazione critica delle potenzialità di implementazione digitale. I casi sono stati selezionati in quanto rappresentativi di musei di piccole dimensioni situati in aree interne, caratterizzati da una forte relazione con il territorio e le comunità locali. L'analisi non si concentra su progetti digitali pienamente implementati, ma sulle pratiche esistenti e sulle potenzialità di sviluppo, in una prospettiva esplorativa tipica delle Digital Humanities applicate ai contesti periferici. L'analisi comparativa dei casi studio è orientata in particolare a tre dimensioni trasversali — accessibilità, mediazione digitale e partecipazione — considerate come indicatori qualitativi della capacità dei musei di operare come infrastrutture di conoscenza e memoria nei contesti delle aree interne.L'obiettivo non è misurare l'impatto quantitativo delle tecnologie, ma analizzarne il potenziale epistemico e culturale.Al fine di rendere esplicita la posizione analitica del contributo e rafforzare la comparabilità tra i casi studio, si propone una sintesi delle principali dimensioni di analisi — accessibilità, mediazione digitale e partecipazione — applicate ai due contesti museali (cfr. Tabella 1)

Dimensione	**Geomuseo Monte Arci (Masullas)**	**Museo PARC (Genoni)**
Accessibilità	Fruizione mediata da strumenti digitali (VR, mappe narrative) per superare limiti fisici del territorio	Dispositivi inclusivi (LIS, audio, easy-to-read, Braille) orientati a accessibilità universale
Mediazione digitale	Ricostruzioni geologiche e territoriali; integrazione tra paesaggio e narrazione	Ricostruzioni paleontologiche e ambienti immersivi per interpretazione scientifica
Partecipazione	Potenziale sviluppo di narrazioni territoriali e storytelling comunitario	Maggiore integrazione con didattica, scuole e pubblici diversificati
Ruolo del digitale	Estensione del territorio e interpretazione del paesaggio	Mediazione cognitiva e inclusione dei pubblici
Criticità	Limitata infrastrutturazione digitale sistemica	Necessità di integrazione in reti di ricerca e standardizzazione dati

Tabella 1. Sintesi comparativa dei casi studio

4. I CASI STUDIO

I casi studio non sono presentati come esempi di piena implementazione tecnologica, ma come contesti nei quali pratiche già attive e scenari di sviluppo digitale realistici consentono di osservare criticamente il ruolo del digitale come infrastruttura culturale e sociale. L'analisi si concentra in particolare su accessibilità, mediazione e partecipazione, come scenari di applicazione sostenibile.

4.1 IL GEOMUSEO MONTE ARCI "STEFANO INCANI" DI MASULLAS

Il Geomuseo Monte Arci "Stefano Incani" di Masullas si colloca in un territorio profondamente caratterizzato dalla presenza del Monte Arci, uno dei principali giacimenti di ossidiana del Mediterraneo occidentale. Il museo articola il proprio racconto integrando prospettive geologiche configurandosi come un esempio paradigmatico di patrimonio intrinsecamente interdisciplinare e territorialmente situato.In questo contesto, il museo non opera esclusivamente come spazio espositivo, ma come dispositivo territoriale di mediazione culturale, capace di connettere saperi scientifici e memorie locali. Le mappe narrative digitali e le ricostruzioni tridimensionali dei geositi permettono inoltre di superare i limiti fisici e ambientali del territorio, rendendo fruibili contesti difficilmente accessibili per ragioni morfologiche o motorie. In questa prospettiva, la tecnologia non sostituisce l'esperienza diretta, ma la integra, offrendo percorsi interpretativi alternativi e personalizzabili. Strumenti quali audioguide multilingue, visori di realtà virtuale e ricostruzioni immersive svolgono un ruolo concreto di supporto alle attività didattiche, facilitando la comprensione di processi complessi – dalla formazione geologica ai cicli di estrazione e

circolazione dell'ossidiana nel Neolitico – e favorendo forme di apprendimento esperienziale, particolarmente efficaci in ambito scolastico. Dal punto di vista dell'accessibilità, l'adozione di interfacce semplificate, contenuti adattivi e strumenti di traduzione contribuisce a includere pubblici con disabilità motorie, sensoriali o linguistiche. In questo senso, il digitale si configura come strumento di equità culturale e di ampliamento dei diritti di fruizione.Il caso del GeoMuseo Monte Arci si inserisce inoltre in un più ampio sistema territoriale legato alla valorizzazione dell'ossidiana, in cui rientra anche il Museo dell'Ossidiana di Pau. Tale relazione evidenzia come il digitale possa fungere da infrastruttura di connessione tra istituzioni diffuse, contribuendo alla costruzione di narrazioni territoriali integrate e multilivello.

4.2 IL MUSEO PALEONTOLOGICO PARC DI GENONI

Il Museo Paleontologico PARC di Genoni rappresenta un presidio scientifico e culturale di riferimento per la paleontologia sarda e per la divulgazione nelle aree interne. Le collezioni fossilifere e le attività educative offrono un contesto particolarmente favorevole allo sviluppo di pratiche digitali orientate alla mediazione, all'accessibilità e alla partecipazione. In questo caso, il digitale si configura non come semplice supporto espositivo, ma come mediatore cognitivo, capace di facilitare la comprensione dei reperti e dei contesti paleoambientali. La digitalizzazione delle collezioni e la creazione di archivi accessibili contribuiscono a estendere l'esperienza museale oltre la visita fisica, rafforzando il dialogo tra museo, scuola e ricerca e riducendo le barriere geografiche tipiche dei contesti periferici. Le tecnologie immersive e le ricostruzioni tridimensionali permettono di restituire ambienti e organismi non più esistenti, trasformando l'osservazione in un'esperienza attiva e interpretativa. In ambito didattico, tali strumenti favoriscono modalità di apprendimento esperienziale e partecipativo, facilitando l'interazione con contenuti scientifici complessi. Particolare rilevanza assume il tema dell'accessibilità, intesa come dimensione strutturale della mediazione museale. L'integrazione di dispositivi digitali inclusivi — quali contenuti multilingue, supporti audiovisivi accessibili e strumenti adattivi per diversi bisogni sensoriali e cognitivi — contribuisce a ridurre le barriere linguistiche, percettive e cognitive. In questa prospettiva, il digitale non opera come soluzione compensativa, ma come dispositivo di equità culturale, ampliando il diritto alla fruizione del patrimonio. L'adozione di strumenti quali audioguide digitali, contenuti in lingua dei segni e dispositivi per la lettura tattile evidenzia come l'accessibilità possa essere integrata nei processi di mediazione in modo sistemico. Allo stesso tempo, l'impiego di linguaggi semplificati e supporti visivi per l'orientamento nello spazio museale contribuisce a rendere l'esperienza più inclusiva anche per pubblici con esigenze cognitive diversificate. Infine, l'inserimento del museo in reti digitali di ricerca e divulgazione rafforza il suo ruolo come nodo attivo nella produzione e circolazione del sapere scientifico. Il caso del PARC mostra come anche contesti di piccole dimensioni possano sviluppare pratiche digitali sostenibili e significative, contribuendo a ridefinire il museo come infrastruttura culturale accessibile, partecipata e territorialmente radicata nel quadro delle Digital Humanities.

4.3 IN SINTESI

In entrambi i casi, il digitale emerge come strumento operativo, capace di supportare l'apprendimento migliorare l'accessibilità e facilità di lavoro con le scuole. Consente inoltre di valorizzare patrimoni fragili e localizzati permettendo anche di ridurre l'isolamento delle aree interne. Non quindi tecnologia fine a sé stessa, ma aiuto concreto per rendere i musei infrastrutture culturali vive, inclusive e pienamente integrate nel territorio.Nei due casi analizzati, il digitale emerge come infrastruttura operativa capace di supportare l'apprendimento, migliorare l'accessibilità e rafforzare il rapporto tra museo e territorio. In entrambi i contesti, le tecnologie digitali non si configurano come elementi accessori, ma come dispositivi di mediazione culturale che consentono di valorizzare patrimoni localizzati e ridurre l'isolamento delle aree interne. Il digitale si definisce dunque come condizione abilitante per la costruzione di musei accessibili, partecipati e integrati nel contesto territoriale.

5. IL DIGITALE COME INFRASTRUTTURA DI MEMORIA PARTECIPATA

Nei due casi analizzati, il digitale può essere interpretato come infrastruttura di memoria piuttosto che come semplice supporto tecnologico. Pratiche di digital storytelling, raccolta di testimonianze orali e produzione di contenuti audiovisivi consentono la co-generazione della memoria, rafforzando il legame tra museo e comunità e riconfigurando il patrimonio come processo sociale condiviso (Giaccardi, 2012).Podcast, mappe interattive e micro-narrazioni digitali rappresentano strumenti accessibili e

sostenibili per estendere il museo oltre i suoi confini fisici, intervenendo sulle dimensioni spaziali e narrative della conoscenza culturale e contribuendo a ridurre le disuguaglianze territoriali (Bodenhamer, Corrigan & Harris, 2010). Tali pratiche si inseriscono in un più ampio quadro di sperimentazione delle citizen humanities, in cui la partecipazione dei pubblici non è accessoria, ma strutturale alla produzione del sapere culturale (Ridge, 2014).Questi processi, tuttavia, richiedono una governance stabile, competenze adeguate e strategie di sostenibilità che non sono sempre disponibili nei contesti periferici. La riflessione sulle infrastrutture digitali museali deve pertanto accompagnarsi a una valutazione critica delle condizioni organizzative e istituzionali che ne rendono possibile l'effettiva implementazione (Berry & Fagerjord, 2017).Nei casi analizzati, l'adozione dei principi FAIR (Findable, Accessible, Interoperable, Reusable) non va intesa come elemento aggiuntivo o puramente tecnico, ma come componente strutturale delle pratiche di mediazione e produzione della conoscenza. In particolare, la digitalizzazione delle collezioni, la produzione di contenuti multimediali e le pratiche di narrazione territoriale possono essere reinterpretate alla luce dei principi FAIR, in quanto processi orientati alla reperibilità, accessibilità e riusabilità dei dati culturali.L'adozione dei principi FAIR (Findable, Accessible, Interoperable, Reusable) costituisce un elemento chiave per la sostenibilità dei progetti digitali museali. I dati prodotti — testimonianze orali, contenuti multimediali, narrazioni territoriali — possono essere organizzati in repository aperti, favorendo il riuso per finalità di ricerca, didattica e valorizzazione culturale. L'Open Science, in questo contesto, non è solo una scelta tecnica, ma una postura etica orientata all'accessibilità e alla condivisione della conoscenza.In questo senso, i principi FAIR contribuiscono a strutturare il museo come infrastruttura informativa aperta, interoperabile e orientata alla produzione condivisa della conoscenza.

6. DISCUSSIONE

L'analisi evidenzia come i musei delle aree interne possano diventare laboratori privilegiati per sperimentare pratiche DH orientate alla partecipazione e all'equità culturale. Tuttavia, permangono criticità legate alla sostenibilità economica, alla disponibilità di competenze digitali e alla necessità di una governance stabile. La creazione di reti museali digitali e il coinvolgimento delle istituzioni accademiche emergono come strategie fondamentali per superare tali limiti.

7. CONCLUSIONI

Il contributo ha mostrato come il digitale, inteso come infrastruttura epistemica e sociale, possa contribuire a ridefinire il ruolo dei musei delle aree interne, superando una concezione meramente conservativa del patrimonio. Attraverso l'analisi dei casi del GeoMuseo Monte Arci "Stefano Incani" di Masullas e del PARC Museo di Genoni, emerge come pratiche digitali orientate alla partecipazione e alla condivisione possano rafforzare l'accessibilità, la sostenibilità e la rilevanza scientifica di queste istituzioni.Dal punto di vista delle Digital Humanities, i musei periferici si configurano come contesti privilegiati per sperimentare modelli di public history, citizen humanities e memoria digitale, in cui la produzione di conoscenza è distribuita e situata. Il contributo evidenzia inoltre la necessità di strategie di governance e di reti collaborative tra musei e istituzioni accademiche per garantire la preservazione, l'interoperabilità e il riuso dei dati culturali.In questo senso, il contributo intende offrire non solo un'analisi applicativa, ma anche un apporto teorico al dibattito delle Digital Humanities, proponendo i musei delle aree interne come laboratori privilegiati per sperimentare modelli di produzione della conoscenza situata, partecipata e digitalmente mediata.In questa prospettiva, i musei delle aree interne non rappresentano un ambito marginale, ma un terreno centrale per riflettere criticamente sul rapporto tra patrimonio, digitale e società, offrendo un contributo significativo al dibattito contemporaneo nelle Digital Humanities.I musei delle aree interne non costituiscono soltanto un ambito applicativo delle Digital Humanities, ma un contesto privilegiato per la produzione di modelli teorici situati, capaci di ridefinire il rapporto tra patrimonio, tecnologia e comunità.

BIBLIOGRAFIA

Bodenhamer, D. J., Corrigan, J., Harris, T. M. (2010). *The Spatial Humanities: GIS and the Future of Humanities Scholarship*. Bloomington: Indiana University Press.

Bonacchi, C. (2014). *Understanding the Public Experience of Archaeology*. London: Routledge.

Berry, D. M., Fagerjord, A. (2017). *Digital Humanities: Knowledge and Critique in a Digital Age*. Cambridge: Polity Press.

Cauvin, T. (2016). *Public History: A Textbook of Practice*. London: Routledge.

Drucker, J. (2014). *Graphesis: Visual Forms of Knowledge Production*. Cambridge (MA): Harvard University Press.

Giaccardi, E. (2012). *Heritage and Social Media: Understanding Heritage in a Participatory Culture*. London: Routledge.

Ridge, M. (2014). *Crowdsourcing Our Cultural Heritage*. Farnham: Ashgate.

Candela, L., Castelli, D., & Pagano, P. (2020). *Data Infrastructures for Open Science.* Springer.

European Commission. (2018). *Turning FAIR into Reality. Publications Office of the European Union*.

Tosca, S., & Klastrup, L. (2019). *Transmedial Worlds in Everyday Life*. Routledge.

Pubblici remoti: la Realtà Virtuale come infrastruttura culturale di accesso

ELISA SMERALDO
UNIVERSITÀ DI ROMA LA SAPIENZA – ELISA.SMERALDO@UNIROMA1.IT

ABSTRACT (ITALIANO)

Nel contesto culturale contemporaneo, i processi di digitalizzazione stanno ridefinendo le modalità di fruizione del patrimonio culturale e aprendo possibilità inedite per l'accessibilità. Nonostante tali sviluppi, le arti performative presentano ancora barriere pervasive per molte persone con disabilità. Questo contributo prende in analisi come le innovazioni tecnologiche possono riconfigurare le condizioni di accesso al teatro, favorendo una partecipazione culturale più inclusiva. Il caso di studio presentato, sviluppato con il Teatro Stabile di Torino, riguarda la realizzazione di due performance in realtà virtuale destinate a pazienti ricoverati in hospice e ai loro caregiver, con l'obiettivo di esplorare forme di fruizione teatrale alternative alla co-presenza fisica, garantendo l'accesso al di là delle limitazioni legate alla mobilità. Da un punto di vista metodologico, il progetto adotta un approccio sperimentale e prevede il coinvolgimento di psicologi, operatori sanitari e volontari, consultati nella definizione del protocollo di raccolta dati e successivamente coinvolti nella sperimentazione dei contenuti al fine di raccogliere riscontri qualitativi da integrare nell'analisi. La ricerca esamina come gli ambienti immersivi mediano l'esperienza, l'attenzione e la percezione di presenza per pubblici che sarebbero altrimenti esclusi dalla cultura teatrale, in dialogo con la letteratura inerente al teatro digitale e al processo di mediatizzazione delle arti performative. I risultati preliminari, pur evidenziando alcune criticità legate ai contenuti proposti, indicano che la VR può facilitare il coinvolgimento narrativo e la percezione di presenza anche in assenza di co-presenza fisica, valorizzando la fruizione culturale in condizioni di isolamento.
Parole chiave: realtà virtuale; teatro digitale; accessibilità; patrimonio culturale immateriale; mediatizzazione.

ABSTRACT (ENGLISH)

Remote Audiences: Virtual Reality as Cultural Access Infrastructure
Within contemporary cultural context, digitality plays a pivotal role in shaping new modes of experiencing cultural heritage, opening possibilities for accessibility. Nonetheless, people with disabilities still encounter pervasive barriers to the arts. This paper examines how technological innovation can reconfigure these conditions in the performing arts, with the objective of fostering a more inclusive cultural participation. The case-study here presented was developed in collaboration with the National Theatre of Turin and consists in the realization of two VR performances designed for patients in hospice care and their families, with the objective of exploring modes of theatre fruition different than physical co-presence, granting access beyond limitations concerning mobility issues. Methodologically, the project adopts an experimental approach and involves psychologists, health care professionals and volunteers, who were consulted in the definition of the data collection protocol and, subsequently, included in the experimentation in order to gather qualitative feedback to incorporate in the analysis. The research examines how immersive environments mediate presence and attention for publics who would otherwise be excluded from theatre culture, in dialogue with literature on digital theatre and on the mediatization of the performing arts. While the preliminary results highlight some criticalities regarding the contents proposed, they also indicate that VR can facilitate narrative involvement and the perception of presence even in the absence of physical copresence, valorising cultural engagement in contexts of isolation.
Keywords: virtual reality; digital theatre; accessibility; intangible cultural heritage; mediatization.

1. INTRODUZIONE

Le tecnologie digitali hanno determinato trasformazioni significative in ambito umanistico, da un lato nella produzione del sapere, dall'altro mutando le modalità di produzione, fruizione e distribuzione delle opere artistiche e dei beni culturali. In questo scenario, l'adozione di servizi di accessibilità inediti da

parte di enti culturali di diversa natura si colloca tra le applicazioni più significative delle nuove tecnologie, soprattutto laddove tali servizi consentono di raggiungere pubblici tendenzialmente marginalizzati e impossibilitati a partecipare ad attività culturali per ragioni fisiche, sociali o contestuali. La partecipazione di queste comunità non solo costituisce un requisito fondamentale per una cultura realmente democratica, ma determina un beneficio per l'intera comunità: la valorizzazione del patrimonio culturale non può infatti limitarsi alla sua preservazione, poiché è intrinsecamente legata alle condizioni di fruizione e al contesto sociale che ne determinano la vitalità e la produzione di significato (Arenghi *et al.*, 2019). L'accessibilità e l'inclusione culturale sono dunque risorse collettive capaci di generare conoscenza, benessere e coesione sociale. La ricerca qui presentata si colloca nell'ambito delle arti performative e indaga l'utilizzo della Realtà Virtuale (VR) come strumento per rendere la produzione teatrale accessibile anche a pubblici impossibilitati a recarsi fisicamente a teatro, con l'obiettivo di valutare se la VR possa essere impiegata come infrastruttura culturale di accesso, e con quali benefici e limiti. Dopo aver delineato un quadro sulla performance digitale e sulle trasformazioni introdotte dai processi di mediatizzazione, il contributo illustra un caso di studio realizzato in collaborazione con il Teatro Stabile di Torino. Il progetto ha coinvolto pazienti ricoverati negli hospice di Torino e provincia, insieme ai loro caregiver, con l'obiettivo di testare una modalità di fruizione teatrale che favorisca accesso e partecipazione culturale in contesti di cura. I risultati preliminari indicano che l'immersività offerta dalla VR può favorire un più intenso coinvolgimento narrativo e una percezione della presenza scenica anche in assenza di co-presenza fisica, valorizzando la fruizione culturale in condizioni di isolamento. Al tempo stesso, sono emerse criticità legate alla natura dei contenuti proposti e alla loro adeguatezza nei confronti dei pubblici coinvolti, un aspetto cruciale da considerare nell'ottica di possibili applicazioni future da parte di istituzioni teatrali.

2. QUADRO TEORICO

Le contaminazioni digitali nelle arti performative non sono una novità; tuttavia, il teatro ha assistito a una moltiplicazione di sperimentazioni tecnologiche soprattutto negli ultimi decenni. Una definizione di teatro digitale deve tenere conto dell'eterogeneità di tali pratiche, che spaziano dall'integrazione delle tecnologie nel processo creativo – generando forme di intermedialità (Masura 2020) o mixed reality (Benford & Giannachi 2011) – fino allo sviluppo di nuovi canali per la diffusione delle performance (Sullivan 2020). Steve Dixon propone di definire la performance digitale come "all performance works where computer technologies play a key role rather than a subsidiary one in content, techniques, aesthetics, or delivery forms" (Dixon 2007: 3). In questo ampio spettro, la performance in Realtà Virtuale (VR) rappresenta una declinazione specifica della dialettica tra teatro e digitale. La VR è particolarmente rilevante perché, pur operando in un ambiente mediatizzato, "produce effetti nel qui ed ora" (Pizzo 2003: 79), instaurando un rapporto con la performance che "non appare come una dialettica dei contrari quanto come un più delicato e sfumato attrarsi tra simili" (Ivi: 22). Come osserva Coco, "theater and virtual reality, suggest a common unity of intent aimed at realizing the utopia of a phantasmatic encounter" (Coco 2024: 331). Studi provenienti da diversi ambiti disciplinari (mediologia, informatica, scienze cognitive e psicologia) convergono nel sostenere che la VR possa intensificare il senso di presenza e il coinvolgimento, generando risposte analoghe a quelle sperimentate in contesti non virtuali (Slater *et al.* 2009; Fuchs *et al.* 2011; Slater, Sanchez-Vives 2016; Kilteni *et al.* 2012). In questo ambito di ricerca, il concetto di presenza si riferisce a una qualità esperienziale legata alla percezione soggettiva di essere situati in un luogo, uno stato cognitivo e attenzionale che può emergere anche in ambienti virtuali (Mestre 2018; Wilkinson *et al.* 2021). Negli studi della performance, e in particolare nel ramo di ricerca relativo al contesto mediatico contemporaneo, la 'presenza', o *liveness*, acquisisce un'accezione diversa, seppur similmente determinata dalla percezione di contiguità, in questo caso con l'evento performativo cui si assiste. In particolare, Auslander (1999) ha delineato una cornice fenomenologica per l'analisi della performance *live*, successivamente ampliata grazie a ricerche empiriche rivolte alla risposta del pubblico (Barker 2012; Sanden 2013). In una recente pubblicazione di Gemini e Brilli (2023), gli autori hanno proposto una sistematizzazione delle diverse forme di *liveness* in base al grado di continuità spazio-temporale con l'evento e alla percezione di tale continuità da parte dello spettatore. Integrando la teoria della performance con prospettive provenienti dall'archeologia, Giannachi, Kaye e Shanks (2012) hanno condotto un'analisi critica della presenza, che non si limita all'esperienza individuale ma include anche le dimensioni relazionali e di co-presenza, concepite entro un'ecologia dell'ambiente in cui spazialità e temporalità sono strettamente intrecciate. In questo quadro, presenza e *liveness* si configurano come

categorie relazionali e stratificate, la cui definizione si complica con la mediatizzazione della performance, che introduce nuove forme di continuità e discontinuità spazio-temporale tra evento, medium e spettatore.
Con la pandemia da Covid-19, la sperimentazione con il digitale è divenuta una necessità operativa (Charlesworth 2022; Monteverdi 2023; Brilli *et al.* 2023; Aebischer & Nicholas 2024), evidenziando il potenziale delle tecnologie immersive nel raggiungere pubblici remoti o con limitata mobilità. In questo senso, la VR può essere intesa come una modalità alternativa di produzione della presenza, distinta dalla co-presenza fisica ma potenzialmente capace di generare forme analoghe di coinvolgimento esperienziale tramite processi di rimediazione (Bolter & Grusin 1999). Con il progressivo allentamento delle restrizioni pandemiche, la sperimentazione digitale ha conosciuto una parziale rifocalizzazione verso ambiti più circoscritti. Nondimeno, le tecnologie immersive possono configurarsi come infrastrutture di accesso culturale per pubblici remoti, contribuendo a ridurre barriere temporanee o strutturali alla fruizione teatrale, quali quelle legate a condizioni di degenza prolungata.

3. METODOLOGIA E DISEGNO DI RICERCA

Il progetto è stato strutturato sulla base di due casi di studio relativi a due differenti contenuti teatrali. Il primo consiste in un breve estratto dell'*Amleto* con la regia di Leonardo Lidi, le cui riprese sono state svolte da un palchetto del teatro al fine di restituire il punto di vista tradizionale dello spettatore. Il secondo, diretto da Diego Pleuteri, è una breve *pièce*[1] dedicata a Piero Gobetti, intellettuale antifascista torinese, registrata all'interno del suo studio presso il *Centro studi Piero Gobetti.* In quest'ultimo caso la telecamera è stata collocata al centro della scena, generando un grado più elevato di immersività. Per valutarne l'impatto, è stata inoltre realizzata una versione in un tradizionale formato 2D, così da proporre ai partecipanti entrambe le modalità di fruizione e rilevare eventuali differenze nella ricezione. Dal punto di vista metodologico, è stata adottata una metodologia mista, che integra la compilazione di alcuni brevi questionari prima e dopo l'esperienza di fruizione delle performance VR, con interviste discorsive semistrutturate relative alla familiarità pregressa con il teatro e con l'uso del visore, e alla percezione dell'esperienza (comprensione, attenzione, senso di presenza narrativa, coinvolgimento emotivo e valutazione complessiva). Inoltre, è stato adottato un approccio di ricerca partecipativa, riconoscendo il ruolo degli operatori sanitari e dei volontari come portatori di competenze situate nel contesto di cura. Anzitutto, durante la fase di definizione del disegno di ricerca sono stati consultati i coordinatori e gli psicologi degli hospice, al fine di identificare strumenti di rilevazione adeguati e di valutare la fattibilità dell'intervento nel rispetto delle condizioni cliniche dei pazienti. La definizione del protocollo delle interviste, così come la scelta dei questionari impiegati, è stata dunque realizzata in collaborazione con il personale degli enti aderenti alla ricerca. Tale fase è stata seguita da giornate di osservazione preliminare in loco, utili per conoscere gli operatori, comprendere l'organizzazione del contesto e definire protocolli operativi compatibili con le esigenze dell'assistenza quotidiana. Inoltre, sebbene la co-progettazione dei contenuti non rientrasse negli obiettivi del progetto – poiché la produzione artistica è rimasta sotto la responsabilità del Teatro e non è stata oggetto di negoziazione – la ricerca ha previsto un coinvolgimento continuativo degli operatori sanitari, che sono stati invitati a fornire riscontri qualitativi attraverso interviste discorsive e *focus group* paralleli alle interviste con i pazienti. L'integrazione di queste competenze professionali ha consentito di contestualizzare più accuratamente i risultati preliminari e di individuare criticità operative rilevanti per l'implementazione futura di progetti analoghi, contribuendo a una produzione di conoscenza situata (Haraway 1988) e a una co-produzione epistemica tra ricercatori accademici e attori istituzionali del contesto di cura. L'intero protocollo di ricerca, comprensivo delle procedure di raccolta dati, dei criteri di inclusione ed esclusione e delle modalità di somministrazione delle interviste, è stato preventivamente valutato e approvato dal Comitato di Bioetica dell'Università di Torino, garantendo la conformità alle normative vigenti e la tutela dei partecipanti in un contesto clinico particolarmente sensibile. L'attuazione del progetto ha evidenziato alcune criticità strutturali. In particolare, il numero di pazienti disponibili alla partecipazione è risultato limitato, sia per le condizioni cliniche variabili sia per la necessità di contemperare l'intervento con i tempi e le priorità dell'assistenza. Tale variabilità ha comportato un campione numericamente ridotto e caratterizzato da eterogeneità nei tempi di fruizione e nel livello di attenzione. Inoltre, il contesto stesso dell'hospice ha introdotto un alto grado di imprevedibilità, rendendo complessa la raccolta di dati replicabili secondo una scansione

[1] *Appunti per una rivoluzione liberale*, di Diego Pleuteri, visionabile al link: https://youtu.be/H3kveiuVC5A.

temporale fissa. Il numero di questionari raccolti non è sufficiente, allo stato attuale, per sostenere un'analisi statistica con significatività inferenziale: sono stati coinvolti, al momento della stesura del contributo, 17 pazienti, 12 parenti e 30 tra operatori e volontari, distribuiti su quattro hospice a Torino e provincia. Per questa ragione i dati quantitativi saranno utilizzati in senso esplorativo, come indicazioni preliminari utili a identificare le tendenze e le ricorrenze emerse. Al tempo stesso, i limiti del campione hanno favorito una maggiore profondità nell'analisi qualitativa, permettendo di ricostruire con precisione le condizioni soggettive di fruizione e l'impatto emotivo dell'esperienza. Nonostante i limiti operativi, la combinazione tra osservazione diretta, strumenti esplorativi e contributi qualitativi degli operatori ha consentito di sviluppare un quadro preliminare coerente, utile a orientare eventuali iniziative future e a suggerire linee guida applicabili in contesti affini.

4. RISULTATI PRELIMINARI E DISCUSSIONE

Una discussione sistematica dei risultati dello studio risulta prematura, poiché la raccolta dati è tuttora in corso e l'analisi si trova in una fase preliminare. Ciononostante, emergono già alcune ricorrenze e temi rilevanti. Le prime osservazioni riguardano tre principali aree: l'usabilità e tollerabilità del visore; la comprensione e il carico cognitivo legato ai contenuti; e gli effetti relazionali.

Dal punto di vista del medium, sebbene siano state rilevate alcune criticità legate all'esperienza d'uso del visore, gli episodi di *cybersickness* hanno riguardato un numero estremamente limitato di partecipanti, i quali presentavano già sintomi pregressi quali vertigini, affaticamento visivo o instabilità posturale prima dell'esposizione. Al fine di ridurre le difficoltà operative, i partecipanti sono stati assistiti durante l'intero processo, comprendendo la selezione dei contenuti, la configurazione dell'audio, una spiegazione semplificata dell'utilizzo del controller per la rotazione del punto di vista (particolarmente rilevante per i partecipanti allettati e impossibilitati a muovere il capo) e il supporto nel posizionamento del visore. È stato inoltre esplicitato che l'assistenza poteva essere richiesta in qualsiasi momento e che eventuali fastidi dovevano essere segnalati tempestivamente per interrompere o sospendere l'esperienza. Prima dell'esposizione, sono stati rilevati eventuali sintomi potenzialmente associati alla *cybersickness*, al fine di orientare la scelta dei contenuti o sconsigliare la partecipazione quando opportuno. I risultati preliminari dei questionari pre e post indicano che, in assenza di sintomi pregressi, non sono stati riportati effetti avversi; tra i partecipanti con sintomi iniziali, solo un numero relativamente ridotto ha preferito interrompere la visione per senso di disagio (ad esempio affaticamento visivo o vertigini), senza aggravamenti successivi. Tuttavia, la limitata autonomia nell'uso del dispositivo e, in alcuni casi, la riluttanza a richiedere assistenza hanno generato difficoltà operative, quali impostazioni visive o sonore inadeguate e un utilizzo errato del controller, che ha richiesto l'intervento dell'operatrice. Tali criticità si inseriscono in una più ampia riflessione sul *digital divide*, inteso come fenomeno multidimensionale che comprende disuguaglianze nell'accesso alle tecnologie, nelle competenze digitali, nelle modalità d'uso e nei benefici derivanti dall'adozione tecnologica (Dijk 2020; Warschauer 2003). In ambito assistenziale, queste disuguaglianze risultano ulteriormente amplificate da fattori legati all'età, spesso elevata nei pazienti ricoverati in hospice, alle condizioni fisiche e al contesto istituzionale: tali variabili assumono qui dunque un rilievo particolarmente critico, suggerendo la necessità di protocolli strutturati di mediazione tecnologica e di un supporto umano continuativo.

In relazione ai contenuti, l'analisi deve essere condotta su più livelli, considerando l'eterogeneità dei materiali proposti e le diverse forme di immersività. È significativo che il consenso generale tra gli operatori sanitari – medici, infermieri, psicologi – e tra i volontari degli hospice, soggetti non professionali ma formati al contesto assistenziale, abbia evidenziato un giudizio critico sui contenuti selezionati. In particolare, la scena selezionata dell'*Amleto*, pur rielaborata in chiave comica e arricchita da riferimenti contemporanei, è stata percepita come "troppo confusionaria" e dinamica; mentre il contenuto su Piero Gobetti, seppur descritto come "emotivamente più tranquillo", è stato considerato come eccessivamente complesso sul piano tematico e linguistico. Dal punto di vista della ricezione, l'immersività spaziale del caso Gobetti ha spesso prodotto una maggiore attenzione all'ambiente scenico a scapito del dialogo, con alcune eccezioni rilevanti in cui i partecipanti hanno dichiarato una maggiore concentrazione sulle parole in virtù dell'isolamento percettivo prodotto dal visore. Questo dato suggerisce una duplice implicazione. Da un lato, l'immersività sembra rendere più complessa la ricezione del dialogo, e in diversi casi la comprensione del contenuto è stata conseguita grazie alla mediazione interpretativa dell'operatrice e a un contesto preliminare, nonché agevolata dalla duplice modalità di fruizione (immersiva e su schermo) progettata per confrontare le diverse forme di rimediazione. Dall'altro, per partecipanti con un basso

livello di istruzione, l'attenzione allo spazio, ai movimenti degli attori e agli oggetti scenici (fotografie d'epoca, arredi, libri, lettere, una macchina da scrivere) ha mantenuto un elevato grado di attrattività dell'esperienza, che probabilmente sarebbe risultato ridotto in una fruizione esclusivamente bidimensionale. Pur tenendo conto delle considerazioni critiche degli operatori, che suggeriscono l'adozione di performance meno complesse in applicazioni future, la ricezione si è rivelata fortemente eterogenea. Nel caso della *pièce* su Gobetti, alcuni partecipanti hanno manifestato interesse per il contenuto storico, indipendentemente dal capitale culturale posseduto, sottolineando il valore dell'esperienza come occasione di scoperta (una signora, dopo la visione, ha affermato con entusiasmo: "Io l'ho trovato bellissimo, bello perché alla scoperta di cose nuove, cose che io non avevo mai appurato, ma anche non mi avevano mai fatto notare"). Altri hanno attivato processi di memoria autobiografica, in particolare in relazione agli oggetti scenici, come la macchina da scrivere, evidenziando la capacità dell'ambiente immersivo di evocare ricordi personali. Analogamente, nel caso dell'*Amleto*, alcuni partecipanti hanno reagito con divertimento o rilassamento, mentre altri hanno espresso difficoltà di comprensione o disinteresse per le scelte registiche. Nel complesso, tali risultati suggeriscono che il carico cognitivo generato dall'ambiente immersivo e dai contenuti narrativi è altamente soggettivo e dipendente da variabili individuali e contestuali. Questi dati mettono in discussione approcci paternalistici alla progettazione dei contenuti per pubblici vulnerabili, mostrando come anche in contesti di fragilità emergano molteplici preferenze estetiche e narrative. Un caso emblematico è rappresentato dalla richiesta di una paziente di 63 anni malata di cancro, che ha chiesto se ci fosse anche "qualcosa di Pirandello, tipo *L'uomo dal fiore in bocca*", un atto unico il cui protagonista è destinato a morire a breve: questa richiesta indica come possa emergere in alcuni il desiderio di esperienze catartiche e tematicamente affini alla propria condizione esistenziale. In prospettiva applicativa, i risultati suggeriscono l'opportunità di offrire un ventaglio diversificato e personalizzabile di contenuti, analogamente alla scelta repertoriale tipica dell'esperienza teatrale dal vivo, al fine di rispettare l'autonomia e l'agentività dei fruitori: non si può interpretare la vulnerabilità come incapacità interpretativa, come è stato discusso nell'ambito dei *disability studies*. Dunque, pur essendo basati su un numero limitato di partecipanti e su un'analisi preliminare, i risultati suggeriscono che le tecnologie immersive possano generare esperienze culturali significative anche in contesti di fragilità, contribuendo a modalità alternative di fruizione più accessibili e inclusive. In particolare, la VR consente una temporanea dislocazione dall'ambiente clinico, offrendo uno spazio esperienziale distinto che alcuni partecipanti hanno percepito come positivo. I questionari pre e post, insieme all'osservazione diretta e alle interviste discorsive, indicano una tendenza alla riduzione di emozioni negative (come tristezza o ansia) e un'elevata attribuzione di stupore, soprattutto nei partecipanti che sperimentavano per la prima volta il visore, che costituiscono la grande maggioranza. Un elemento emerso con chiarezza riguarda il ruolo cruciale della mediazione relazionale: le percezioni e le aspettative dei partecipanti risultano fortemente influenzate dai discorsi e dalle conoscenze pregresse sull'esperienza VR. Questo conferma che l'esperienza immersiva non può essere considerata unicamente come interazione uomo-macchina, ma come un dispositivo sociotecnologico situato. Al contempo, la dimensione relazionale si configura come una potenzialità: la fruizione condivisa (tra pazienti e caregiver) ha spesso attivato momenti di reminiscenza, dialogo e negoziazione di significato, suggerendo che l'esperienza immersiva può funzionare come catalizzatore di interazioni sociali e memorie culturali. Questi risultati indicano che la VR, se accompagnata da adeguate pratiche di mediazione umana, può sostenere forme di partecipazione culturale che vanno oltre il consumo individuale, aprendo prospettive per modelli di fruizione relazionale anche per chi è impossibilitato a prendere parte alla co-presenza fisica dello spazio teatrale tradizionale. In conclusione, sebbene l'indagine si collochi in una fase preliminare e si basi su un campione numericamente limitato, i risultati suggeriscono che la realtà virtuale possa costituire uno strumento rilevante di rimediazione teatrale per raggiungere pubblici remoti e ampliare le modalità di accesso alle arti performative in contesti di marginalità. Tale potenziale, tuttavia, non può essere considerato intrinseco alla tecnologia, ma dipende in maniera determinante dalle condizioni di implementazione, dalla qualità dei contenuti e dalla presenza di adeguate forme di mediazione umana. L'uso della VR in contesti di cura solleva infatti questioni critiche legate alla sostenibilità tecnologica – in termini di dipendenza da specifici hardware, obsolescenza e scalabilità – nonché al rischio di generare nuove barriere, qualora l'accesso venga concepito esclusivamente come disponibilità del dispositivo e non come pratica relazionale. Alla luce di questi elementi, la VR può essere interpretata non come una soluzione neutra o universalmente inclusiva, bensì come un dispositivo socio-tecnologico capace di attivare un'ecologia complessa, in cui si intrecciano percezione sensoriale, agentività interpretativa e relazioni di cura. È

proprio all'interno di questa ecologia, e non nella tecnologia in sé, che si gioca la possibilità di una fruizione culturale significativa e non escludente.

BIBLIOGRAFIA

Aebischer, P., & Nicholas, R. (A c. Di). (2024). Adaptation and resilience in the performing arts. The pandemic and beyond. Manchester University Press.

Arenghi, A., Baraldi, L., & Garofolo, I. (2019). "Accessibility to cultural heritage between tangible and intangible". In Conservation/Consumption. Preserving the Tangible and Intangible Values, (A c. Di) Fiorani, D., Franco, G., Kealy, L., Musso, S., & Calvo-Salve, M. European Association for Architectural Education.

Auslander, P. (1999). Liveness: Performance in a Mediatized Culture. Routledge.

Barker, M. (2012). Live to your local cinema: the remarkable rise of livecasting. Palgrave Mcmillan. Benford, S., & Giannachi, G. (2011). Performing Mixed Reality. MIT Press.

Brilli, S., Gemini, L., & Giuliani, F. (2023). Theatre without theatres: Investigating access barriers to mediatized theatre and digital liveness during the covid-19 pandemic. Poetics, 97, 101750.

https://doi.org/10.1016/j.poetic.2022.101750

Bolter, J. D., & Grusin, R. (1999). Remediation. Understanding New Media. MIT Press.

Charlesworth, H. L. (2022). Nightmare Time and a Case Study for Digital Theatre During the COVID-19 Pandemic. Networking Knowledge: Journal of the MeCCSA Postgraduate Network, 15(1), Articolo 1.

https://ojs.meccsa.org.uk/index.php/netknow/article/view/660

Coco, G. (2024). Virtual and Real: Ideas and Tools for an Enhanced Theatre in an Empathic Direction.

Mimesis Journal, 13(2), 337–337. https://doi.org/10.13135/2389-6086/10001 Dick, J. van. (2020). The Digital Divide. Polity Press.

Dixon, S. (2007). Digital Performance: A history of new media in theatre, dance, performance art, and installation. The MIT press.

Fuchs, P., Moreau, G., & Guitton, P. (A c. Di). (2011). Theoretical and pragmatic approach to virtual reality

(0 ed., pp. 33–68). CRC Press. https://doi.org/10.1201/b11612-8

Gemini, L., & Brilli, S. (2023). Gradienti di Liveness: Performance e comunicazione dal vivo nei contesti mediatizzati. Angeli.

Giannachi, G., Kaye, N., & Shanks, M. (A c. Di). (2012). Archaeologies of presence: Art, performance and the persistence of being. Routledge.

Haraway, D. (1988). Situated Knowledges: The Science Question in Feminism and the Privilege of Partial

Perspective. Feminist Studies, 14(3), 575–599. https://doi.org/10.2307/3178066

Kilteni, K., Groten, R., & Slater, M. (2012). The Sense of Embodiment in Virtual Reality. Presence:

Teleoperators and Virtual Environments, 21(4), 373–387. https://doi.org/10.1162/PRES_a_00124

Masura, N. (2020). Digital Theatre. The Making and Meaning of Live Mediated Performance. Palgrave Macmillan.

Mestre, D. (2018). Presence in Virtual Reality: Insights from fundamental and applied research. Electronic Imaging, 3, 4331-4335. 10.2352/ISSN.2470-1173.2018.03.ERVR-433

Monteverdi, A. M. (2023). Residenze Digitali: Lo spettacolo si fa on line. Umanistica Digitale, (15), Articolo

15. https://doi.org/10.6092/issn.2532-8816/16888

Pizzo, A. (2003). Teatro e mondo digitale: Attori, scena e pubblico. Marsilio.

Sanden, Paul. (2013). Liveness in modern music: musicians technology and the perception of performance. Routledge.

Slater, M., Lotto, B., Arnold, M. M., & Sanchez-Vives, M. V. (2009). How we experience immersive virtual environments: The concept of presence and its measurement. Anuario de Psicología, 40.

Slater, M., & Sanchez-Vives, M. V. (2016). Enhancing Our Lives with Immersive Virtual Reality. Frontiers in Robotics and AI, 3. https://doi.org/10.3389/frobt.2016.00074

Sullivan, E. (2020). Live to your living room: Streamed theatre, audience experience, and the Globe's A

Midsummer Night's Dream. Participations: Journal of Audience & Receptions Studies, 17, 92–119.
Warschauer, M. (2003). Technology and Social Inclusion. Rethinking the Digital Divide. The MIT Press.
Wilkinson, M., Brantley, S., & Feng, J. (2021). A Mini Review of Presence and Immersion in Virtual Reality. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 1099-1103. https://doi.org/10.1177/1071181321651148

[4/C] Formazione, didattica e laboratori DH

Gli Hunger Games del LabCD. Un diario di sopravvivenza

ENRICA SALVATORI[1], VITTORE CASAROSA[2], FRANCESCO RICCIARDI[3]
[1]UNIVERSITÀ DI PISA, AIPH - ENRICA.SALVATORI@UNIPI.IT
[2]ISTI CNR - CASAROSA@ISTI.CNR.IT
[3]UNIVERSITÀ DI PISA - F.RICCIARDI9@STUDENTI.UNIPI.IT

ABSTRACT (ITALIANO)

Il contributo presenta un caso di studio sulla conservazione dei progetti digitali di ricerca attraverso l'esperienza del Laboratorio di Cultura Digitale dell'Università di Pisa (LabCD), attivo dal 2011 nell'ambito delle *Digital Humanities*. A partire dalla storia e dalle pratiche del Laboratorio, il saggio analizza le difficoltà strutturali incontrate nel garantire continuità, accessibilità e riuso a un patrimonio digitale eterogeneo per finalità, tecnologie e livelli di maturità, sviluppato in un contesto fortemente sperimentale. Vengono messi in evidenza problemi comuni a molti centri universitari: dipendenza da infrastrutture informatiche sottodimensionate, vulnerabilità agli attacchi esterni, discontinuità delle competenze tecniche e assenza di politiche istituzionali di lungo periodo per la conservazione. Il contributo discute quindi alcune strategie pragmatiche adottate dal LabCD, tra cui il *web archiving* dei siti come oggetti complessi, l'estrazione e la preservazione selettiva dei dati e il deposito di risorse strutturate in *repository* disciplinari. Il caso mostra come, in assenza di soluzioni complessive, la conservazione dei progetti digitali richieda scelte consapevoli, flessibili e orientate alla sostenibilità e al riuso futuro.
Parole chiave: Digital Humanities; conservazione digitale; web archiving; progetti di ricerca; repository

ABSTRACT (ENGLISH)

The paper presents a case study on the preservation of digital research projects through the experience of the Laboratory of Digital Culture at the University of Pisa (LabCD), active since 2011 in the field of Digital Humanities. Drawing on the history and practices of the Lab, the contribution analyzes the structural difficulties involved in ensuring continuity, accessibility, and reuse for a heterogeneous digital heritage developed in a strongly experimental context. It highlights issues common to many university research centers, including reliance on under-resourced IT infrastructures, exposure to external cyber threats, discontinuity of technical expertise, and the lack of long-term institutional preservation policies. The paper then discusses a set of pragmatic strategies adopted by LabCD, such as web archiving of websites as complex objects, selective extraction and preservation of data, and the deposit of structured resources in disciplinary repositories. The case study shows that, in the absence of comprehensive solutions, digital project preservation currently requires conscious, flexible, and sustainability-oriented choices aimed at future reuse.
Keywords: Digital Humanities; digital preservation; web archiving; research projects; repositories

1. INTRODUZIONE

In un contributo scientifico standard si individua un problema - che in questo caso sarebbe la conservazione di progetti digitali di contenuto e forma variabile -, si illustra lo stato dell'arte e il metodo applicato, traendo infine conclusioni più o meno propositive a livello delle soluzioni adottate e adottabili. Il caso che si presenta è un po' diverso. Illustrando la storia del Laboratorio di Cultura Digitale dell'Università di Pisa (fondato nel 2011, da qui in poi LabCD), si vuol portare innanzitutto all'attenzione dei lettori le difficoltà incontrate negli ultimi anni nel mantenimento e nella conservazione dei suoi prodotti digitali. L'idea è quella di offrire un esempio delle problematiche, non solo squisitamente tecniche, che possono incontrare i centri di ricerca universitari attivi nell'ambito delle Digital Humanities che intendano garantire un certo grado di "vita" ai loro progetti. Anche se gli autori sono consapevoli del dibattito che si sta svolgendo in questi ultimi anni sul tema della sostenibilità digitale (Edmond & Morselli, 2020; Goddard & Seeman, 2019; Smithies *et alii*, 2019), la finalità della proposta non è tanto quella di dare un contributo innovativo e propositivo alla discussione, quanto invece di mettere in luce come, nella concretezza del lavoro quotidiano, non esista una "buona" soluzione pratica, ma solo un insieme di

strategie, fatalmente flessibili e fortemente disomogenee, da adottare a seconda delle circostanze e soprattutto come ci sia ancora bisogno di enfatizzare l'importanza della conservazione dei progetti digitali. La speranza è che questo diario di bordo, nell'illustrare una continua lotta per la sopravvivenza, serva a fare crescere in realtà simili la consapevolezza sulla mancanza di soluzioni tecniche complessive e di infrastrutture solide e di supporto, quantomeno a livello universitario, e su quanto questo freni, di fatto, la sperimentazione.

2. LA MIETITURA

Il LabCD dell'Università di Pisa è nato nel 2011 come Centro interdipartimentale di formazione e ricerca, a seguito dello sviluppo delle ricerche e dell'attività didattica del corso di laurea di Informatica Umanistica - triennale e magistrale - del medesimo ateneo. Nel corso dei suoi 15 anni di vita vi hanno collaborato docenti di cinque dipartimenti e numerosi studiosi indipendenti, riuniti dal desiderio di sviluppare progetti, strumenti e conoscenze che unissero l'ambito umanistico e quello informatico. Il Laboratorio ha funzionato (e continua a funzionare) come una sorta di bottega digitale, un *DigiCraft,* nel senso che lavora a progetti digitali integrando le diverse capacità e competenze di docenti, esperti indipendenti, dottorandi, laureandi e tirocinanti. Il suo DNA è sempre stato quindi quello della sperimentazione, che gli deriva fin dalle origini dal corso di laurea in Informatica Umanistica (nato nel 2002) e dall'allora polo informatico dedicato all'ambito umanistico (CISIAU).
Molte idee progettuali sono state costruite e più o meno rapidamente distrutte, altre sono restate in piedi un po' più a lungo: la "mietitura" - prendendo a prestito il termine che nei romanzi e nei film degli *Hunger Games* indica la scelta degli eroi (i "tributi")[1] - lasciava molto spazio a iniziative spontanee e anche a proposte poco strutturate, nella convinzione che fosse proprio il brulicare delle idee a rendere fertile il terreno della sperimentazione. I "caduti" sono stati quindi numerosi, sebbene eroici: la strada della sperimentazione sull'e-learning, ad esempio (dal podcasting alla costruzione di MOOC), non ha trovato l'accoglienza necessaria per essere trasformata in una strategia di Ateneo; un innovativo progetto su Second Life (Denard *et alii* 2011) si scontrò invece con l'impossibilità burocratica di acquistare un'"isola" con Linden Dollar. Possiamo ascrivere queste perdite alle difficoltà insite in tutte le attività pionieristiche che all'epoca, per quanto attiene le *Digital Humanities*, andavano a braccetto con la stessa battaglia che il settore stava intraprendendo per il proprio riconoscimento (l'AIUCD nasceva nel 2011). Altri combattenti, messi in campo forse in tempi più maturi, si fecero invece strada con maggiore successo: il progetto Erasmus Plus *A Digital Journey in Europe* (2015-2018) poi gemmato nell'Erasmus Plus *Shelfie* (2021-2023) in sostegno al processo di digitalizzazione delle scuole, il sostegno dato allo sviluppo di EVT1 tramite l'*edizione digitale del Codice Pelavicino* (2014-2020) e la digitalizzazione e pubblicazione con criteri di Public History del patrimonio dell'Archivio Comunale di Vicopisano (Salvatori, 2023)[2] .

3. L'ARENA

Nel momento in cui scriviamo (aprile 2026) il sito del LabCD e tutti i siti connessi sono in gran parte irraggiungibili: l'oscuramento, dovuto a un attacco informatico esterno e malevolo, è certamente temporaneo, ma illumina un problema strutturale. I siti sono mantenuti dal polo informatico dell'area umanistica che fa capo al servizio informatico dell'Ateneo. La struttura risponde giornalmente a numerose richieste ed esigenze provenienti da diversi dipartimenti ed è - come usuale in Italia - sottodimensionata a livello di personale. Il contesto strutturale comporta quindi che, in caso di attacco informatico esterno, si crei automaticamente una procedura di *triage*, in cui i siti vengono "rianimati" con una tempistica che dipende da una valutazione di priorità operata dalla struttura stessa. Le realtà che sono percepite necessarie o più utili al funzionamento delle strutture universitarie hanno ovviamente la precedenza e i progetti sperimentali di ambito culturale, che non erogano servizi "concreti", ricevono un codice verde o bianco.
Nel moltiplicarsi di questi episodi e nell'abbondanza dei progetti digitali creati non solo dal LabCD, viene "naturalmente" preso in considerazione anche il peso accademico del responsabile scientifico del progetto, fatalmente variabile a seconda del momento e condizionato anche dall'inevitabile *turn-over.*

[1] *The Hunger Games* è un romanzo distopico scritto da Suzanne Collins (2008), primo di una trilogia di grande successo, formata anche da *Catching Fire* (2009) e *The Mockingjay* (2010). Ne sono stati tratti quattro film usciti tra 2012 e 2015.
[2] Si vedano https://shelfie.labcd.unipi.it/, https://adigitaljourney.labcd.unipi.it/, https://archivicodigitale.labcd.unipi.it/.

Queste condizioni hanno reso indispensabile per il LabCD indirizzare risorse ed energie alla elaborazione di una strategia di conservazione che potesse ricorrere a realtà esterne all'Università, in grado di garantire sufficiente spazio di archiviazione e una discreta salvaguardia ai progetti sopravvissuti in un'arena flagellata da attacchi continui e da un'attenzione interna altalenante.

4. LA CORNUCOPIA

Data la vocazione fortemente sperimentale del LabCD appena descritta, non sono mai state definite a priori le specifiche tecniche delle proposte progettuali da promuovere. Se usare Wordpress o Omeka o altro CMS, se invece creare siti statici o puntare sull'interazione tramite l'uso di linguaggi lato server (PHP, Python, ASP.net) e database (MySQL, PostgreSQL), dipende da come i docenti, gli esperti e gli studenti valutano di volta in volta gli obiettivi di ricerca, e competenze disponibili nel team e le eventuali richieste e abilità informatiche dei committenti (in caso di richiesta da terzi).
Gli scopi stessi del LabCD hanno quindi prodotto un patrimonio di "superstiti" molto disomogeneo per scopi e caratteristiche tecniche interne, queste ultime spesso determinanti per la loro sopravvivenza: se i siti statici sono tanto stabili e longevi quanto poco aggiornabili e attraenti, quelli dinamici richiedono competenze per la loro modifica che non sono costantemente disponibili; se nei CMS è necessario caricare continuamente nuovi plug-in per rendere più sicuri i prodotti, alcuni aggiornamenti "massicci" richiedono l'intervento dei sistemisti, con la conseguente attivazione della procedura di triage già descritta.

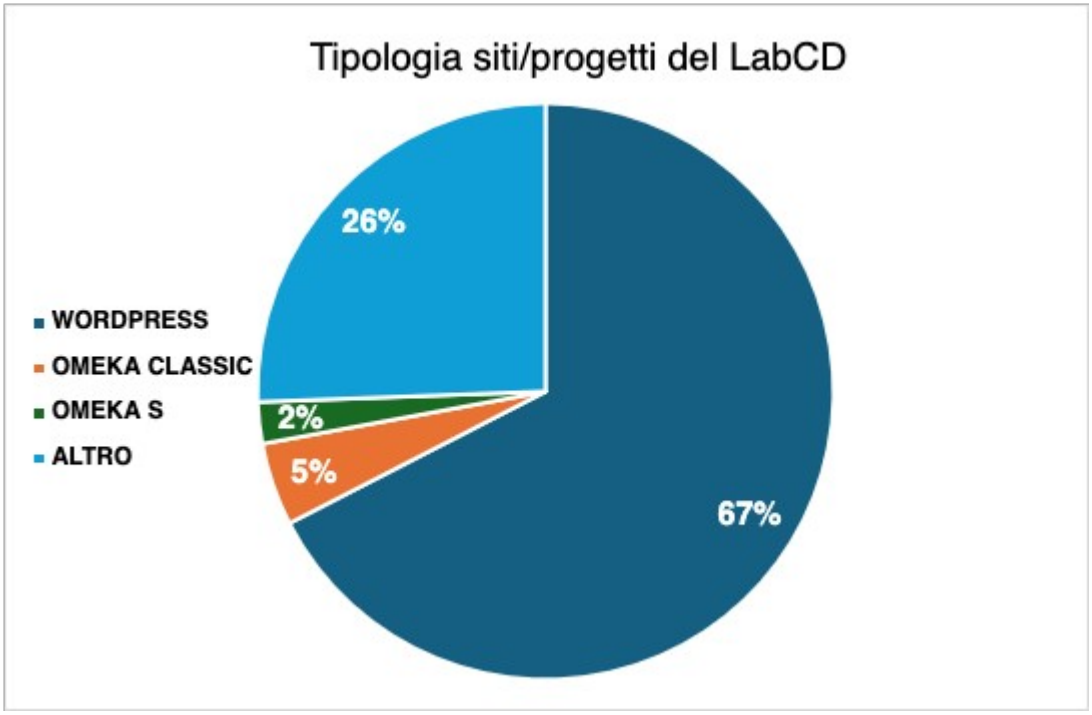


Fig. 1 Distribuzione per tipologie dei siti del LabCD.

Tale disomogeneità è stata affrontata, in questa fase di revisione e di "messa in sicurezza", innanzitutto operando un censimento e una tipizzazione dei progetti gestiti, valutando poi quali "condannare a morte" con le nostre stesse mani, perché valutati incompleti o poco significativi. In seguito abbiamo guardato all'insieme degli strumenti e delle risorse che avevamo a disposizione, sia per catturare i progetti, scaricare e organizzare i dati, sia per archiviarli in qualche *repository*.
Un primo tentativo fu fatto con una tesi triennale di Informatica Umanistica, copiando sostanzialmente le idee della Wayback Machine. Furono sviluppate due piattaforme distinte: una di tipo server dedicata all'archiviazione vera propria dei siti relativi ai progetti del LabCD, la seconda realizzata in WordPress per l'accesso ai contenuti archiviati (Bertellotti 2019-2020). La piattaforma di archiviazione, utilizzabile soltanto da amministratori registrati (inizialmente approvati da un "super amministratore") permetteva di indicare un sito web da "archiviare", fornendo la relativa URL e specificando pochi parametri. Il sito veniva quindi archiviato in un deposito locale, in formato WARC. WARC è uno standard internazionale di

archiviazione web (ISO 28500:2017) concepito per combinare in un singolo file l'insieme delle risorse digitali alla base di una risorsa web (HTML, immagini, script, fogli di stile, metadati etc.) insieme alle informazioni necessarie per la loro successiva riproduzione. A differenza di formati più semplici, WARC può conservare sia i contenuti catturati sia i metadati associati alle richieste HTTP, consentendo di preservare uno snapshot di un sito web in un determinato momento, navigabile poi come se fosse il sito originale (al momento dello snapshot).

La seconda piattaforma (la Wayback Machine, realizzata in WordPress) permetteva a tutti gli utenti (senza bisogno di registrazione) di accedere all'archivio dei siti Web salvati. Individuato il progetto di interesse, un visualizzatore di file WARC ripristinava le pagine salvate e permetteva di navigare nel sito come se fosse l'originale. La Wayback Machine forniva varie funzioni per navigare tra i siti salvati (per data, per categoria, per descrizione), inclusa una finestra di ricerca alla Google. La sezione forse piu' importante era quella di "Salva il tuo sito". Veniva presentato un modulo da riempire con tutte le informazioni richieste per il salvataggio di un nuovo sito, modulo che veniva poi inviato a un amministratore della piattaforma di archiviazione per procedere possibilmente all'archiviazione di un nuovo sito.
L'idea (la speranza) per questo progetto era che tutto il sistema diventasse uno "strumento" dell'Ateneo, con i siti web salvati su un repository aperto al pubblico. Ma tutto ciò non è accaduto. Panem non ha recepito la supplica.

Un tentativo attualmente in corso sta utilizzando ArchiveBox, uno strumento open source per l'archiviazione web che funge da crawler e anche da archivio in modalità self hosted. Il software accetta una lista di URL e genera snapshot completi delle pagine includendo più formati oltre al WARC (HTML statici, PDF, wget) in maniera ridondante. ArchiveBox esegue una scansione del sito, raccogliendo tutte le risorse rilevanti e organizzandole in una struttura di directory con dati e metadati, e consente l'upload automatico dei file WARC su Internet Archive, facilitando il processo di archiviazione dei progetti e permettendo la consultazione tramite Wayback Machine (web.archive.org).
Nonostante lo snapshot permetta di visualizzare come era il sito in un dato momento, non permette però di utilizzare il sito (se dinamico) come se fosse l'originale, rendendo quindi non disponibili dati conservati nel database del sito. Tale approccio è quindi adeguato per i siti statici, ma presenta limiti significativi per quelli dinamici. Per ovviare a questa limitazione, e salvare in ogni caso i dati del progetto, è stato deciso di effettuare i dump dei vari database. A questo scopo è stato necessario individuare le tecnologie utilizzate per lo sviluppo dei singoli progetti e verificare l'eventuale uso di un Content Management System (CMS).
Per i siti realizzati con un CMS (ad esempio WordPress) è stato effettuato l'accesso come amministratore ed è stato generato un dump dei contenuti in formato XML. Successivamente, tramite appositi script Python, questi sono stati convertiti in una serie di file CSV, selezionando infine quei file contenenti dati ritenuti rilevanti ai fini dell'archiviazione.
Per i siti dinamici realizzati senza un CMS è stato effettuato l'accesso diretto al server via SSH, in modo da poter accedere direttamente al database e poterne effettuare un dump completo. In tal caso il dump potrà poi essere conservato nel formato originale (SQL), preservando la struttura e le relazioni tra tabelle, o convertito in formato CSV per facilitare la consultazione, perdendo però le relazioni logiche tra i dati.
Come si vede le soluzioni messe in campo sono state diverse, per alcune funzioni efficaci ma nel complesso nessuna veramente soddisfacente. Mantenendo la metafora degli Hunger Games spesso una soluzione "buona", nella facile previsione di una inevitabile scomparsa di un sito dinamico che offre funzionalità diverse è quella dello smembramento consapevole del progetto con estrazione di documenti e di dati. L'abbiamo adottata, ad esempio, per l'edizione digitale del Codice Pelavicino per proteggere "almeno" il cuore pulsante dell'edizione: il corpus dei documenti codificati. L'edizione, lo ricordiamo, è basata su immagini, con codifica digitale in formato XML/TEI e visualizzazione tramite il software open source EVT1, mentre una serie di ricchi materiali di corredo è pubblicata in un sito in Wordpress e la ricerca dinamica sulle named entities è possibili in un terzo ambiente Web. La fusione delle tre parti in un unico medium di visualizzazione è difficilmente realizzabile. Fiduciosi del fatto che l'insieme possa comunque sopravvivere abbastanza a lungo per trovare una soluzione più armoniosa, abbiamo pubblicato parte dei materiali di corredo su una rivista scientifica e scelto di concentrare gli sforzi di conservazione (per un eventuale riuso) nei 536 file XML che compongono l'edizione; contemporaneamente abbiamo

anche depositato sul repository ILC-CNR for CLARIN-IT i dataset dei delle named entities codificate (Salvatori & Casarosa, 2024; Salvatori Casarosa & Santoni, 2024).

5. CANTO DELLA VITTORIA CERCASI

Il problema che qui abbiamo presentato non è squisitamente tecnico ma culturale. Se esistono oggi dei repository sostenuti da fondi pubblici, come Zenodo o ILC-CNR for CLARIN-IT, questi sono finalizzati soprattutto a preservare i dati della ricerca, pubblicazioni, software, poster, slide, immagini, dataset, altri materiali correlati, e non realtà articolate come quelle qui descritte. La scelta di preservare interi siti, statici o dinamici che siano, deve passare attraverso il riconoscimento che è proprio la loro composizione interna, il legame creato tra dati, oggetti, e funzioni, a dover essere tutelato. Dato che si tratta di una scelta culturale e dato anche che non esiste una strategia europea o nazionale da questo punto di vista, dovrebbe essere l'istituzione che produce il progetto digitale stesso ad avvertire la "responsabilità" della conservazione. Nel contesto in cui operiamo una tale consapevolezza è miraggio lontano. All'università di Pisa è stato aperto di recente (2025) un datacenter con possibilità di storage su cloud in cui sono disponibili 5 TB per progetto, per un massimo 3 anni, prorogabili. Si tratta di un servizio certamente utile per la ricerca ma, anche in questo caso, strettamente legato ai dati della ricerca e finalizzato a un uso specifico, entro le équipe stesse di ricerca. Non è un archivio né una biblioteca digitale, non preserva né diffonde la memoria della ricerca, ne custodisce solo temporaneamente i dati.
E così i "prodotti" delle DH, in particolare quelli più difficilmente "smembrabili" e impacchettabili in dataset, sono destinati a morire, a scomparire uno dopo l'altro, uccisi da un *hacker* e non rianimati nel *triage*, legati all'inevitabile scomparsa (per pensionamento, morte o cambio di sede) dei creatori, o affidati per disperazione a strutture commerciali anch'esse effimere. Segno indubbio di una società che valorizza il progetto digitale finché lo usa - ossia finché è vivo -, finché lo diverte, fino a che sopravvive alla successiva stagione dei giochi.

BIBLIOGRAFIA

Denard, H., Salvatori, E. & Simi, M. (2011). Learning by Building in SL: A Reflection on an Interdisciplinary and International Experience. In G. Vincenti & J. Braman (A c. Di), *Multi-User Virtual Environments for the Classroom: Practical Approaches to Teaching in Virtual Worlds,* (pp. 134–158). IGI Global.

Bertellotti, F. (2019). *La conservazione dei siti Web del LabCD: un servizio per l'Ateneo* [Laurea triennale]. Corso di laurea in Informatica Umanistica. Università di Pisa.

Edmond, J., & Morselli, F. (2020). Sustainability of digital humanities projects as a publication and documentation challenge. Journal of Documentation, *76*(5), 1019–1031.

Goddard, L., & Seeman, D. (2019). Negotiating sustainability: Building digital humanities projects that last. In *Doing more digital humanities* (pp. 38–57). Routledge.

Salvatori, E. (2023). Costruendo un Digicraft: Storia, ambizioni e sfide del Laboratorio di Cultura Digitale. In E. Salvatori, S. Pelagatti, & C. Mannari (A c. Di), *Pionieri fra due culture. Informatica umanistica a Pisa in onore di Maria Simi* (p. articolo 9). Simonelli.

Smithies, J., Westling, C., Sichani, A.-M., Mellen, P., & Ciula, A. (2019). Managing 100 Digital Humanities Projects: Digital scholarship and archiving in King's Digital Lab. *Digital Humanities Quarterly*, *13*(1).

Salvatori, Enrica and Casarosa, Vittore, 2024, *Codice Pelavicino,* ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli", http://hdl.handle.net/20.500.11752/OPEN-1012

Salvatori, Enrica; Casarosa, Vittore and Santoni, Dario, 2024, *Enriched Data from Codice Pelavicino,* ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli", http://hdl.handle.net/20.500.11752/OPEN-1011

ISO (2017). *ISO 18500:2017 - Information and documentation –WARC file format.* International Organization for Standardization. https://www.iso.org/standard/68004.html

It’s Buildin’ Time: A Project-based Approach for Teaching Data Management to Cultural Heritage Graduate Students

SEBASTIAN BARZAGHI[1], ALBA D’ELIA[2], ESTHER MONTSERRAT GIORDANO[3], GIULIA GUIDARELLI[4], GIANLUCA PETROSILLO[5], ELISABETTA SABATTINI[6], FIAMMETTA SABBA[7], LUCIA SARDO[8]

[1]DEPARTMENT OF CLASSICAL PHILOLOGY AND ITALIAN STUDIES, UNIVERSITY OF BOLOGNA, BOLOGNA (BO), ITALY – SEBASTIAN.BARZAGHI2@UNIBO.IT
[2]DEPARTMENT OF CULTURAL HERITAGE, UNIVERSITY OF BOLOGNA, RAVENNA (RA), ITALY - ALBA.DELIA@STUDIO.UNIBO.IT
[3]DEPARTMENT OF CULTURAL HERITAGE, UNIVERSITY OF BOLOGNA, RAVENNA (RA), ITALY - ESTHER.GIORDANO@STUDIO.UNIBO.IT
[4]DEPARTMENT OF CULTURAL HERITAGE, UNIVERSITY OF BOLOGNA, RAVENNA (RA), ITALY - GIULIA.GUIDARELLI3@STUDIO.UNIBO.IT
[5]DEPARTMENT OF CULTURAL HERITAGE, UNIVERSITY OF BOLOGNA, RAVENNA (RA), ITALY - GIANLUCA.PETROSILLO2@STUDIO.UNIBO.IT
[6]DEPARTMENT OF CULTURAL HERITAGE, UNIVERSITY OF BOLOGNA, RAVENNA (RA), ITALY - ELISABETTA.SABATTINI@STUDIO.UNIBO.IT
[7]DEPARTMENT OF CULTURAL HERITAGE, UNIVERSITY OF BOLOGNA, RAVENNA (RA), ITALY - FIAMMETTA.SABBA@UNIBO.IT
[8]DEPARTMENT OF CULTURAL HERITAGE, UNIVERSITY OF BOLOGNA, RAVENNA (RA), ITALY - LUCIA.SARDO@UNIBO.IT

ABSTRACT (ENGLISH)

This paper presents the results of the course *Digital Humanities and Data Management for Cultural Heritage*, integrated into the *Library and Archive Science* master’s at the University of Bologna. While traditional information science curricula provide a robust foundation in bibliographic theory, a pedagogical gap exists regarding the practical management of digital assets according to FAIR principles. To address this, the course adopted a project-based approach, structured around one possible conceptualisation of the Data Management lifecycle, for encouraging students to “build” new knowledge with the data provided. Through a 30-hour module, students from traditional humanities backgrounds alternated theoretical foundations on Digital Humanities, Data Management and Open Science with hands-on sessions using tools such as GitHub, OpenRefine, and Voyant Tools. Evaluation focused on original projects using textual 20th-century historical sources as data. Five students’ projects are then discussed as case studies. Qualitative feedback from the students highlights satisfaction with the theory-practice alternation, while revealing a need for a more unified computational ecosystem. Future developments include the curriculum's evolution toward a Python-based framework.
Keywords: Library and Information Science; Data Management; Digital Humanities Pedagogy; FAIR Principles; Project-based Learning

ABSTRACT (ITALIANO)

È tempo di costruzione: un approccio basato su progetti per insegnare la gestione dei dati agli studenti di laurea magistrale in Beni Culturali. Questo articolo presenta i risultati del corso *Digital Humanities e Data Management per i Beni Culturali*, integrato nella laurea magistrale in *Scienze del Libro e del Documento* presso l’Università di Bologna. Sebbene i curricula tradizionali nell’ambito delle scienze dell’informazione forniscono una solida base nella teoria bibliografica, esiste un divario pedagogico per quanto riguarda la gestione pratica delle risorse digitali secondo i principi FAIR. Per affrontare tale lacuna, il corso ha adottato un approccio basato sulla realizzazione di progetti, strutturato attorno al ciclo di vita del Data Management, in modo da incoraggiare gli studenti a “costruire” nuova conoscenza con i dati disponibili. Attraverso un modulo di 30 ore, gli studenti hanno alternato basi teoriche su Digital Humanities, Data Management e Open Science a sessioni pratiche utilizzando strumenti come GitHub, OpenRefine e Voyant Tools. La valutazione si è concentrata su progetti originali, presentati qui come casi di studio, che hanno utilizzato fonti storiche testuali del XX secolo come dati. Il feedback qualitativo degli studenti evidenzia la soddisfazione per l'alternanza tra teoria e pratica, rivelando al contempo la necessità di un ecosistema computazionale più unificato. Gli sviluppi futuri prevedono l'evoluzione del curriculum verso un framework basato su Python.
Parole chiave: Scienze dell’Informazione; Gestione dei dati; Educazione in Digital Humanities; Principi FAIR; Apprendimento Project-based

1. INTRODUCTION

The Master's degree in *Library and Archive Science* (*Scienze del Libro e del Documento*)[1] at the *Department of Cultural Heritage* (campus of Ravenna, University of Bologna, Italy) aims to train professionals equipped with the technical and scientific competencies required to manage paper, digital, and multimedia documents. In an increasingly complex information landscape, these figures must operate effectively across the entire documentary workflow, from the preservation of historical heritage to the management of contemporary digital assets. Thus, while the curriculum offers a robust foundation in library science, bibliography, and archives, the rapid digitisation of cultural heritage (CH) calls for practical skills to be integrated into the professional journey of CH workers.

The field of *Digital Humanities* (DH), broadly defined as the intersection of digital technologies and humanities disciplines (Poole, 2017), has introduced innovative ways of working with digital resources in the humanities domains. This knowledge can be particularly useful, given the newfound abundance of digital data (Bevan, 2015) and the ever present reproducibility crisis sweeping across academic circles (Baker, 2016). Moreover, knowing how to do proper *Data Management* (DM) and *Open Science* (OS) has become an essential part of good practice in DH (Borgman, 2010) and in the *Library and Information Science* (LIS) domain as well (Koltay, 2017).

Despite this shift, DM instruction has been historically overlooked in the humanities (Dressel, 2017), although more recently digital methods have been increasingly adopted into the curricula of some institutions (Yao & Xian, 2022). Indeed, training in DM for researchers has been encouraged, although the same cannot be said for humanities students (Wiljes & Cimiano, 2019). While students in LIS programs - including *Library and Archive Science* - often possess solid theoretical knowledge in information theory, cataloguing, and bibliographic classification, they frequently lack the practical skills necessary to manage CH data correctly to make them *Findable*, *Accessible*, *Interoperable* and *Reusable* - in other words, compliant with the *FAIR principles* (Wilkinson et al., 2016).

To address these challenges in the context of the course illustrated at the beginning, this paper presents the results of the first year of a new elective course called *Digital Humanities and Data Management for Cultural Heritage*[2]. The original contribution of this paper lies in the proposal of a "low-resource" pedagogical model, designed within a an academic context characterised by limited technological and human resources, as well as time constraints and an overall lack of consciousness towards data and their management. By focusing on practical activities related to CH data production, use and dissemination, the course encourages students to be "builders" of new, reusable knowledge resources (Warwick, 2015).

2. RELATED WORK

Over the last several decades, scholarly research in the humanities has undergone a profound transformation, becoming increasingly electronic and data-driven (Borgman, 2010a). As a consequence, the role of LIS specialists has evolved: modern information professionals find themselves performing more and more analytical tasks such as data analysis and text mining (Komova et al., 2023). Moreover, they are expected to be able to make data more accessible, reusable and well-preserved for the long term (Carlson et al., 2013).

The academic response to the need for DM education in LIS (and the humanities in a broader sense) has generally followed two distinct paths: standalone courses and one-shot workshops (Carlson et al., 2013). A significant reference is the work of Wiljes & Cimiano (2019), who propose a full-semester course at Bielefeld University designed to empower students through a competency-based, hands-on approach that leverages group discussions, individual presentations, and live software demonstrations. Similarly, Piorun et al. (2012) present a multi-institutional initiative to address the challenges of managing large datasets by developing a curriculum mapped to the National Science Foundation data management requirements and structured into flexible modules. Furthermore, initiatives like *Library Carpentry* (Baker et al., 2016), *Programming Historian* (Crymble, 2012) and guides offered by institutions and consortia, such as the

[1] https://corsi.unibo.it/2cycle/LibraryScience
[2] https://www.unibo.it/en/study/course-units-transferable-skills-moocs/course-unit-catalogue/course-unit/2024/502386

Digital Research Infrastructure for the Arts and Humanities (DARIAH) (Tóth-Czifra, 2019a), have emerged as agile and accessible gateways that provide introductory DM and software skills specifically tailored for humanities professionals.

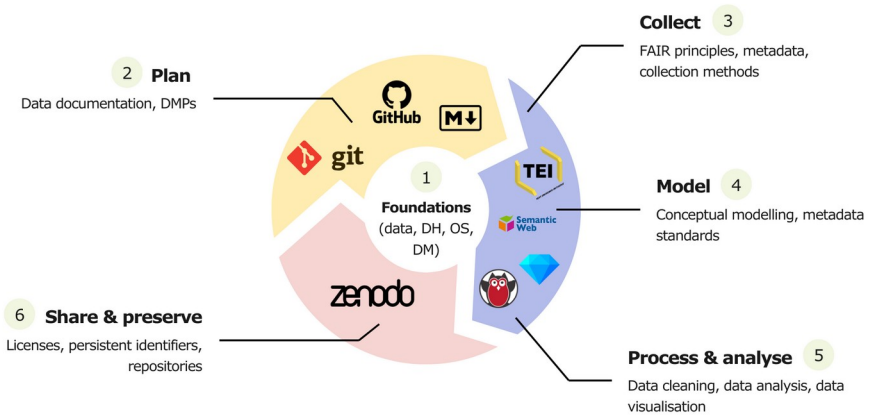


Figure 1. The course structure mapping the DM lifecycle phases to the specific theoretical concepts and software tools taught in the module

However, despite this growth in training materials for established researchers, formal courses specifically designed for undergraduate and graduate students remain rare (Smits & Teperek, 2020). This is especially common in humanities departments, due to a lack of access to technological infrastructures, financial resources, and expertise in teaching DM to those involved in humanities, amongst other things (Ruediger & MacDougall, 2023; Tóth-Czifra, 2019b).
Moreover, in the wake of the all-encompassing rise of Artificial Intelligence, the landscape of DM and DH pedagogy has been shaken up further by the emergence of Large Language Models (LLMs) (Chun & Elkins, 2023) and their use in the classroom, including data analysis (Guo, 2024) and collaborative knowledge production (Joubin, 2024). However, limitations such as hallucinations and output inaccuracy are real and need to be addressed through a careful, critical approach (Adorni et al., 2025; Wessels, 2025).
While students express willingness to learn, the lack of practical courses in humanities curricula that teach how to manage data critically and effectively, while within infrastructural limitations, remains a challenge to be addressed. This paper, hopefully, provides an example that can show how it is possible to tackle this problem and enrich the discourse around DH and DM education in the humanities.

3. METHODOLOGY

The course for the academic year 2024/2025 was designed as a 30-hour low-resource and intensive module organised around the DM lifecycle as conceptualised by University of Bologna's data stewards (Gualandi et al., 2022), and developed with a project-based, learning-by-doing approach (Georgopoulou et al., 2025) that is a staple in DH pedagogy (Thomas & Solomon, 2014; Garwood & Poole, 2019). To ensure a balanced acquisition of skills, the teaching method alternated between theoretical lectures and hands-on laboratory sessions using free open-source tools. All course materials, including slides and datasets, are shared via *Virtuale*[3] (the institutional learning platform of the University of Bologna), documented on a dedicated *GitHub* repository[4], and uploaded on *Figshare* for long-term preservation (Barzaghi, 2026). During the practical lessons, the students tackled real-world case studies based on subsets of the *Mythologiae* dataset[5], a collection of over 4,000 descriptions of artworks depicting scenes from classical mythology.
As shown in Figure 1, the lessons (15 in total) can be roughly aggregated in six thematic units:

[3] https://virtuale.unibo.it/
[4] https://github.com/dhdmch/2024-2025
[5] https://mythologiae.unibo.it/

- **Foundations.** Introduction to data as representations of entities used for knowledge creation (Andersson et al., 2024; Leonelli, 2016) and their specific meaning in the humanities domains (Gualandi et al., 2023), and to the paradigms of DH (Fitzpatrick, 2012), OS (UNESCO, 2021; Edmond & Tóth-Czifra, 2018) and DM (Briney, 2023) (Briney et al., 2020). Students are then introduced to *Git* and GitHub as tools for version control and project management (Ram, 2013; Perez-Riverol et al., 2016; Blischak et al., 2016).
- ***Plan* phase.** Students learn to document their data (Rocca-Serra et al., 2022), draft simple *Data Management Plans* (DMP) (Michener, 2015; Borer et al., 2009) and use *Markdown* to produce portable and human-readable documentation (Ovadia, 2014).
- ***Collect* phase.** Focus shifts to the FAIR principles and the importance of *metadata* as structured data documentation (Australian Research Data Commons, 2020; Ulrich et al., 2022). Students then learn to use *Zotero* (Mueen Ahmed & Dhubaib, 2011) for bibliographic data collection.
- ***Model* phase.** The students are introduced to formal *data modelling* in the humanities (Tomasi, 2018). This includes text encoding via the *Text Encoding Initiative* (TEI) XML scheme (Buzzoni, 2020; Dee, 2014) and semantic graph construction through the *Resource Description Framework* (RDF) data model (Blaney, 2017), using vocabularies like *Dublin Core* (Kunze & Baker, 2007), *Schema.org* (Guha et al., 2016) and Getty's *Art & Architecture Thesaurus* (AAT) (Soergel, 2011).
- ***Process* & *analyse* phases.** The students use *OpenRefine* (Ham, 2013) for data cleaning, following the *Tidy Data* principles (Wickham, 2014) and other guidelines for making the data more reusable (White et al., 2013). Analytical skills are developed through exploratory data analysis (Behrens, 1997), text analysis (Ittoo et al., 2016) and data visualisation (Graham, 2017; Midway, 2020) using *Voyant Tools* (Sampsel, 2018).
- ***Share & preserve* phases**. The final phase covers legal frameworks, Creative Commons licenses (Kim, 2007), and persistent identifiers, such as *Digital Object Identifiers* (DOIs) and *Open Researcher and Contributor IDs* (ORCIDs), through repositories like *Zenodo* (Sicilia et al., 2017).

The culmination of the course is an original project consisting of a series of tasks - including the creation and management of data, their manipulation, and their dissemination on the Web - which constitute some of the main activities in a DH project (Dubrovina et al., 2021) and workflow (Peroni & Tomasi, 2022).

4. RESULTS

Although students were free to select any data aligned with their specific interests, a default dataset was provided as a starting point. The dataset provided consisted of two historical travelling guides of Rome: "*Le chiese di Roma: guida storica e artistica"*, written by Diego Angeli in 1904; and "*Le chiese di Roma: guida storico-artistica"*, written by Aristide D. Tani in 1921. Both documents were provided as unstructured data in *.txt* and *.pdf* formats. From a domain perspective, these guides represent an important source of Roman travel literature: indeed, thorough history, the focus of the description in the travel guide as a literary object shifted from medieval *mirabilia* to scientific dissertations in the Renaissance, and finally to the mass-market itineraries of the 19th and early 20th centuries. For an expert in the field, these texts contain a wealth of valuable data related to topography, art history, prosopography and bibliography.
To ensure a common base of evaluation, each project was assessed against a learning outcomes framework centered around three axes, each evaluated with a score ranging from 1 to 10: ***documentation*** (proficiency in data provenance and planning, verified through the completeness of README files and a formal DMP written in Markdown); ***publication*** (proficiency in version control and data dissemination, verified through the use of GitHub and *GitHub Pages*); ***preservation*** (proficiency in OS practices, verified through the deposition of outputs on Zenodo). The three scores were then summed up to determine the final score to be assigned. Students were free to use any tool they deemed to be the best to carry out the activity (even those not covered in the course, if they wished) and follow any approach they wanted. As verifiable results, the following projects are briefly illustrated:

- **SIDE QUEST – MAIN QUEST.**[6] The project investigates the geospatial distribution of churches in Rome as described in Diego Angeli's work. Initially, data from Angeli's text was tabulated (for a total of 60 churches), with each church subjected to a geocoding process via the *Nominatim* API. Churches missing from modern maps required further manual research through *Wikipedia* and *Wikidata* to establish accurate assessments on coordinates and conservation status. For the 48 churches with valid coordinates, a *Nearest Neighbor Analysis* (NNA) was performed in *QGIS*. Results, visualised via *Kepler.gl*, indicate a distribution with a slight tendency toward aggregation, though statistically insignificant.
- **QUER PASTICCIACCIO.**[7] The project addresses the challenge of establishing a method to extract, reconcile, and encode artistic and historical information contained in Diego Angeli's work. 12 entity types were identified in the text, including architectural features, sculptural elements, and specific pictorial cycles. First, data were manually extracted and converted into a CSV file containing information such as church names, architectural agents, and internal sculptural elements. Second, the dataset underwent a cleaning process using OpenRefine, where identifiable entities were reconciled against authority files, whenever possible. Finally, the text was encoded in XML-TEI, using the cleaned CSV as a reference.
- **STRANGEWAYS, HERE WE COME.**[8] The project explores the archival and library history within Aristide D. Tani's work, focusing on data about the conservation status of documentary complexes housed within ecclesiastical institutions. Initial data exploration was conducted via Voyant Tools, using frequencies to find relevant archival mentions to be extracted and converted into a structured tabular format. These data were then processed in OpenRefine for cleaning and authority reconciliation with the *Virtual International Authority File* (VIAF). Finally, the cleaned dataset was manually modelled into RDF triples using *Dublin Core*, *Schema.org*, and the *FRBR-aligned Bibliographic Ontology* (FaBio) (Peroni & Shotton, 2012).
- **CDR-TMLA.**[9] The project explores the stylistic and structural differences between the texts of the two guides, using three specific case studies, to compare and understand the specificities of each author's style. First, the text was encoded in XML-TEI to identify entities relevant to the analysis, such as people, locations, events, and dates, reconciled whenever possible with Wikidata. Then, a quantitative analysis was conducted on the texts using Voyant Tools to extract metrics that could be used for comparison, such as vocabulary density and word frequencies. These findings were organised in a CSV file to support a comparative qualitative assessment that revealed distinct authorial strategies.
- **Roma-Sotterranea.**[10] The project explores the feasibility of mapping Rome's underground religious architectures using Diego Angeli's historical guide as a primary source. The methodological pipeline began with a manual keyword extraction to identify relevant sites, resulting in an initial CSV dataset enriched with Wikidata identifiers and accessibility status. The second phase involved XML-TEI encoding, to both organise the text into sections, paragraphs, and page breaks, and mark relevant artworks, materials, authors, and dates. A second CSV dataset was then generated specifically for artworks found in the hypogea, containing metadata such as subject IDs, dates, and provenance information.

5. DISCUSSION AND CONCLUSIONS

The first edition of the course was evaluated through two questionnaires. While the elective nature and institutional novelty of the course resulted in a small initial cohort (a total of 5 students, with 3 attending in person), thus preventing a representative quantitative analysis, the qualitative feedback gathered provides interesting information nonetheless. An initial survey revealed that, while students were driven by a strong interest in learning how to leverage digital tools for managing CH data, their baseline familiarity with DH, DM and OS was medium-low, as most entered the master's program from strictly

[6] Code: https://github.com/malemg/sqmq_project. Data: Giordano, 2025. License: CC BY-SA 4.0.utm_source=ig_web_copy_link&igsh=MzRlODBiNWFlZA==
[7] Code: https://github.com/elisabestia/qrpstcccc_project_dhdm. Data: Sabattini, 2025a; Sabattini, 2025b. Licence: CC0 1.0.
[8] Code: https://github.com/ggdrll/shwc_Tani_project. Data: Guidarelli, 2025. License: CC0 1.0.
[9] Code: https://github.com/Albadelia/cdr_tmlca_examproject_dhdm. Data: d'Elia, 2025. License: CC-BY 4.0.
[10] Code: https://github.com/glptr-gh/Roma-Sotterranea. Data: Petrosillo, 2025. License: CC0 1.0.

traditional humanities backgrounds. A mandatory post-course evaluation highlighted high student satisfaction, particularly regarding the alternation between theory and practice and the group-based laboratory sessions. The success of the students' projects serves as a promising case study, showing that the integration of this type of approach in teaching DM could significantly benefit humanities students, giving them practical skills and providing a model that can be used even in cases where technological, human and time resources are scarce.

Naturally, some limitations emerged as well. While using various open-source software tools was impactful for students who never used them before, it also highlighted the lack of a cohesive ecosystem: each tool was its own separate beast to tame, with its own interface and functionalities. Furthermore, students occasionally encountered limitations within these tools that would have required software development skills to circumvent, such as modifying certain visualisations' features produced with Voyant Tools.

Building on these shortcomings, in the second edition of the course (A.Y. 2025/2026), while maintaining the same structure, we opted for using *Python*, *Pandas*, and *Jupyter Notebooks* (freely hosted on *Google Colab*), offering students greater agency to modify and extend their DM workflows. Preliminary evidence supports this decision: the results of the mandatory post-course survey (administered to 11 attending students) showed a high satisfaction rate, despite the steeper learning curve for Python programming. Moreover, the analysis of the students' projects (shown in a growing list in the README of the course repository[11]) reveals that a more code-centric approach allows for more complex data manipulation, analysis, and management.

ACKNOWLEDGMENTS

During the preparation of this work, the authors used Grammarly and DeepL for text translation, spelling and grammar checks. Moreover, Google Gemini (Gemini 3.1 Pro) was used for paraphrasing and rewording certain sentences for improved clarity.

REFERENCES

Adorni, G., & Grosso, D. (2025, September). Generative AI in Digital Humanities Education: A Human-Centered Problem-Solving Approach. In *2025 IEEE International Conference on Cyber Humanities (IEEE-CH)* (pp. 1-6). IEEE. https://doi.org/10.1109/IEEE-CH65308.2025.11279290

Andersson, L., Huvila, I., & Sköld, O. (2024). An Introduction to Paradata. In I. Huvila, L. Andersson, & O. Sköld (Eds), *Perspectives on Paradata: Research and Practice of Documenting Process Knowledge* (pp. 1–14). Springer International Publishing. https://doi.org/10.1007/978-3-031-53946-6_1

Australian Research Data Commons (2020). *ARDC Metadata Guide*. https://doi.org/10.5281/zenodo.6459832

Baker, J., Moore, C., Priego, E., Alegre, R., Cope, J., Price, L., Stephens, O., Strien, D. van, & Wilson, G. (2016). Library Carpentry: Software skills training for library professionals. *LIBER Quarterly: The Journal of the Association of European Research Libraries*, *26*(3), 141–162. https://doi.org/10.18352/lq.10176

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, *533*(7604), 452–454. https://doi.org/10.1038/533452a

Barzaghi, S. (2026). Lecture slides for the Digital Humanities and Data Management for Cultural Heritage course (Version 1). *figshare*. https://doi.org/10.6084/m9.figshare.31168180.v1

Behrens, J. T. (1997). Principles and procedures of exploratory data analysis. *Psychological Methods*, *2*(2), 131–160. https://doi.org/10.1037/1082-989X.2.2.131

Bevan, A. (2015). The data deluge. *Antiquity*, *89*(348), 1473–1484. https://doi.org/10.15184/aqy.2015.102

Blaney, J. (2017). Introduction to the Principles of Linked Open Data. *Programming Historian*. https://programminghistorian.org/en/lessons/intro-to-linked-data

[11] https://github.com/dhdmch/2025-2026

Blischak, J. D., Davenport, E. R., & Wilson, G. (2016). A Quick Introduction to Version Control with Git and GitHub. *PLOS Computational Biology*, *12*(1), e1004668. https://doi.org/10.1371/journal.pcbi.1004668

Borer, E. T., Seabloom, E. W., Jones, M. B., & Schildhauer, M. (2009). Some Simple Guidelines for Effective Data Management. *The Bulletin of the Ecological Society of America*, *90*(2), 205–214. https://doi.org/10.1890/0012-9623-90.2.205

Borgman, C. L. (2010a). *The Digital Archive: The Data Deluge arrives in the Humanities*. https://escholarship.org/uc/item/2cc978tf

Borgman, C. L. (2010b). The Digital Future is Now: A Call to Action for the Humanities. *Digital Humanities Quarterly*, *3*(4). https://escholarship.org/uc/item/0fp9n05s

Briney, K. (2023). *The Research Data Management Workbook*. https://doi.org/10.7907/z6czh-7zx60

Briney, K. A., Coates, H., & Goben, A. (2020). Foundational Practices of Research Data Management. *Research Ideas and Outcomes*, *6*, e56508. https://doi.org/10.3897/rio.6.e56508

Buzzoni, M. (2020). Towards the construction of a stemma. In P. Roelli (Ed.), *Handbook of Stemmatology: History, Methodology, Digital Approaches* (pp. 139–207). De Gruyter. https://www.degruyterbrill.com/document/doi/10.1515/9783110684384-004/html

Carlson, J., Johnston, L., Westra, B., & Nichols, M. (2013). Developing an Approach for Data Management Education: A Report from the Data Information Literacy Project. *International Journal of Digital Curation*, *8*(1), 204–217. https://doi.org/10.2218/ijdc.v8i1.254

Chun, J., & Elkins, K. (2023). The crisis of artificial intelligence: A new digital humanities curriculum for human-centred AI. *International Journal of Humanities and Arts Computing*, *17*(2), 147-167. https://doi.org/10.3366/ijhac.2023.0310

Crymble, A. (2012, July). *The Programming Historian*. https://uhra.herts.ac.uk/id/eprint/18303/

d'Elia, A. (2025). *Cdr_tmlca_20250227_v1.0*. https://doi.org/10.5281/zenodo.14933952

D. Hirsch, B. (Ed.). (2012). *Digital Humanities Pedagogy: Practices, Principles and Politics*. Open Book Publishers. https://doi.org/10.11647/OBP.0024

Dee, S. (2014). Learning the TEI in a Digital Environment. *Journal of the Text Encoding Initiative*, *Issue 7*. https://doi.org/10.4000/jtei.968

Dressel, W. (2017). Research Data Management Instruction for Digital Humanities. *Journal of eScience Librarianship*, *6*(2), e1115. https://doi.org/10.7191/jeslib.2017.1115

Dubrovina, L., Lobuzina, K., Onyshchenko, O., & Boriak, H. (2021). Digital Humanitarian Project as a component of digital humanities. *Science and Innovation, 17(1)*, 54-63. https://doi.org/10.15407/scine17.01.054

Edmond, J., & Tóth-Czifra, E. (2018, December). *Open Data for Humanists, A Pragmatic Guide*. https://shs.hal.science/halshs-02115443

Fitzpatrick, K. (2012). The Humanities, Done Digitally. In M. K. Gold (Ed.), *Debates in the Digital Humanities* (p. 0). University of Minnesota Press. https://doi.org/10.5749/minnesota/9780816677948.003.0002

Garwood, D. A., & Poole, A. H. (2019). Pedagogy and public-funded research: An exploratory study of skills in digital humanities projects. *Journal of Documentation*, *75*(3), 550–576. https://doi.org/10.1108/JD-06-2018-0094

Georgopoulou, M. S., Troussas, C., Triperina, E., & Sgouropoulou, C. (2025). Approaches to Digital Humanities Pedagogy: A systematic literature review within educational practice. *Digital Scholarship in the Humanities*, *40*(1), 121–137. https://doi.org/10.1093/llc/fqae054

Giordano, E. M. (2025). *Dataset geospaziali delle chiese di Roma—Progetto SQMQ (v1.0)* [Data set]. *Zenodo*. https://doi.org/10.5281/zenodo.15918009

Graham, E. (2017). Introduction: Data Visualisation and the Humanities. *English Studies*, *98*(5), 449–458. https://doi.org/10.1080/0013838X.2017.1332021

Gualandi, B., Caldoni, G., & Marino, M. (2022). *Research Data Management: Data Lifecycle* [Graphic]. *Zenodo*. https://doi.org/10.5281/zenodo.7249051

Gualandi, B., Pareschi, L., & Peroni, S. (2022). What do we mean by "data"? A proposed classification of data types in the arts and humanities. *Journal of Documentation*, *79*(7), 51–71. https://doi.org/10.1108/JD-07-2022-0146

Guha, R. V., Brickley, D., & Macbeth, S. (2016). Schema.org: Evolution of structured data on the web. *Commun. ACM*, *59*(2), 44–51. https://doi.org/10.1145/2844544

Guidarelli, G. (2025). *Strangeways, here we come—Vicende conservative di archivi e biblioteche nelle chiese di Roma* [Data set]. *Zenodo*. https://doi.org/10.5281/zenodo.14934067

Guo, Q. (2024). Prompting change: ChatGPT's impact on digital humanities pedagogy–a case study in art history. *International Journal of Humanities and Arts Computing*, *18*(1), 58-78. https://doi.org/10.3366/ijhac.2024.0321

Ham, K. (2013). OpenRefine (version 2.5). Http://openrefine.org. Free, open-source tool for cleaning and transforming data. *Journal of the Medical Library Association : JMLA*, *101*(3), 233–234. https://doi.org/10.3163/1536-5050.101.3.020

Ittoo, A., Nguyen, L. M., & van den Bosch, A. (2016). Text analytics in industry: Challenges, desiderata and trends. *Computers in Industry*, *78*, 96–107. https://doi.org/10.1016/j.compind.2015.12.001

Joubin, A. A. (2024). Enhancing the trustworthiness of generative artificial intelligence in responsive pedagogy in the context of humanities higher education. In *General Aspects of Applying Generative AI in Higher Education: Opportunities and Challenges* (pp. 207-220). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-65691-0_11

Kim, M. (2007). The Creative Commons and Copyright Protection in the Digital Era: Uses of Creative Commons Licenses. *Journal of Computer-Mediated Communication*, *13*(1), 187–209. https://doi.org/10.1111/j.1083-6101.2007.00392.x

Koltay, T. (2017). Data literacy for researchers and data librarians. *Journal of Librarianship and Information Science*, *49*(1), 3–14. https://doi.org/10.1177/0961000615616450

Komova, M., Petrushka, A., Penchuk, I., & Tarasov, D. (2023). *Digital Humanities in the Concept of Training Specialists in Information, Library and Archival Science*.

Kunze, J. A., & Baker, T. (2007). *The Dublin Core Metadata Element Set* (Request for Comments No. RFC 5013). Internet Engineering Task Force. https://doi.org/10.17487/RFC5013

Leonelli, S. (2016). *Data-Centric Biology*. University of Chicago Press.

Michener, W. K. (2015). Ten Simple Rules for Creating a Good Data Management Plan. *PLOS Computational Biology*, *11*(10), e1004525. https://doi.org/10.1371/journal.pcbi.1004525

Midway, S. R. (2020). Principles of Effective Data Visualization. *Patterns*, *1*(9). https://doi.org/10.1016/j.patter.2020.100141

Mueen Ahmed, K. K., & Dhubaib, B. E. A. (2011). Zotero: A bibliographic assistant to researcher. *Journal of Pharmacology and Pharmacotherapeutics*, *2*(4), 304–305. https://doi.org/10.4103/0976-500X.85940

Ovadia, S. (2014). Markdown for Librarians and Academics. *Behavioral & Social Sciences Librarian*, *33*(2), 120–124. https://doi.org/10.1080/01639269.2014.904696

Perez-Riverol, Y., Gatto, L., Wang, R., Sachsenberg, T., Uszkoreit, J., Leprevost, F. da V., Fufezan, C., Ternent, T., Eglen, S. J., Katz, D. S., Pollard, T. J., Konovalov, A., Flight, R. M., Blin, K., & Vizcaíno, J. A. (2016). Ten Simple Rules for Taking Advantage of Git and GitHub. *PLOS Computational Biology*, *12*(7), e1004947. https://doi.org/10.1371/journal.pcbi.1004947

Peroni, S., & Shotton, D. (2012). FaBiO and CiTO: Ontologies for describing bibliographic resources and citations. *Journal of Web Semantics*, *17*, 33-43. https://doi.org/10.1016/j.websem.2012.08.001

Peroni, S., & Tomasi, F. (2022). Approaching Digital Humanities at the University: A Cultural Challenge. *Journal of Art Historiography*, December 2022(27S), 27s/SPFT1. https://arthistoriography.files.wordpress.com/2022/11/03_peroni.pdf

Petrosillo, G. (2025). Roma_Sotterranea (1.0.1) [Data set]. *Zenodo*. https://doi.org/10.5281/zenodo.17553908

Piorun, M. E., Kafel, D., Leger-Hornby, T., Najafi, S., Martin, E. R., Colombo, P., & LaPelle, N. R. (2012). Teaching Research Data Management: An Undergraduate/Graduate Curriculum. *Journal of eScience Librarianship*, *1*(1). https://doi.org/10.7191/jeslib.2012.1003

Poole, A. H. (2017). The conceptual ecology of digital humanities. *Journal of Documentation*, *73*(1), 91–122. https://doi.org/10.1108/JD-05-2016-0065

Ram, K. (2013). Git can facilitate greater reproducibility and increased transparency in science. *Source Code for Biology and Medicine*, *8*(1), 7. https://doi.org/10.1186/1751-0473-8-7

Rocca-Serra, P., Sansone, S.-A., Gu, W., Welter, D., Abbassi Daloii, T., & Portell-Silva, L. (2022). *D2.1 FAIR Cookbook*. https://doi.org/10.5281/zenodo.6783564

Ruediger, D., & MacDougall, R. (2023). *Are the Humanities Ready for Data Sharing?*. ITHAKA S+ R. https://sr.ithaka.org/wp-content/uploads/2023/03/SR-Issue-Brief-Data-Sharing-in-the-Humanities-03062023.pdf

Sampsel, L. J. (2018). Voyant Tools. *Music Reference Services Quarterly*, *21*(3), 153–157. https://doi.org/10.1080/10588167.2018.1496754

Sabattini, E. (2025a). dataset_Angeli_DHDM_refined [Data set]. *Zenodo*. https://doi.org/10.5281/zenodo.15697626

Sabattini, E. (2025b). dataset_DHDM_0 [Data set]. *Zenodo*. https://doi.org/10.5281/zenodo.15697603

Sicilia, M.-A., García-Barriocanal, E., & Sánchez-Alonso, S. (2017). Community Curation in Open Dataset Repositories: Insights from Zenodo. *Procedia Computer Science*, *106*, 54–60. https://doi.org/10.1016/j.procs.2017.03.009

Smits, D. A. B., & Teperek, M. (2020). Research Data Management for Master's Students: From Awareness to Action. *Data Science Journal*, *19*(1). https://doi.org/10.5334/dsj-2020-030

Soergel, D. (1995). The Art and Architecture Thesaurus (AAT): A Critical Appraisal. *Visual Resources*. https://www.tandfonline.com/doi/abs/10.1080/01973762.1995.9658306

Thomas, L., & Solomon, D. (2014). Active Users: Project Development and Digital Humanities Pedagogy. *CEA Critic*, *76*(2), 211–220. https://muse.jhu.edu/pub/1/article/550524

Tomasi, F. (2018). Modelling in the Digital Humanities: Conceptual Data Models and Knowledge Organization in the Cultural Heritage Domain. *Historical Social Research / Historische Sozialforschung. Supplement*, *31*, 170–179. https://www.jstor.org/stable/26533637

Tóth-Czifra, E. (2019a). DARIAH Pathfinder to Data Management Best Practices in the Humanities. *DARIAH-Campus*. https://campus.dariah.eu/resources/pathfinders/dariah-pathfinder-to-data-management-best-practices-in-the-humanities

Tóth-Czifra, E. (2019b). The risk of losing thick description: Data management challenges Arts and Humanities face in the evolving FAIR data ecosystem. https://shs.hal.science/halshs-02115505v1

Ulrich, H., Kock-Schoppenhauer, A.-K., Deppenwiese, N., Gött, R., Kern, J., Lablans, M., Majeed, R. W., Stöhr, M. R., Stausberg, J., Varghese, J., Dugas, M., & Ingenerf, J. (2022). Understanding the Nature of Metadata: Systematic Review. *Journal of Medical Internet Research*, *24*(1), e25440. https://doi.org/10.2196/25440

UNESCO (2021). UNESCO Recommendation on Open Science. UNESCO Digital Library. https://doi.org/10.54677/MNMH8546

Warwick, C. (2015). Building theories or theories of building? A tension at the heart of digital humanities. A new companion to digital humanities, 538-552. http://dx.doi.org/10.1002/9781118680605.ch37

Wessels, B. (2025). Developing an AI framework for learning in higher education: a humanities perspective from English Literature. *International Journal of Educational Technology in Higher Education*, *22*(1), 65. https://doi.org/10.1186/s41239-025-00562-w

White, E. P., Baldridge, E., Brym, Z. T., Locey, K. J., McGlinn, D. J., & Supp, S. R. (2013). *Nine simple ways to make it easier to (re)use your data* (No. e7v2). PeerJ Inc. https://doi.org/10.7287/peerj.preprints.7v2

Wickham, H. (2014). Tidy Data. *Journal of Statistical Software*, *59*, 1–23. https://doi.org/10.18637/jss.v059.i10

Wiljes, C., & Cimiano, P. (2019). Teaching Research Data Management for Students. *Data Science Journal*, *18*, 38. https://doi.org/10.5334/dsj-2019-038

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR

Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18

Yao, W., & Xiao, P. (2022). What contributes to a qualified digital humanities librarian and ideal digital humanities pedagogy? An exploratory qualitative study. *The Journal of Academic Librarianship*, *48*(6), 102524. https://doi.org/10.1016/j.acalib.2022.102524

Integrazione di tecnologie per la traduzione e CLIL nella creazione di risorse didattiche per il sardo

GIANFRANCO FRONTEDDU[1], IGOR DEIANA[2]

[1]UNIVERSITÀ DEGLI STUDI DI CAGLIARI GIANFRANCO.FRONTEDDU@UNICA.IT
[2]UNIVERSITÀ DEGLI STUDI DI CAGLIARI IGOR.DEIANA@UNICA.IT[1]

ABSTRACT (ITALIANO)

Il contributo propone una riflessione metodologica sull'uso integrato delle tecnologie per la traduzione nella produzione di materiali didattici per l'insegnamento del sardo in ambito scolastico, con particolare riferimento alla metodologia CLIL (*Content and Language Integrated Learning*). A partire dal quadro normativo nazionale e regionale che promuove l'inserimento del sardo in orario curricolare (L. 482/1999; L.R. 22/2018; avvisi INSULAS e FRAILES), il lavoro affronta la problematica della scarsità di risorse didattiche aggiornate per una lingua a risorse limitate. Dopo una sintesi dei principali approcci alla traduzione automatica (basata su regole, su corpus e neurale), il contributo analizza il ruolo del post-editing come pratica di mediazione linguistica in linea con il QCER.
In questo contesto, viene discusso l'uso del sistema di traduzione automatica *rule-based* Apertium italiano–sardo, integrato con strumenti di traduzione assistita *open source*, in particolare OmegaT, per la creazione di Unità di Apprendimento CLIL. La proposta metodologica è illustrata attraverso un caso di studio sperimentato nell'ambito del Master di II livello per la Didattica della Lingua Sarda dell'Università di Cagliari, relativo a un'UdA di storia per la scuola secondaria di primo grado. L'esperienza mostra come la combinazione tra traduzione automatica e TAO consenta di aumentare significativamente velocità ed efficienza nella produzione dei materiali, favorendo il riuso, l'adattamento e la standardizzazione di risorse digitali in lingua sarda. Se adottata su larga scala, la metodologia permette inoltre la condivisione delle memorie TMX generate, utili sia alla coerenza terminologica sia allo sviluppo dell'ecosistema digitale del sardo, anche in prospettiva di addestramento di sistemi IA e motori NMT.
Parole chiave: Traduzione automatica; Traduzione assistita (TAO); Post-editing; CLIL; Didattica del sardo

ABSTRACT (ENGLISH)

Integrating translation technologies and CLIL in the creation of teaching resources for Sardinian.
This paper offers a methodological reflection on the integrated use of translation technologies in the production of teaching materials for Sardinian in schools, with particular reference to the CLIL (Content and Language Integrated Learning) methodology. Drawing on the national and regional regulatory framework promoting the inclusion of Sardinian in the curriculum (Law 482/1999; Regional Law 22/2018; INSULAS and FRAILES notices), the paper addresses the issue of the scarcity of up-to-date teaching resources for a language with limited resources. After a summary of the main approaches to machine translation (rule-based, corpus-based, and neural), the paper analyzes the role of post-editing as a linguistic mediation practice in line with the Common European Framework of Reference for Languages (CEFR). In this context, the use of the rule-based machine translation system Apertium Italian–Sardinian, integrated with open-source computer-assisted translation tools, particularly OmegaT, for the creation of CLIL Learning Units is discussed. The proposed method is illustrated through a case study developed as part of the Master's Degree in Teaching the Sardinian Language at the University of Cagliari, involving a history course for middle school students. This experience demonstrates how the combination of machine translation and TAO significantly increases the speed and efficiency of material production, promoting the reuse, adaptation, and standardization of digital resources in the Sardinian language. If adopted on a large scale, the methodology also allows for the sharing of generated TMX memories, which are useful both for terminological consistency and for the development of the Sardinian digital ecosystem, including for training AI systems and NMT engines.

[1] L'articolo è il frutto di un lavoro congiunto e di una stretta collaborazione degli autori. All'interno di una concezione ed elaborazione comune e condivisa in tutti i suoi aspetti, sono tuttavia da attribuire a Gianfranco Fronteddu i paragrafi 3, 4 e 6 e a Igor Deiana il paragrafo 2. I paragrafi 1, 5 e 7 sono da attribuire a entrambi gli autori.

Keywords: Machine translation; Computer-assisted translation (CAT); Post-editing; CLIL; Teaching Sardinian

1. INTRODUZIONE

Il contributo propone una riflessione sull'uso integrato della traduzione automatica e della traduzione assistita per la produzione di materiali didattici finalizzati all'uso veicolare della lingua sarda per lo studio di contenuti disciplinari secondo la metodologia CLIL (*Content and Language Integrated Learning*). Muovendo dal quadro normativo e dalle politiche linguistiche regionali, il lavoro analizza il potenziale didattico di strumenti open source per lingue minoritarie. In particolare, viene presentata una metodologia applicativa che combina traduzione automatica basata su regole, *post-editing* e *CAT tools*. L'approccio è illustrato attraverso un caso di studio ideato e presentato in occasione delle attività del laboratorio "Laboratorio di didattica delle lingue e delle nuove tecnologie e metodologia CLIL", tenuto dagli autori del contributo, nell'ambito del Master di II livello per la Didattica della Lingua Sarda dell'Università degli Studi di Cagliari.

2. PRODUZIONE DI MATERIALE DIDATTICO PER LA DIDATTICA DEL SARDO IN CLIL

L'inserimento dell'insegnamento del sardo in orario curriculare in ambito scolastico trae fondamento giuridico sia dalla normativa statale a tutela delle minoranze linguistiche (L. 482/1999)[2] che dalla politica linguistica regionale (L.r. 22/2018)[3](Marra 2021). Le iniziative proposte dalla Regione Autonoma della Sardegna (d'ora in avanti RAS) per il potenziamento dell'offerta formativa sono articolate nelle linee INSULAS ("INSegnamento Unico Lingue A Scuola") e FRAILES ("Fucine di Lingue sarde") che finanziano, rispettivamente, progetti per l'insegnamento e l'utilizzo veicolare del sardo in orario curricolare e laboratori didattici extracurricolari nelle scuole dell'infanzia, primaria e secondaria di primo e secondo grado (RAS 2025). Se già negli ultimi decenni le scuole dell'Isola hanno visto lo sviluppo di proposte didattiche per la lingua sarda, le iniziative più recenti hanno contribuito ad arricchire il panorama dell'offerta (Mereu, 2024). Sono un valido esempio alcune attività dei corsi INSULAS che si basano sulla metodologia CLIL, attraverso la quale l'apprendimento di una lingua straniera o minoritaria avviene in modo più efficace quando la stessa è veicolare di contenuti disciplinari; infatti, in questo modo si favorisce l'integrazione tra competenze linguistiche, metalinguistiche e conoscitive (Coyle et al., 2010). In questo quadro, l'uso della lingua sarda come lingua veicolare consente di collegare l'apprendimento formale delle strutture e del lessico con l'analisi di contenuti culturali e disciplinari, dando luogo a un percorso didattico che supera la semplice acquisizione di regole e che favorisce la costruzione di conoscenze integrate.

L'introduzione di questi percorsi rende evidente la necessità di disporre di materiali didattici per il sardo aggiornati e adeguati alla moderna didattica delle lingue, dei quali, purtroppo, al momento, vi è poca disponibilità (Micali 2023, Grosso et al. 2024). Come ricorda infatti Mereu (2024): "I materiali didattici a disposizione per l'insegnamento del sardo (e in sardo) sono ancora molto limitati e, in genere, si ricorre a dei testi già esistenti, adattati alle esigenze scolastiche". A questa criticità si aggiunge il fatto che la creazione del materiale necessario allo svolgimento delle attività richiede, oltre a specifiche competenze glottodidattiche, un notevole dispendio di tempo anche quando si trae spunto da testi autentici o ricavati da libri di testo in adozione.

Alla luce del quadro appena descritto, come sarà illustrato nei paragrafi 3 e 4, l'uso combinato della Traduzione Automatica in combinazione con la Traduzione Assistita dal Computer, insieme alle risorse offerte dalla didattica delle lingue (sillabi e referenziali), possono essere impiegati come supporti alla generazione di risorse per l'insegnamento curricolare del sardo.

[2] Stato Italiano. (1999). Legge 15 dicembre 1999, n. 482 – Norme in materia di tutela delle minoranze linguistiche storiche (G.U. n. 297, 20 dicembre 1999). Url: https://www.mim.gov.it/lingue-di-minoranza (cons. 31/12/2025)
[3] Regione Autonoma della Sardegna. (2018). Legge regionale 3 luglio 2018, n. 22 — Disciplina della politica linguistica regionale. Url: https://leggiregionali.regione.sardegna.it/legge-regionale?data=3-7-2018&numero=22 (cons. 31/12/2025)

3. TRADUZIONE AUTOMATICA PER LA LINGUA SARDA: APERTIUM

La *traduzione automatica (d'ora in avanti TA)* o in inglese *Machine Translation* (MT) è l'espressione più diffusa per indicare i sistemi informatici che traducono da una lingua a un'altra, con o senza intervento umano, precedente o successivo al processo di traduzione (Monti, 2023).
Per classificare le principali tipologie di TA si distinguono i sistemi basati su regole (*rule-based machine translation*, d'ora in poi RBMT), in cui esperti umani definiscono esplicitamente risorse linguistiche e regole di trasferimento, e i sistemi basati su corpus (*statistical machine translation*, d'ora in poi SMT), il cui funzionamento dipende dall'addestramento su grandi quantità di dati paralleli utilizzati per la costruzione di modelli di traduzione e modelli linguistici (Martín-Mor et al., 2016). La traduzione automatica neurale (*neural machine translation*, d'ora in avanti NMT) rappresenta un'evoluzione di tali approcci ed è generalmente considerata lo stato dell'arte della TA (Castilho et al., 2017). Basata su reti neurali artificiali e modelli di *deep learning*, la NMT è in grado di elaborare testi completi apprendendo rappresentazioni linguistiche astratte, analogamente ai grandi modelli linguistici (*Large Language Models*; d'ora in avanti LLM) (Shormani, 2024; Wu et al., 2016). L'impiego di architetture *transformer* ha migliorato la fluidità e la coerenza testuale delle traduzioni, grazie alla capacità di catturare *pattern* di equivalenza tra lingue attraverso vasti corpora empirici (Elhamayed & Nour, 2025). Tuttavia, nonostante i progressi dell'Intelligenza Artificiale, d'ora in poi IA, la TA non ha ancora raggiunto un livello pienamente equivalente alla traduzione umana, soprattutto su testi estesi e a livello documentale (Mattioda et al., 2023).
La distinzione tra RBMT, SMT e NMT e la conoscenza delle rispettive caratteristiche computazionali, consente di comprenderne limiti e potenzialità e di adottare strategie d'uso adeguate (Haddow et al., 2022; Lakew et al., 2018). Studi comparativi mostrano, ad esempio, che i sistemi RBMT tendono a garantire una maggiore correttezza morfosintattica, ma risultano rigidi sul piano stilistico, mentre i sistemi NMT producono *output* più fluidi ma con errori lessicali specifici (Koponen et al., 2019).
In questo quadro, il *post-editing* (*Machine Translation Post-Editing*, MTPE) assume un ruolo centrale, inteso come revisione dell'output della TA (Balling et al., 2014). Si distinguono un *post-editing leggero*, volto alla correzione degli errori che ostacolano la comprensione, e un *post-editing completo*, che mira a un livello qualitativo comparabile a quello della traduzione umana. Si ricorda che invece, a monte, il *pre-editing* interviene sul testo di partenza tramite linguaggi controllati per prevenire errori (Riediger & Galati, 2023).
L'uso integrato di TA e *post-editing* consente una significativa riduzione di tempi e costi rispetto alla traduzione umana tradizionale. Studi e report settoriali indicano risparmi economici fino al 47% e riduzioni dei tempi di consegna fino al 50–60%, con incrementi di produttività compresi tra +38% e +66% (Kapusta et al., 2021). Tali vantaggi risultano particolarmente rilevanti nel contesto delle lingue minoritarie, dove la generazione rapida di contenuti digitali rappresenta una priorità strategica e metodologica (Terribile, 2024).
Nell'ambito della lingua sarda, un sistema di riferimento è Apertium, piattaforma di traduzione automatica basata su regole di trasferimento superficiale, particolarmente adatta alla traduzione tra lingue romanze (Tyers et al., 2017; Fronteddu et al., 2017). Il sistema è costituito da un motore centrale e da risorse linguistiche codificate in XML, tra cui dizionari morfologici monolingui e bilingui e file di regole, organizzate secondo una struttura modulare che riflette le tre fasi tipiche dei traduttori indiretti: analisi, trasferimento e generazione (Tyers et al., 2017; Fronteddu et al., 2017). Attualmente, per il sardo sono disponibili le coppie italiano–sardo e catalano–sardo, in entrambe le direzioni di traduzione (Tyers et al., 2017; Fronteddu et al., 2017). Le caratteristiche tecniche di Apertium hanno reso necessaria l'adozione di una norma ortografica coerente e sufficientemente attestata, individuata nella *Limba Sarda Comuna* (LSC),[4] funzionale alla costruzione di risorse digitali e di un corpus adeguato (Fronteddu et al., 2017; Tyers et al., 2017). In termini di qualità, Apertium italiano–sardo presenta un tasso di errore pari al 9,9% su testi di carattere generale (Tyers et al., 2017).
Negli ultimi anni, l'ecosistema delle tecnologie per la traduzione automatica del sardo si è ampliato includendo servizi basati su modelli neurali, come MinT (Pombal et al., 2025), e progetti di ricerca quali LIMBA (Carta et al., 2024) e A BERT (Carta et al., 2025), orientati allo sviluppo di risorse digitali e strumenti di apprendimento basati su IA.

[4] Limba Sarda Comuna (LSC). Url: https://www.regione.sardegna.it/documenti/1_72_20060418160308.pdf (cons. 05/02/2026)

In un contesto didattico come nel caso oggetto del presente studio, Apertium italiano–sardo risulta particolarmente adatto poiché, grazie all'integrazione con CAT tools come OmegaT (vedi paragrafo 4), consente un *post-editing* più controllabile e funzionale alla produzione efficiente di materiali didattici di base.

4. TRADUZIONE ASSISTITA (*COMPUTER-ASSISTED TRANSLATION*)

L'integrazione di strumenti di TAO risulta complementare all'uso della TA. La definizione di Traduzione Assistita dal Computer (*Computer Assisted Translation*) si riferisce all'utilizzo di programmi informatici a supporto della figura del traduttore nel processo di traduzione (Alsaawi, 2025).
Un caso di riferimento è quello di OmegaT,[5] che attualmente è uno dei programmi *open-source*[6] per la TAO, o Cat Tools, predisposti per la gestione di progetti per il sardo. Secondo quanto riportato nella guida aggiornata al 2020,[7] OmegaT è un software open source e gratuito sviluppato in Java, pensato per supportare traduttori professionisti e non professionisti nella gestione di progetti di traduzione tramite memorie di traduzione, glossari e corrispondenze parziali (*fuzzy match*) tra segmenti testuali. OmegaT non traduce autonomamente, ma offre un ambiente in cui il traduttore può ri-utilizzare e organizzare risorse linguistiche per migliorare la coerenza e l'efficienza del processo traduttivo. Dal punto di vista funzionale, OmegaT supporta la gestione di memorie di traduzione in formati standard come TMX (*Translation Memory eXchange*) e XLIFF (*XML Localization Interchange File Format*), l'uso di glossari con riconoscimento delle forme flesse, ed è integrabile con *plugin* (ad esempio, per collegarsi a motori di TA quali Apertium), oltre a fornire strumenti di allineamento, facilitando l'uso combinato in ambito didattico tra TAO e MT. Per il contesto del sardo, il supporto di Unicode[8] rende OmegaT particolarmente adatto, perché consente la configurazione manuale di coppie linguistiche e codici lingua anche non standardizzati, senza problemi con l'ortografia sarda.

5. METODOLOGIA

In occasione delle attività svolte durante il "Laboratorio di didattica delle lingue e delle nuove tecnologie e metodologia CLIL" del Master di II livello in Didattica della Lingua Sarda organizzato nell' a.a 2024-2025 dal Dipartimento di Lettere, Lingue e Beni Culturali dell'Università degli Studi di Cagliari, con il co-finanziamento della RAS, è stata ideata e presentata una metodologia volta all'uso di strumenti digitali attualmente disponibili per la creazione di Unità di Apprendimento (d'ora in avanti UdA) da destinare all'insegnamento della lingua sarda e, nello specifico, al suo uso veicolare in modalità CLIL.
La metodologia in questione prevede la creazione di un'UdA, in linea con i descrittori del Quadro Comune Europeo di Riferimento per la conoscenza delle lingue, d'ora in avanti QCER, (Consiglio d'Europa 2002, 2020) e con i principi della didattica per competenze (Balboni, 2019).
Di seguito si presentano le diverse fasi in cui si articola la proposta.
Partendo dall'analisi dei bisogni linguistici degli apprendenti e del contesto di apprendimento, la prima fase prevede la definizione del livello linguistico di riferimento su cui si lavorerà, aspetto fondamentale per stabilire i criteri di progettazione del materiale didattico e per orientare lo sviluppo dell'UdA.
La seconda fase è dedicata alla selezione del testo di partenza, preferibilmente un testo autentico in lingua italiana (divulgativo, disciplinare o enciclopedico), disponibile in formato digitale o digitalizzato tramite strumenti Riconoscimento Ottico dei Caratteri (*Optical Character Recognition*, d'ora in poi OCR).
Segue l'analisi del livello linguistico (fase 3), finalizzata a collocare il testo scelto entro uno dei sei livelli di competenza del QCER. Per la determinazione del livello linguistico dell'italiano, un valido strumento di analisi della complessità testuale è MALT_IT2[9] (Forti et al., 2020) o, in alternativa, servizi di intelligenza artificiale impiegati in modo critico e guidato. Per una più precisa determinazione delle strutture morfosintattiche e del lessico di riferimento del livello individuato è, inoltre, possibile avvalersi di strumenti come sillabi e referenziali linguistici. In particolare, per l'italiano risultano fondamentali le indicazioni fornite da *Il Profilo della lingua italiana* (Spinelli, Parizzi 2010), il quale offre un quadro di riferimento per la selezione del lessico, grazie alle liste lessicali organizzate per aree semantiche e livelli linguistici. Altrettanto utili risultano essere i principali sillabi per l'insegnamento dell'italiano come L2 (Lo

[5] OmegaT. Url: https://omegat.org/ (cons. il 05/02/2026)
[6] Open source. Url: https://opensource.org/osd (cons. il 05/02/2026)
[7] Guida per l'utente OmegaT. Url: https://hackmd.io/@omegat/for_cat_beginners (cons. il 05/02/2026)
[8] Unicode. Url: https://home.unicode.org/ (cons. il 05/02/2026)
[9] MALT_IT2. Url: è disponibile: https://lol.unistrapg.it/malt/ (cons. il 06/02/2026)

Duca, 2006; Benucci, 2007), che presentano un catalogo, indicativo e non esaustivo, di forme per le diverse funzioni comunicative e delle strutture grammaticali pertinenti ai sei livelli di competenza.
Si procede poi, nella quarta fase, con un lavoro preliminare di adattamento del testo di partenza al livello linguistico stabilito nella fase 1. Si interviene sulla struttura sintattica, morfosintattica e lessicale del testo, applicando strategie di semplificazione, riscrittura controllata e facilitazione, in funzione dei destinatari (età e livello linguistico) e degli obiettivi didattici prefissati stabiliti.
In una quinta fase, il testo viene sottoposto a un processo di TA dall'italiano al sardo mediante Apertium italiano-sardo (Fronteddu et al., 2017; Tyers et al., 2017), integrato con l'impiego dalla TAO, OmegaT.
Segue, nella fase 6, il post-editing (Vedi paragrafo 3), durante il quale non ci si limita alla correzione formale dell'output della traduzione automatica, ma si applica una riflessione di natura glottodidattica e acquisizionale analoga a quella adottata nella fase di adattamento del testo italiano. In particolare, la revisione tiene conto non solo della correttezza morfosintattica, ma anche dell'adeguatezza delle scelte lessicali e strutturali rispetto al livello linguistico precedentemente stabilito. In assenza, allo stato attuale, di sillabi e referenziali per la lingua sarda,[10] risulta utile far riferimento in modo comparativo alle risorse disponibili per le principali lingue romanze affini, come il *Plan curricular del Instituto Cervantes* (Instituto Cervantes 2007) per lo spagnolo e i *Référentiels pour le français* (Beacco et al. 2004, Beacco et al. 2007, Beacco et al. 2008, Beacco et al. 2011, Riba 2016).
La fase 7, grazie all'uso di OmegaT, consente la costruzione progressiva di memorie di traduzione e glossari terminologici disciplinari, che possono essere riutilizzati in successive UdA o condivisi per altri progetti. Tale processo contribuisce non solo alla coerenza linguistica dei materiali prodotti, ma anche alla creazione di risorse digitali per il sardo, rafforzandone la presenza nello spazio educativo e rispondendo agli obiettivi di valorizzazione e normalizzazione linguistica perseguiti dalle politiche regionali.
In questa prospettiva, l'integrazione tra CLIL, TA e TAO si configura come un impianto metodologico orientato all'efficienza produttiva, in cui il *post-editing guidato* consente una significativa riduzione dei tempi di elaborazione dei materiali didattici. Qualora la metodologia venga adottata come supporto alla produzione di materiali didattici per il sardo su larga scala, le memorie di traduzione in formato TMX generate attraverso l'uso di OmegaT potranno essere ri-utilizzate e condivise in progetti successivi, favorendo una maggiore coerenza terminologica e testuale. Tali risorse potranno inoltre costituire basi dati utili per l'addestramento di sistemi di IA e di motori di NMT, contribuendo allo sviluppo progressivo dell'ecosistema digitale della lingua sarda.

6. "SA PESADA DE SU FASCISMU": UN CASO DI STUDIO

Come caso di studio utile a mostrare un esempio applicativo di questa metodologia, si propone il modello pratico condiviso con i corsisti del Master di II livello in Didattica della Lingua Sarda. Si tratta di UdA di storia, svolta in modalità CLIL, rivolta a studenti del terzo anno della scuola secondaria di primo grado. L'UdA è intitolata *Sa Pesada de su Fascismu* ("L'ascesa del Fascismo").
Coerentemente con l'impianto metodologico sopra delineato, la fase iniziale del lavoro ha previsto la definizione del livello di riferimento e la selezione dei testi funzionali alla progettazione dell'UdA. A tal fine, sono stati individuati testi espositivi di tema storico tratti dal terzo volume del manuale *Fino a noi* (Fossati et al., 2020), attualmente in adozione nella scuola secondaria di primo grado. I testi sono stati successivamente digitalizzati mediante strumenti OCR per consentirne l'elaborazione didattica e traduttiva. Il livello linguistico stabilito prendendo in considerazione l'età e il grado scolastico frequentato dai potenziali destinatari dell'UdA (12-13 anni, scuola secondaria di primo grado) è il B1 del QCER.
Lo strumento utilizzato nella terza fase per determinare il livello dei testi secondo il QCER è la piattaforma MALT_IT2 (Forti et al., 2020). I testi sono risultati essere prevalentemente di livello B1, sebbene presentino anche tratti lessicali e strutturali riconducibili al livello B2 e, in misura marginale, al livello C1. Per questo motivo si è reso necessario un lavoro di semplificazione e facilitazione che si è limitato prevalentemente a interventi di linearizzazione e chiarificazione della struttura testuale e di

[10] Per quanto ad oggi il sardo non disponga di sillabi e referenziali, analogamente a quanto si osserva per altre lingue romanze, si segnala, tuttavia come nell'ambito dei progetti PeDESS *Pedagogia e Didattica per l'Educazione linguistica nella Scuola e nella comunità Sarda* (finanziato dall'Unione europea su fondi NextGenerationEU) e PLUREs: *PLUringuismo in Rete* (finanziato nell'ambito dell'alleanza "European Digital UniverCity" - EDUC) del Dipartimento di Lettere, Lingue e Beni Culturali dell'Università degli Studi di Cagliari sia stata avviata un'attività di ricerca finalizzata proprio alla loro progressiva elaborazione di questi strumenti (Deiana et al., 2025).

semplificazione di alcune scelte lessicali che non risultavano pertinenti per il livello stabilito per l'UdA ossia il B1.
Si è poi proceduto alla creazione di un nuovo progetto di traduzione su OmegaT, nel quale il testo è stato segmentato in unità di traduzione e successivamente sottoposto a un processo di TA su Apertium ita-srd, disponibile nella piattaforma tramite *plug-in*. Come illustrato nella Tabella 1,[11] l'*output* di Apertium è di buona qualità e il processo di *post-editing* sul prodotto della TA è legato soprattutto a scelte stilistiche o a eventuali vuoti traduttologici e ambiguità, tipicamente associati a un traduttore *rule-based* di tipo superficiale qual è Apertium (vedi paragrafo 3).

Source Text	**Apertium ita-srd**	***Post-editing***
La nascita del fascismo in Italia ha avuto luogo nel clima di crisi politica, sociale ed economica che seguì la Prima guerra mondiale. Alla fine del conflitto, il Paese si trovò attraversato da profonde tensioni.	Sa nàschida de su fascismu in Itàlia at tentu logu in su clima de crisi polìtica, sotziale e econòmica chi aiat sighidu sa Prima gherra mundiale. A sa fine de sa cuntierra, su Paisu si fiat agatadu rugradu dae profundas tensiones.	**Su fascismu est nàschidu in Itàlia** in su clima de crisi polìtica, sotziale e econòmica **a pustis de** sa Prima gherra mundiale. **A s'agabbu** de sa cuntierra, **in su Paisu b'aiat àpidu abolotu mannu.**

Tabella 1. Un estratto del testo fonte e le versioni corrispondenti di Apertium e *post-editing.*

Dal progetto di traduzione è stato ricavata una memoria di traduzione in formato TMX di 184 segmenti con un glossario in formato CSV di 10 espressioni legate alla storia del fascismo (es. it. "fasci di combattimento" → srd. "*fàscios de cumbata*"). La memoria è stata poi esportata e distribuita ai corsisti del Master per la traduzione di testi finalizzati alla creazione di altre tre UdA, sviluppate su tematiche storiche e rivolte allo stesso target di discenti, con un risparmio di tempo stimato del 40%.
La memoria di traduzione, integrata dagli aggiornamenti derivati dai successivi progetti di traduzione, conta un totale di 344 segmenti, e un incremento del glossario a 23 espressioni.
Per quanto concerne la metodologia di valutazione del prodotto della TA in riferimento al *post-editing*, sono state utilizzate metriche automatiche basate sulla distanza testuale già utilizzate per la valutazione preliminare di Apertium ita-srd (Tyers et. al, 2017; Fronteddu et. al 2017). Nello specifico si è fatto ricorso al *Word Error Rate* (WER), che misura il numero minimo di operazioni (sostituzioni, inserzioni e cancellazioni) necessarie per trasformare l'*output* della traduzione automatica nel testo di riferimento, normalizzate sulla lunghezza di quest'ultimo (Lee et al., 2023) e al *Position-independent Error Rate* (PER) (Nießen et al., 2000), che valuta la sovrapposizione lessicale tra i due testi senza considerare l'ordine delle parole. Si è rivelato particolarmente utile in presenza di variazioni sintattiche; infatti una valutazione preliminare su un campione di 344 segmenti ha evidenziato un WER relativamente elevato (≈50%), dovuto principalmente a interventi di riformulazione stilistica (vedi sopra), mentre il PER risulta significativamente inferiore (≈30%), indicando una buona copertura lessicale da parte del sistema. Ciò conferma come l'*output* di Apertium costituisca una base adeguata per il *post-editing* in ambito didattico.

7. CONCLUSIONE

L'esperienza descritta mostra come l'integrazione tra CLIL, TA e TAO e competenze glottodidattiche possa costituire una strategia efficace per la produzione di risorse didattiche in sardo. Tale organizzazione consente di ottimizzare significativamente i tempi di produzione, grazie alla generazione automatica di bozze e al riuso sistematico delle risorse, concentrando l'intervento umano nella fase di revisione, fase che può avvalersi anche degli strumenti e delle risorse messe a disposizione dalla glottodidattica moderna. In questa prospettiva, le memorie di traduzione (TMX) e i glossari costruiti durante il processo risultano riutilizzabili e condivisibili in contesti successivi, favorendo coerenza terminologica e accelerando la produzione di nuovi materiali. Inoltre, tali risorse possono costituire basi dati utili per l'addestramento di sistemi di IA e NMT, contribuendo allo sviluppo progressivo dell'ecosistema digitale della lingua sarda.

[11] Nella seconda colonna relativa all'output di Apertium, alcune porzioni di testo sono evidenziate tramite sottolineatura per segnalare i segmenti oggetto di intervento. Tali elementi corrispondono alle forme riportate in grassetto nella colonna del *post-editing*, dove sono state modificate per migliorarne adeguatezza linguistica e scorrevolezza.

La metodologia qui presentata consente quindi non solo di adattare materiali esistenti, ma anche di produrre in tempi rapidi nuove risorse, contribuendo allo sviluppo dell'offerta didattica che per il sardo è ancora dotata di pochi strumenti.

Come sviluppo futuro, è auspicabile l'estensione del modello attraverso studi di caso fondati su sperimentazioni didattiche in contesti reali di classe, che consentano non solo di verificarne l'efficacia operativa, ma anche di raccogliere evidenze empiriche e *feedback* da parte dei docenti in merito all'utilità e alla funzionalità percepita della metodologia proposta.

RINGRAZIAMENTI

Ringraziamo Unica per questa opportunità. Per ragioni di trasparenza, comunichiamo di aver fatto ricorso all'uso di supporti IA (ChatGPT), per analisi dei dati e formattazione delle citazioni.
Ringraziamo il gruppo di ricerca che ha coordinato il Master di II livello in Didattica della Lingua Sarda, con particolare riconoscimento alla prof.ssa Giulia Murgia.

BIBLIOGRAFIA

Alsaawi, A. M. (2025). Impact of Computer-Aided Translation Tools on the Product of Future Translators. Journal of Language Teaching and Research, 16(5), 1738–1748. https://doi.org/10.17507/jltr.1605.31 Balboni, P. E. (2019). Le sfide di Babele. Insegnare le lingue nelle società complesse. Quarta edizione. Torino: UTET.

Balling, L. W., Carl, M., & O'Brian, S. (2014). Post-editing of machine translation: Processes and applications. Cambridge Scholars Publishing. https://books.google.com/books?hl=en&lr=&id=W3IxBwAAQBAJ&oi=fnd&pg=PR5&dq=info:5WiD9E8QKKoJ:scholar.google.com&ots=fallTwuBvj&sig=d22N0WiK49EkhdaHf2Cvqeg7ezo

Beacco, J. C., Bouquet, S., Porquier R. (2004). Un référentiel. Niveau B2 pour le francais. Parigi. Didier.

Beacco, J. C., Porquier R. (2007). Un référentiel. Niveau A1 pour le francais. Parigi. Didier.

Beacco, J. C., Lepage, S., Porquier R., Riba, P. (2008). Un référentiel. Niveau A2 pour le francais. Parigi. Didier.

Beacco, J. C., Blin, B., Houles, E., Lepage, S., Riba, P. (2011). Un référentiel. Niveau B1 pour le francais. Parigi. Didier.

Benucci, A. (2007). Sillabo di italiano per stranieri. Una proposta del Centro Linguistico dell'Università per Stranieri di Siena. Perugia, Guerra Edizioni.

Castilho, S., Moorkens, J., Gaspari, F., Calixto, I., Tinsley, J., & Way, A. (2017). Is Neural Machine Translation the New State of the Art? The Prague Bulletin of Mathematical Linguistics, 108(1), 109–120. https://doi.org/10.1515/pralin-2017-0013

Consiglio d'Europa (2002). Quadro comune europeo di riferimento per le lingue: apprendimento, insegnamento, valutazione. Firenze. La Nuova Italia-Oxford.

Consiglio d'Europa (2020). Quadro comune europeo di riferimento per le lingue: apprendimento, insegnamento, valutazione. volume complementare. Italiano LinguaDue, 12(2). DOI: https://doi.org/10.13130/2037-3597/15120

Coyle, D., Hood, P., & Marsh, D. (2010). Content and Language Integrated Learning. Cambridge University chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://assets.cambridge.org/97805211/30219/excerpt/9780521130219_excerpt.pdf

Deiana, I., Grosso, G. I., Marra, A., Murgia, G., Puddu, N. (2025). L'insegnamento della lingua sarda Strumenti e proposte didattiche. Presentazione al workshop L'insegnamento delle lingue minoritarie in contesti educativi: pratiche, problemi, prospettive. LVIII Congresso internazionale SLI. Grammaticalizzazione.11-13 settembre 2025. L'Aquila.

Elhamayed, S. A., & Nour, M. (2025). Overview of deep learning and large language models in machine translation: A special perspective on the Arabic language. Journal of Electrical Systems and Information Technology, 12(1), 27. https://doi.org/10.1186/s43067-025-00211-2

Forti, L., Grego Bolli, G., Santarelli, F., Santucci, V., & Spina, S. (2020). MALT-IT2: A New Resource to Measure Text Difficulty in Light of CEFR Levels for Italian L2 Learning. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (A c. Di), Proceedings of the Twelfth Language Resources and

Evaluation Conference (pp. 7204–7211). European Language Resources Association. https://aclanthology.org/2020.lrec-1.890/

Fossati, L., Rossi, M., & Bianchi, P. (2020). Fino a noi. Storia, società e cittadinanza. Milano: Zanichelli.

Fronteddu, G., Font, H. A. i, & Tyers, F. M. (2017). Machine translation from Catalan to Sardinian: A translation tool for a language in the process of standardisation. Linguamática, 9(2), Articolo 2. https://doi.org/10.21814/lm.9.2.255

Grosso, G. I., Marra, A., Murgia, G. (2024). Certifying a minority language: a first description of the Sardinian case. MOSAIC, 15, 105-119. DOI: 10.69117/MOSAIC.15.1.24.006

Haddow, B., Bawden, R., Miceli Barone, A. V., Helcl, J., & Birch, A. (2022). Survey of Low-Resource Machine Translation. Computational Linguistics, 48(3), 673–732. https://doi.org/10.1162/coli_a_00446

Instituto Cervantes (2007). Plan curricular del Instituto Cervantes. Niveles de referencia para el español. Madrid, Edelsa | Editorial Español lengua extranjera. https://cvc.cervantes.es/ensenanza/biblioteca_ele/plan_curricular/

Kapusta, J., Benko, Ľ., Munkova, D., & Munk, M. (2021). Analysis of Edit Operations for Post-editing Systems. International Journal of Computational Intelligence Systems, 14(1), 1. https://doi.org/10.1007/s44196-021-00048-3

Kenny, D. (2022). Machine translation for everyone: Empowering users in the age of artificial intelligence. https://doi.org/10.5281/ZENODO.6653406

Koponen, M., Salmi, L., & Nikulin, M. (2019). A product and process analysis of post-editor corrections on neural, statistical and rule-based machine translation output. Machine Translation, 33(1), 61–90. https://doi.org/10.1007/s10590-019-09228-7

Lakew, S. M., Erofeeva, A., Negri, M., Federico, M., & Turchi, M. (2018). Transfer Learning in Multilingual Neural Machine Translation with Dynamic Vocabulary. In M. Turchi, J. Niehues, & M. Frederico (A c. Di), Proceedings of the 15th International Conference on Spoken Language Translation (pp. 54–61). International Conference on Spoken Language Translation. https://doi.org/10.48550/arXiv.1811.01064

Lee, S. et al. (2023). A Survey on Evaluation Metrics for Machine Translation. Mathematics, 11(4), 1006.

Lo Duca, M. G. (2006). Sillabo di italiano L2. Roma, Carocci.

Marra, A. (2021). L'insegnamento della lingua sarda. In Luise, M, C., Vicario, F. (A c. Di.), Le lingue regionali a scuola. Competenze e certificazione didattica. UTET, Torino,197-224.

Martín-Mor, A., Sánchez-Gijón, P., & Huerta, R. P. i. (2016). Tradumàtica: Tecnologies de la traducció. Eumo Editioral.

Mattioda, M. M., Molino, A., & Cinato, L. (2023). L'intelligenza artificiale per la traduzione: Orizzonti, pratiche e percorsi formativi. mediAzioni, 39, A1–A16. https://doi.org/10.6092/issn.1974-4382/18784

Micali, I. (2023). Quale didattica per le lingue minoritarie? Approcci e modelli per un'educazione plurilingue. Italiano LinguaDue, 15/1, 730-746.

Mereu, D. (2024). Una prospettiva ecologica allo studio delle lingue di minoranza: Il sardo e le sue varietà. PRT. https://iris.unito.it/handle/2318/1988912

Monti, J. (2023). Per una didattica della traduzione automatica. mediAzioni, 39, A55–A83. https://doi.org/10.6092/issn.1974-4382/18787

Nießen, S., Och, F. J., Leusch, G., & Ney, H. (2000). An evaluation tool for machine translation: Fast evaluation for MT research. In Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000). Athens.

Pombal, J., Guerreiro, N. M., Rei, R., & Martins, A. F. T. (2025). Adding Chocolate to Mint: Mitigating Metric Interference in Machine Translation (arXiv:2503.08327). arXiv. https://doi.org/10.48550/arXiv.2503.08327

Regione Autonoma della Sardegna. (2018). Legge regionale 3 luglio 2018, n. 22 — Disciplina della politica linguistica regionale (B.U. Regione Sardegna 05/07/2018, n. 32). Testo vigente sulla tutela, promozione e diffusione delle lingue storiche, inclusa la lingua sarda. https://leggiregionali.regione.sardegna.it/legge-regionale?data=3-7-2018&numero=22

Regione Autonoma della Sardegna. (2006). Limba Sarda Comuna — Norme linguistiche di riferimento per la lingua sarda. Deliberazione adottata dalla Giunta regionale (norma ortografica ufficiale per la lingua sarda). https://www.regione.sardegna.it/documenti/1_72_20060418160308.pdf

Regione Autonoma della Sardegna. (2025), PIANO DI POLITICA LINGUISTICA 2025-2029 - RELAZIONE. https://delibere.regione.sardegna.it/delibera/116675

Riba, P. (2016). Niveau C1/C2 pour le francais. Élément pour un référentiel. Parigi. Didier.

Riediger, H., & Galati, G. (2023). La traduzione nell'era dell'IA: Nuovi ruoli, nuove competenze, nuova formazione. mediAzioni, 39, A35–A54. https://doi.org/10.6092/issn.1974-4382/18786

Shormani, M. Q. (2024). Generative linguistics contribution to artificial intelligence: Where this contribution lies? (arXiv:2410.20221). arXiv. https://doi.org/10.48550/arXiv.2410.20221

Spinelli, B., Parizzi, F. (2010). Profilo della lingua italiana. Livelli di riferimento del QCER A1, A2, B1, B2. Firenze, La Nuova Italia.

Stato Italiano. (1999). Legge 15 dicembre 1999, n. 482 – Norme in materia di tutela delle minoranze linguistiche storiche (G.U. n. 297, 20 dicembre 1999).https://www.mim.gov.it/lingue-di-minoranza

Terribile, S. (2024). Is post-editing really faster than human translation? Translation Spaces, 13(2), 171–199. https://doi.org/10.1075/ts.22044.ter

Tyers, F. M., Alòs i Font, H., Fronteddu, G., & Martín-Mor, A. (2017). Rule-Based Machine Translation for the Italian–Sardinian Language Pair. The Prague Bulletin of Mathematical Linguistics, 108(1), 221–232. https://doi.org/10.1515/pralin-2017-0022

Wang Et Al. 2022 | PDF | Machine Learning | Receiver Operating Characteristic. (s.d.). Recuperato 6 febbraio 2026, da https://it.scribd.com/document/882065912/Wang-et-al-2022

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., … Dean, J. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation (arXiv:1609.08144). arXiv. https://doi.org/10.48550/arXiv.1609.08144

[5/A] Linked Open Data, modelli semantici, infrastrutture

Reti sociali e linked data: persone e co-occorrenze documentarie nel Portale delle fonti per la storia della Repubblica italiana

HERBERT NATTA[1], GIANLUCA ROSSI[2], ROBERTA MAGGI[3]

[1] EUROPEAN RESEARCH INFRASTRUCTURE FOR HERITAGE SCIENCE E-RIHS ERIC, ITALY – HERBERT.NATTA@E-RIHS.EU
[2] FONDAZIONE LUIGI MICHELETTI, ITALY – GIANLUCA.ROSSI@GE.IMATI.CNR.IT
[5] ISTITUTO DI MATEMATICA APPLICATA E TECNOLOGIE INFORMATICHE IMATI-CNR, ITALY – MAGGI@AREA.GE.CNR.IT

ABSTRACT (ITALIANO)

Il contributo propone un'analisi esplorativa delle reti di relazioni tra le persone coinvolte nel periodo che va dalla caduta del fascismo alla fine della prima legislatura della Repubblica italiana (1943–1953), a partire dai Linked Open Data pubblicati nel Portale delle fonti per la storia della Repubblica italiana. L'ipotesi di ricerca è che le relazioni di co-occorrenza documentaria, ovvero il riferimento a persone all'interno degli stessi documenti, possano essere utilizzate come proxy significativo delle interazioni tra i principali attori che hanno contribuito alla costruzione dello Stato repubblicano.
Dal punto di vista metodologico, il lavoro si avvantaggia della disponibilità delle fonti come linked data per sviluppare l'analisi delle reti sociali (social network analysis). I dati vengono interrogati tramite SPARQL endpoint, sfruttando il modello semantico del portale per estrarre nodi (persone) e archi (relazioni indirette basate sulla condivisione di fonti), con ponderazione dei legami in base alla frequenza di co-occorrenza. Le reti risultanti vengono quindi analizzate mediante metriche di centralità e visualizzazioni grafiche, con l'obiettivo di individuare figure-chiave, cluster e comunità.
Il contributo intende dimostrare come infrastrutture digitali basate su Linked Open Data possano non solo supportare la consultazione delle fonti storiche, ma anche abilitare nuove forme di interrogazione analitica basate su metodi computazionali. In questo senso, la social network analysis non è proposta come strumento sostitutivo dell'interpretazione storica, bensì come dispositivo euristico capace di suggerire nuove domande di ricerca e di rendere visibili strutture relazionali difficilmente ricostruibili attraverso approcci tradizionali.
Parole chiave: Social Network Analysis; Linked Open Data; Web semantico; Storia italiana; Dataviz

ABSTRACT (ENGLISH)

Social networks and linked data: people and documentary co-occurrences in the Portal of Sources for Italian Republic History. The contribution proposes an exploratory analysis of the networks of relationships among the individuals involved in the period from the fall of Fascism to the end of the first legislature of the Italian Republic (1943–1953), based on the Linked Open Data published in the Portal of Sources for the History of the Italian Republic. The research hypothesis is that documentary co-occurrence relationships—that is, references to individuals within the same documents—can be used as a meaningful proxy for the interactions among the main actors who contributed to the construction of the republican state.
From a methodological perspective, the work benefits from the availability of sources as linked data in order to develop social network analysis. The data are queried through a SPARQL endpoint, leveraging the portal's semantic model to extract nodes (persons) and edges (indirect relationships based on shared sources), with ties weighted according to the frequency of co-occurrence. The resulting networks are then analyzed using centrality metrics and graphical visualizations, with the aim of identifying key figures, clusters, and communities.
The contribution aims to demonstrate how digital infrastructures based on Linked Open Data can not only support the consultation of historical sources but also enable new forms of analytical inquiry based on computational methods. In this sense, social network analysis is not proposed as a substitute for historical interpretation, but rather as a heuristic device capable of suggesting new research questions and making visible relational structures that are difficult to reconstruct through traditional approaches.

Keywords: Social Network Analysis; Linked Open Data; Semantic Web; Italian Republican History; Dataviz

1. INTRODUZIONE

La digitalizzazione del patrimonio documentario, e il conseguente sviluppo di infrastrutture digitali dedicate alla condivisione di dati e informazioni, rende disponibile una base di conoscenza utile all'analisi, ricostruzione e interpretazione delle dinamiche storiche (Belteki *et al*, 2025).
In particolare, la diffusione di archivi digitali basati su modelli semantici e Linked Open Data (LOD) consente non solo l'estensione dell'accesso alla consultazione delle fonti, ma anche specifiche modalità di interrogazione analitica dei dati (Daquino & Tomasi, 2017).
In questo contesto, il Portale delle fonti per la storia della Repubblica italiana[1] (cfr. Tardella *et al*, 2024) rappresenta un caso di studio di particolare interesse, in quanto aggrega e rende interoperabili descrizioni archivistiche eterogenee relative alla storia politico-istituzionale dell'Italia repubblicana, offrendo un accesso strutturato tramite interfacce di ricerca e SPARQL endpoint[2].
Il periodo che va dalla caduta del fascismo alla fine della prima legislatura della Repubblica italiana (1943–1953) è una fase cruciale di ricostruzione istituzionale, riorganizzazione amministrativa e ridefinizione dei rapporti politici e sociali. In questi anni si consolidano i nuovi assetti costituzionali, si ridefiniscono le strutture dello Stato e si intrecciano le azioni di una pluralità di soggetti (politici, funzionari, intellettuali, rappresentanti istituzionali) che contribuiscono alla costruzione della Repubblica. L'ampia attenzione storiografica dedicata al periodo ha messo in luce le principali dinamiche che hanno caratterizzato questa significativa fase di trasformazione istituzionale a diverse scale, dai dettagli biografici alla ricostruzione della rete di relazioni tra gli attori coinvolti.
La disponibilità di dati e documenti inediti conseguente alla pubblicazione del Portale offre nuovi elementi utili all'interpretazione di questo periodo storico, abilitando inoltre l'utilizzo di metodi e tecniche computazionali. L'esplorazione di un *corpus* documentale eterogeneo (Natta et al., 2025), composto da circa 30.000 unità e 5.000 agenti (3.585 persone e 1.627 enti), richiede infatti l'adozione di strategie di analisi finalizzate sia alla restituzione di contenuti organizzati in modo funzionale all'interpretazione storiografica sia al calcolo di metriche quantitative che arricchiscono la base di conoscenza di nuovi elementi ermeneutici.
Il presente contributo si propone quindi come caso d'uso dei dati resi disponibili dal Portale, esplorando in particolare il potenziale dell'analisi delle reti sociali (Social Network Analysis, SNA) come supporto nella ricostruzione delle strutture relazionali tra le persone coinvolte nel periodo di interesse (Padget, 1993; Wetherell 1998).
L'ipotesi di ricerca è che le relazioni di co-occorrenza, intese come la presenza congiunta di riferimenti a più agenti all'interno delle stesse unità documentarie, fondi o complessi documentari, possano costituire un proxy significativo delle interazioni politiche, istituzionali e amministrative che caratterizzano il periodo in esame (Fiscarelli, 2022; Becker et al., 2020). Pur non rappresentando relazioni dirette o intenzionali, tali connessioni permettono di individuare pattern relazionali, densità di interazione e cluster tematici o di connessioni politiche e istituzionali difficilmente osservabili attraverso metodi non computazionali.
Il contributo si propone pertanto un duplice obiettivo. Da un lato, intende valutare la solidità metodologica dell'uso della SNA su dati archivistici descrittivi, discutendone potenzialità e limiti interpretativi; dall'altro, mira a dimostrare come infrastrutture basate su Linked Open Data possano fungere da piattaforme abilitanti per analisi computazionali riproducibili, capaci di generare nuove domande di ricerca storica. In questa prospettiva, l'analisi delle reti non è concepita come strumento sostitutivo dell'interpretazione storiografica, ma come dispositivo euristico volto a rendere visibili strutture relazionali latenti e a supportare una riflessione critica sulle modalità di costruzione e utilizzo delle fonti digitali.

2. METODOLOGIA: DAI DATI ALLA RETE

Il lavoro si articola in tre fasi principali: accesso ed estrazione dei dati, costruzione della rete e analisi delle strutture relazionali.

[1] portalefontirepubblicaitaliana.cnr.it
[2] portalefontirepubblicaitaliana.cnr.it/sparql

I dati utilizzati provengono dal Portale delle fonti per la storia della Repubblica italiana, che raccoglie risorse provenienti da diversi istituti di conservazione nazionali e locali, tra cui archivi storici istituzionali, biblioteche specializzate e fondi privati, garantendo ampia copertura sia dal punto di vista cronologico che tematico e rappresentando trasversalmente lo spettro politico attivo nel periodo di interesse.
L'interrogazione del Portale tramite endpoint SPARQL consente di definire in modo esplicito e riproducibile i criteri di selezione delle entità e delle relazioni, garantendo la trasparenza del processo di estrazione.
Le entità considerate nello studio comprendono gli agenti di tipo persona (irho-p:Person), identificati in base ai metadati semantici associati alle risorse archivistiche, con le quali intrattengono relazioni di responsabilità di diverso tipo: relazione generiche (arco-lt:isRelatedAgentOf), nel caso in cui l'agente rappresenti una voce di indice senza indicazione di responsabilità, soggetto (arco-cd:isSubjectOf), quando l'agente rappresenta l'argomento o uno degli argomenti trattati nel documento, autore (arco-lt:hasAuthor) e soggetto produttore (arco-lt:hasCreator) per specifiche relazioni di responsabilità.
Le interrogazioni sono state progettate per individuare le entità di tipo persona e i livelli (unità documentarie, unità archivistiche, fondi o complessi archivistici) nei quali tali entità sono menzionate o ai quali sono collegate da relazioni di responsabilità. A partire da queste informazioni, sono state estratte relazioni indirette basate sulla co-occorrenza documentaria.
In particolare, la prima *query* è stata funzionale all'estrazione delle relazioni agente-risorsa:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX arco-lt: <https://w3id.org/arco/ontology/arco-lite/>
PREFIX arco-cd: <https://w3id.org/arco/ontology/context-description/>
PREFIX irho-p: <https://w3id.org/italia/republic-history/onto/person/>

SELECT ?agent ?agentLabel ?res ?resLabel
WHERE {
  ?agent arco-lt:isRelatedAgentOf|arco-cd:isSubjectOf|^arco-lt:hasAuthor|^arco-lt:hasCreator ?res .
  ?agent rdfs:label ?agentLabel;
        rdf:type ?agentType.
  ?res rdfs:label ?resLabel.

  FILTER (?agentType = irho-p:Person)
}
```

mentre la seconda è stata finalizzata al conteggio del numero di risorse alle quali ogni agente è collegato, utile alla successiva definizione dei pesi da attribuire ai nodi della rete:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX arco-lt: <https://w3id.org/arco/ontology/arco-lite/>
PREFIX arco-cd: <https://w3id.org/arco/ontology/context-description/>
PREFIX irho-p: <https://w3id.org/italia/republic-history/onto/person/>

SELECT ?agent (count(distinct ?res) as ?nRes)
WHERE {
  ?agent a  irho-p:Person;
                    rdfs:label ?subLabel;
                    arco-lt:isRelatedAgentOf|arco-cd:isSubjectOf|^arco-lt:hasAuthor|^arco-lt:hasCreator ?res.
} group by ?agent
order by desc(?nRes)
```

Sulla base dei dati estratti, è stato costruito un grafo non orientato dove i nodi rappresentano le persone rappresentate nel Portale e gli archi definiscono le relazioni di co-occorrenza documentaria. La scelta di una rete non orientata riflette la natura simmetrica della relazione considerata, che non implica un rapporto gerarchico o causale, ma esclusivamente una condivisione di contesto documentario.
I nodi sono stati pesati considerando il numero di risorse collegate, mentre alle relazioni è stato assegnato un peso in base alla frequenza di co-occorrenza, assumendo che un numero maggiore di occorrenze condivise possa indicare una maggiore intensità o stabilità del legame all'interno delle fonti.

La costruzione della rete è stata sviluppata tramite script Python, elaborando i dati estratti dalle query SPARQL. In particolare, i dati relativi alle relazioni agente-persona sono stati deduplicati (conservando però la frequenza di co-occorrenza) e aggiunti, come archi pesati, al grafo costruito tramite libreria *networkx*. I nodi generati automaticamente sono stati poi arricchiti con il peso relativo al numero di risorse relazionate.
La rete risultante è stata analizzata mediante un insieme selezionato di metriche, scelte per il loro valore interpretativo in ambito storico. In particolare, sono state considerate diverse misure di centralità (centralità di grado, autovettori, closeness e betweenness), per individuare gli attori maggiormente connessi e i soggetti che svolgono un ruolo di mediazione tra diversi gruppi o cluster della rete. Inoltre, l'analisi delle componenti e l'individuazione di comunità (*community detection*), effettuata attraverso il metodo Louvain (Blondel et al., 2008), al fine di mettere in evidenza aggregazioni di attori potenzialmente riconducibili a contesti istituzionali, tematici o politici condivisi.
L'uso delle metriche di rete non è inteso come strumento di validazione automatica delle ipotesi storiche, ma come supporto euristico per l'individuazione di pattern relazionali e per la formulazione di nuove domande di ricerca ed è quindi preliminare a una successiva fase di interpretazione qualitativa, in cui i risultati della SNA siano confrontati con il contesto storico e con la letteratura di riferimento.
Nel complesso, la metodologia adottata mira a valorizzare il potenziale analitico dei dati Linked Open Data archivistici, mantenendo al contempo una distanza critica rispetto ai limiti intrinseci delle fonti e dei modelli computazionali impiegati e intendendo l'analisi delle reti come pratica interpretativa a supporto della ricerca storica.

3. RISULTATI: CENTRALITA' E COMUNITA'

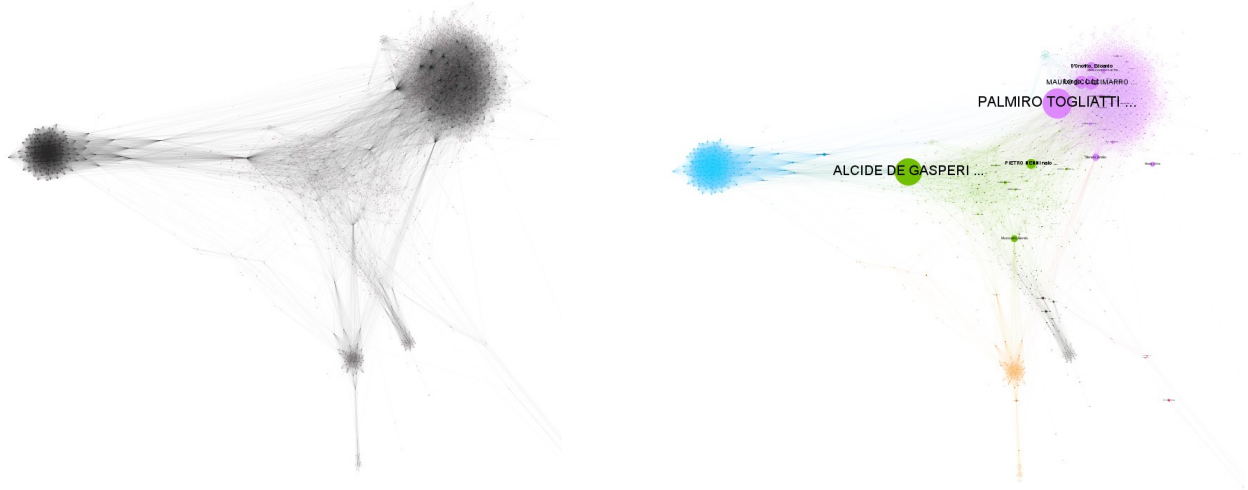


Figura 1 - Le immagini mostrano la configurazione della rete costruita a seguito dell'elaborazione dei dati estratti dal Portale (a sinistra) e il risultato della *community detection* (a destra) con nodi e archi colorati in base alla comunità di appartenenza e i nodi dimensionati in relazione ai valori di *betweenness*.

L'estrazione dei dati dal Portale ha prodotto un dataset di 2.633 entità di tipo persona che rappresentano autori, soggetti produttori, soggetti o hanno relazione generica con 2.846 risorse archivistiche, delle quali 2.807 unità, 13 fondi, 11 serie, 13 sottoserie, 1 complesso di fondi e 1 generica collezione di elementi.
Dal punto di vista qualitativo, gran parte delle relazioni tra agenti e risorse è di tipo generico (16.441), in 5.390 occorrenze l'agente è definito come soggetto del documento, in 1.856 casi ne rappresenta l'autore e in 25 il soggetto produttore.
L'analisi preliminare delle co-occorrenze documentarie ha prodotto una rete monovariata composta da 41.182 relazioni (archi) non direzionate e pesate, dove il peso assegnato a ciascun arco corrisponde al numero di documenti condivisi tra due persone, e 2.633 nodi, pesati considerando il numero complessivo di documenti nei quali compaiono.

L'analisi delle componenti della rete ha rivelato la presenza di 411 nodi *singleton*, che non risultano cioè in relazione con altri nodi: si tratta probabilmente di voci di indice isolate che identificano agenti relazionati unicamente con una risorsa e dunque non presenti in alcuna relazione di co-occorrenza.
Per procedere nell'analisi della rete, questi nodi sono stati espunti selezionando l'unica componente che contiene nodi interconnessi: un grafo composto da 2.190 nodi e 37.851 archi.
La densità della componente risultante (0.015), calcolata come rapporto tra le connessioni effettive e quelle teoricamente possibili, suggerisce una rete sparsa, nella quale molti nodi non sono direttamente connessi. Tuttavia, le metriche relative alla dimensione della rete (diametro di 8 nodi e media dei percorsi più brevi di 2,7) suggeriscono la presenza di cluster e nodi-ponte che, attivando connessioni tra cluster, riducono la distanza complessiva dei nodi sulla rete.
Comparando le due misure di centralità che evidenziano i nodi maggiormente attivi (centralità di grado e degli autovettori), si osserva una marcata polarizzazione dove un esiguo numero di agenti (33, l'1,5%) presenta un elevato numero di connessioni (>200, grado >0.1), e una lunga coda (433 nodi) con meno di 5 *link*, dei quali 100 con un solo arco (nodi foglia), con il 93% dei nodi che presentano una centralità degli autovettori (*eigenvector centrality*) inferiore a 0.13. L'unico agente che presenta elevati valori di centralità in entrambe le metriche è Alcide De Gasperi, che registra il valore massimo di *eigenvector centrality* e il sesto come centralità di grado. In termini di network analysis, la convergenza indica un soggetto che combina ampiezza relazionale (molti contatti) e qualità strutturale dei legami (connessioni con altri attori centrali). In generale, i valori confermano l'ipotesi di una struttura a cluster, nella quale numerosi nodi presentano un'elevata connettività locale, mentre solo pochi attori risultano inseriti in sottostrutture dense e strutturalmente centrali.
Le misure di centralità relative alla prossimità dei nodi (*closeness* e *betweenness)* sembrano confermare questa ipotesi: i valori di *betweenness* individuano chiaramente cinque nodi con valori maggiori (>0.05), dei quali due (Palmiro Togliatti e Alcide De Gasperi) superiori a 0.15, che rappresentano probabilmente i principali agenti con funzione di collegamento tra i cluster. La *closeness* invece presenta una distribuzione vicina alla normale (media e mediana hanno uno scarto di solo 0.02), confermando una sostanziale equidistanza dei nodi nella rete.
Spostando quindi la scala dell'osservazione dal singolo nodo e dalle sue relazioni ai sottografi definiti dalla maggiore interconnessione tra gli stessi agenti, si evidenziano 84 punti di articolazione (circa il 3,8%) suggerendo la presenza di un numero limitato ma significativo di agenti la cui rimozione comprometterebbe la connettività globale della rete. Questo risultato è compatibile con una struttura modulare, in cui sottoreti relativamente coese risultano connesse attraverso snodi specifici. In particolare, i punti di articolazione includono i cinque nodi con i valori di *betweeness* più alti (Palmiro Togliatti, Alcide De Gasperi, Mauro Scoccimarro, Luigi Longo, Pietro Nenni), confermandone il ruolo di nodi-ponte di particolare rilevanza per la struttura della rete e, in particolare, per l'interconnessione dei sottografi (Figura 1).
Infine, particolarmente rilevante per valutare l'organizzazione della rete è la presenza di sottografi completi (*cliques*), dove ogni nodo è collegato a tutti gli altri nodi. Il numero elevato di *cliques* (21.297) è di per sé poco significativo, data la presenza di numerosi nodi foglia; tuttavia, si evidenzia la presenza di *cliques* con un numero di individui particolarmente alto (>100), che fanno emergere la presenza di comunità interconnesse.
La *community detection* identifica in particolare 10 gruppi (Figura 1), dei quali due includono l'80% dei nodi: l'uno, più numeroso (55% dei nodi) e strutturalmente coeso, appare politicamente più orientato (includendo ad esempio Palmiro Togliatti, Edoardo D'Onofrio, Giancarlo Pajetta), mentre il secondo (25% dei nodi) sembra rappresentare un tessuto connettivo di figure istituzionali (tra le quali ad esempio Alcide De Gasperi, Ivanoe Bonomi, Ferruccio Parri).

4. CONCLUSIONI E SVILUPPI FUTURI

L'analisi delle reti sociali ha quindi confermato la presenza di figure chiave con funzione di *hub* relazionali, nonché aggregazioni coerenti con le istituzioni e i contesti tematici e politici del periodo di interesse, offrendo elementi di interesse per ulteriori approfondimenti storiografici.
La ricerca evidenzia inoltre limiti metodologici intrinseci: le co-occorrenze documentarie rappresentano un proxy di relazione e non garantiscono l'esistenza di interazioni dirette; la qualità e la completezza delle fonti archivistiche possono influenzare la rappresentatività della rete; la granularità temporale delle unità

archivistiche può generare distorsioni nella rilevazione dei legami. Questi limiti sottolineano l'importanza di interpretare i risultati con prudenza e di integrarli con approcci qualitativi tradizionali.
L'analisi proposta in questo contributo potrebbe essere inoltre integrata estendendone l'arco temporale, a coprire periodi successivi della storia repubblicana, al fine di osservare, in prospettiva diacronica, l'evoluzione delle reti individuate.
Inoltre, l'integrazione di nuove fonti dati, portali e archivi digitali, incoraggiata e supportata dall'utilizzo dei LOD, potrebbe inoltre aumentare l'estensione della rete, sia in termini quantitativi e di granularità, sia anche qualificando in modo più accurato alcune relazioni, permettendo così di aggiungere alle metriche (quantitative) di co-occorrenza, dati di natura qualitativa relative a relazioni effettivamente esistenti e documentate.
Inoltre, lo sviluppo di strumenti di visualizzazione e interazione che consentano di esplorare dinamicamente le reti e favorire un dialogo tra metodi quantitativi e interpretazione storica permetterebbe di sviluppare la metodologia utilizzata con funzione di storytelling digitale.
In conclusione, questo contributo dimostra come l'uso combinato di Linked Open Data e Social Network Analysis possa non solo supportare la consultazione delle fonti archivistiche, ma anche generare nuove domande di ricerca e offrire strumenti euristici per comprendere relazioni sociali e istituzionali.

RINGRAZIAMENTI

Il contributo presenta un'analisi dei dati raccolti e pubblicati nel quadro del progetto Portale delle fonti per la storia della Repubblica italiana, coordinato dal Consiglio Nazionale delle Ricerche (CNR), al quale gli autori hanno partecipato nel gruppo di lavoro dell'Istituto di Matematica Applicata e Tecnologie Informatiche (IMATI).

BIBLIOGRAFIA

Becker, S. O., Hsiao, Y., Pfaff, S., & Rubin, J. (2020). Multiplex Network Ties and the Spatial Diffusion of Radical Innovations: Martin Luther's Leadership in the Early Reformation. American Sociological Review, 85(5), 857–894. https://doi.org/10.1177/0003122420948059

Belteki, D., Rees, A. J., & Sichani, A.-M. (2025). Datafication and Cultural Heritage Collections Data Infrastructures: Critical Perspectives on Documentation, Cataloguing and Data-sharing in Cultural Heritage Institutions. Journal of Open Humanities Data, 11, 14. https://doi.org/10.5334/johd.277

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008(10), P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

Daquino, M., & Tomasi, F. (2017). Linked Cultural Objects: Dagli standard di catalogazione ai modelli per il web of data. Spunti di riflessione dalla Fototeca Zeri. Umanistica Digitale, No 1 (2017). https://doi.org/10.6092/ISSN.2532-8816/7195

Fiscarelli, A. M. (2022). Social network analysis for digital humanities. In A. Fickers & J. Tatarinov (A c. Di), Digital History and Hermeneutics (pp. 23–42). De Gruyter. https://doi.org/10.1515/9783110723991-002

Natta, H., Tardella, M., Lattanzi, E., Rossi, G., Maggi, R. (2025). Entità in relazione: policies, soluzioni tecnologiche e modelli lessicali per un (eco)sistema informativo integrato. Diversità, Equità e Inclusione: Sfide e Opportunità per l'Informatica Umanistica nell'Era dell'Intelligenza Artificiale, Proceedings del XIV Convegno Annuale AIUCD2025. Verona: AIUCD, p. 663. ISBN 978-88-942535-9-7. DOI 10.6092/unibo/amsacta/8380

Padgett, J. F., & Ansell, C. K. (1993). Robust Action and the Rise of the Medici, 1400-1434. American Journal of Sociology, 98(6), 1259–1319. https://doi.org/10.1086/230190

Salvini, A. (2005). L'analisi delle reti sociali: Risorse e meccanismi. PLUS-Pisa University Press.

Tardella, M., Maggi, R., Lodi, G., Albertoni, R., Natta, H., Rossi, G., Pasciuto, T., Ciandrini, P., Sinopoli, L., Artese, M.T., Gagliardi, I., Lattanzi, E., Ventroni, S., Tizzoni, E., Russo, A., & Porena, M. (2024). Un futuro per la memoria. Strumenti, modelli e sinergie per l'integrazione dei dati nel Portale delle fonti per la storia della Repubblica italiana. Me.Te. Digitali. Mediterraneo in rete tra testi e contesti. Proceedings del XIII Convegno Annuale AIUCD, Catania 28-30 maggio 2024, Università di Catania. Di Silvestro, & A., Spampinato D. (Eds.). ISBN 978-88-942535-8-0. DOI 10.6092/unibo/amsacta/7927

Wetherell, C. (1998). Historical Social Network Analysis. International Review of Social History, 43(S6), 125–144. https://doi.org/10.1017/S0020859000115123

Web Portals on Linked Open Data: Testing Sampo-UI in Projects on Early Modern Bibliographic Data and Chinese Calligraphy and Arts

ARIANNA MORETTI[1], KATARINA LUČIĆ[2], IIRO LASSI ILMARI TIIHONEN[3], JONAS PAUL FISCHER[4]

[1] DEPARTMENT OF CLASSICAL PHILOLOGY AND ITALIAN STUDIES, UNIVERSITY OF BOLOGNA, ITALY – ARIANNA.MORETTI4@UNIBO.IT
[2] DEPARTMENT OF INTERPRETING AND TRANSLATION, UNIVERSITY OF BOLOGNA - KATARINA.LUCIC2@UNIBO.IT
[3] DEPARTMENT OF DIGITAL HUMANITIES, UNIVERSITY OF HELSINKI - IIRO.TIIHONEN@HELSINKI.FI
[4] DEPARTMENT OF DIGITAL HUMANITIES, UNIVERSITY OF HELSINKI - JONAS.FISCHER@HELSINKI.FI

ABSTRACT (ENGLISH)

The publication of cultural heritage data as Linked Open Data increases interoperability but imposes an access barrier for non-technical users. This paper evaluates the use of semantic portals to facilitate access to RDF-serialised data, analysing the Sampo-UI framework through two case studies: (1) WenDAng, WRITE Digital Archive, for exploring data on contemporary Chinese calligraphy and arts with visual components and UI customisations, and (2) a workflow based on Early Modern bibliographic data extracted from the SPARQL service of the Bibliothèque nationale de France. The paper addresses strengths (reuse, JSON-driven configuration, independence between data management and user interface generation) and limitations (configuration complexity, dependence on iterative alignment between configurations and SPARQL queries) for small- to medium-scale datasets in the Digital Humanities.
Keywords: Semantic Web; SAMPO Framework; Web Interfaces; Bibliographic Data; Calligraphy

ABSTRACT (ITALIANO)

Portali Web Basati su Linked Open Data: Sperimentazione di Sampo-UI su Progetti Dedicati ai Dati Bibliografici della Prima Età Moderna e alla Calligrafia e alle Arti Cinesi. La pubblicazione dei dati relativi al patrimonio culturale sotto forma di Linked Open Data aumenta l'interoperabilità, ma crea una barriera di accesso per gli utenti privi di un background tecnico. Questo articolo valuta l'uso dei portali semantici per facilitare l'accesso ai dati serializzati in formato RDF, analizzando il framework Sampo-UI attraverso due casi di studio: (1) WenDAng, WRITE Digital Archive, per l'esplorazione di dati su calligrafia contemporanea cinese con componenti visuali e personalizzazioni UI, e (2) un workflow basato su dati bibliografici relativi alla prima età moderna estratti dallo SPARQL endpoint della Bibliothèque nationale de France. Il paper affronta punti di forza (riuso, configurazione JSON guidata, indipendenza tra la gestione dei dati e la generazione dell'interfaccia utente) e limiti (complessità di configurazione, dipendenza dall'allineamento iterativo tra le configurazioni e le query SPARQL) per dataset di piccola-media scala nell'ambito delle Digital Humanities.
Parole chiave: Semantic Web; SAMPO Framework; Interfacce Web; Dati Bibliografici; Calligrafia

1. INTRODUCTION

Cultural heritage organisations have been increasingly converting and publishing their data as RDF Linked Open Data (LOD) datasets, adhering to Semantic Web principles (Candela, 2025; Marden et al., 2013). While such technologies foster interoperability in data linking, they also introduce an access barrier for end users, who often lack the technical expertise needed to explore SPARQL queries or interpret RDF data. Thus, although many Cultural Heritage (CH) projects provide SPARQL endpoints, this is rarely sufficient to ensure meaningful access to information in semantic datasets. To address this challenge, web interfaces and portal solutions have been developed to make RDF data more accessible. Several frameworks currently support exploration, browsing and analysis without forcing users to engage directly with complex data models or query languages. Their development, however, raises open questions regarding usability, scalability and adaptability to different datasets and user groups.

This paper explores effective approaches to developing web interfaces for exposing RDF data, whether semantic web portals can enhance data discovery and lower the access threshold to semantic content, and to what extent the Sampo model and Sampo-UI framework represent a suitable solution for

distributing and making accessible small-to-medium triplestore datasets to specialised and general audiences.
To answer these questions, the paper builds on the Sampo model and Sampo-UI framework as a reference approach for semantic portal development. After a literature review of approaches and tools for generating user-friendly interfaces to disseminate RDF data, and an introduction to the conceptual and technical background of the Sampo framework, the paper presents two case studies in which Sampo-UI was tested to present research data. The first case study analyses the use of Sampo-UI within the WenDAng project to expose and explore data on contemporary Chinese calligraphy and related artistic practices, and evaluates the framework as a potential long-term user interface for a domain-specific semantic digital archive. The second case study applies Sampo-UI to a data processing workflow based on bibliographic data retrieved from the Bibliothèque nationale de France SPARQL endpoint, illustrating its use in refining and exploring historical datasets.
This article aims to identify both the strengths and limitations of the Sampo approach, thereby contributing to a practical, broader discussion of the role of semantic portals in reducing barriers to accessing RDF data. Guided by these goals, the paper addresses two research questions. (RQ1) What requirements and design choices emerge from the analysis of existing solutions for creating semantic portals that allow users to access and explore Linked Open Data research datasets without requiring SPARQL or knowledge of the underlying data model? (RQ2) To what extent do the Sampo model and the Sampo-UI framework respond to the need for reusable and easily configurable tools for publishing and maintaining web dissemination resources for small-to-medium RDF datasets in Digital Humanities, in light of two case studies (WenDAng and a bibliographic workflow on BnF data)?

2. LITERATURE REVIEW (ALTERNATIVES TO SAMPO)

Beyond tools for federated queries and runtime visualisations on SPARQL endpoints, recent research has focused on web portals for publishing RDF content to both expert users and wider audiences. This literature review considers alternatives for exposing RDF datasets, from exploration interfaces to reproducible portal frameworks, evaluating them on four layers: (1) service stability over time; (2) consistency with Linked Data best practices (URI management and navigation between linked entities); (3) support for generic and expert exploration modes; and (4) replicability across datasets or endpoints. Tools for exploring SPARQL endpoints and lowering the entry threshold – such as RDF Surveyor (Vega-Gorgojo et al., 2017, 2019), LODGVis (Coimbra et al., 2019), and RDF Visualizer (Petrakis et al., 2019) – support inspection of classes, instances and relationships, but are not complete publishing systems and remain mainly runtime. Other tools support dissemination through complementary functions: LODStats (Ermilov et al., 2016) provides structural statistics, MELODY (Renda et al., 2023) supports query-based storytelling, while static approaches such as Jekyll RDF (Arndt et al., 2019) and SPUK (Barzaghi, 2025) generate entity-centric HTML artefacts but not query-based interfaces or runtime perspectives. Endpoint-oriented solutions such as LD Viewer (Lukovnikov et al., 2014) and LodView (Kirillovich & Nikolaev, 2022) provide immediate, minimally configurable web views, including content negotiation, yet they do not offer multi-view analysis possibilities on semantic portals. Infrastructure-level platforms – such as RKBPlatform (Glaser & Millard, 2007), CubicWeb (Simon et al., 2013),[1] OntoWiki (Frischmuth et al., 2015), and CLEF (Daquino et al., 2023) – integrate publication with services, synchronised views, or collaborative curation rather than providing a reusable public framework. In the cultural heritage domain, ResearchSpace (Oldman & Tanase, 2018) and WissKI (Scholz & Goerz, 2012) support specialist modelling and curation, while data.open.ac.uk (Cudre-Mauroux et al., 2016) is a robust institutional reference, but it can't be adopted as a reusable toolkit for small-to-medium projects. Finally, SAMPO-UI (Hyvönen, 2012, 2020, 2023) is a framework for building reproducible semantic portals on knowledge graphs accessible via SPARQL, enabling reusable, configurable perspectives and lowering the technical barrier to disseminating semantic datasets through coordinated views such as faceted search and thematic perspectives.

3. SAMPO PORTALS FOR CULTURAL HERITAGE

Cultural heritage data are highly heterogeneous and language-specific, affecting how they are collected, modelled, and published. Digital Humanities research has focused on breaking data silos to enable a Web of Data based on Linked Data principles (Heath & Bizer, 2011; Hyvönen, 2012), where a main challenge

[1] https://www.cubicweb.org/ (cons. 08/02/2026).

remains semantic interoperability across diverse datasets. The Sampo model addresses this goal by providing guidelines for creating LOD services and semantic portals, aligned with FAIR and W3C best practices (Hyvönen, 2023; Hyvönen et al., 2023). This model is further articulated through three core ideas: (1) a "business model" for Linked Data publication, (2) multiple application perspectives for exploring data, and (3) a two-step interaction paradigm combining faceted search with analysis or visualisation of filtered results (Hyvönen, 2020).
Introduced in 2002 (Hyvönen, 2023), the Sampo model has been tested through several implementations, including portals developed by SeCo (Aalto University) and HELDIG (University of Helsinki) (Hyvönen, 2020). A Sampo portal may be implemented using Sampo-UI, a JavaScript framework providing a landing page for selecting application perspectives (typically entity classes), followed by faceted search and multiple visualisation components. Configuring a Sampo-UI portal requires an RDF triplestore accessible via a SPARQL endpoint. In the latest version, the configuration is entirely JSON-based; the backend is implemented in Node.js, and the frontend is built with React and Redux (Rantala et al., 2023; Ikkala et al., 2021). Early portals following the Sampo model include CultureSampo,[2] integrating data from 22 libraries, archives and museums and supporting citizen participation (Hyvönen et al., 2009), and BookSampo,[3] aggregating Finnish fiction literature from the mid-19th century onwards and attracting around two million users per year (Ahola et al., 2025). Other examples include WarSampo[4] and BiographySampo[5] (Hyvönen et al., 2015; Hyvönen et al., 2019). More recently, projects employing the Sampo model have been developed outside the Finnish context, demonstrating its reusability across new domains, such as Beltrans-ui, exploring the BELTRANS[6] corpus by the Royal Library of Belgium about intra-Belgian literary translations[7], and the OntoVE project (De Bastiani & Fabbris, 2025) by the Ca' Foscari University of Venice and Net7[8], about Germanic Cultural Heritage items linked to the Veneto region.

4. METHODOLOGY

Despite operational differences between WenDAng and the BnF workflow, in both projects the Sampo-UI portal configuration is the final step of a broader pipeline that must first produce a stable RDF data service, including extraction, semantic modelling, mapping to RDF, data harmonisation, semantic enrichment, publication via a SPARQL endpoint, data validation, and quality checks.
Sampo-UI runs in a Node.js environment and is configuration-driven: portal logic is defined by JSON configuration files and associated SPARQL queries (implemented as JavaScript modules), rather than through extensive custom coding. The codebase is organised into three main folders: *configs* (portal configuration, perspectives, instance pages: structure, perspectives, facets, displayed properties), *server* (backend and SPARQL directory, including queries against the RDF triplestore populating facets, result lists, instance pages aligned with perspective configs), and *client* (frontend configuration and components: localisation, reusable and project-specific JS components, CSS, UI customisation). Configuration and querying are developed iteratively, adjusting JSON and SPARQL as tests reveal mismatches of any type, such as cardinality, data types, labels, or performance. The minimal replicable unit consists of a SPARQL endpoint, versioned queries, and JSON configurations for perspectives, facets, and views. The separation between the data layer and UI supports reuse across datasets while keeping modelling choices and transformations traceable.

5. THE WENDANG AND THE EARLY MODERN BNF CASE STUDIES

The WenDAng case study examines the adoption of the Sampo-UI framework in WenDAng – WRITE Digital Archive, a project that represents and provides access to data on contemporary Chinese calligraphy and arts. The archive is affiliated with the European Research Council Starting Grant research project called *WRITE – New forms of calligraphy in China: a contemporary culture mirror* (GA n. 949645)[9],

[2] https://www.kulttuurisampo.fi/?lang=en (cons. 08/02/2026).
[3] https://analyysi.kirjasampo.fi/enComhis's/ (cons. 08/02/2026).
[4] https://www.sotasampo.fi/en/ (cons. 08/02/2026).
[5] https://seco.cs.aalto.fi/projects/biografiasampo/en/ (cons. 08/02/2026).
[6] https://www.kbr.be/en/projects/beltrans/ (cons. 22/04/2026).
[7] https://github.com/kbrbe/beltrans-ui (cons. 22/04/2026).
[8] https://www.netseven.it/ (cons. 22/04/2026).
[9] https://writecalligraphyproject.eu/ (cons. 19/04/2026).

which investigates new forms of calligraphy that have emerged in contemporary China. The WRITE project's main research question concerns whether contemporary calligraphic practices can still be understood as calligraphy in the traditional sense, and to what extent such a transformation is identifiable. WenDAng therefore serves not only as a public-facing digital archive but also as a research infrastructure supporting data-driven analysis within the project. The design and development of WenDAng largely align with the Sampo model guidelines, which can be used to verify how the archive implements Linked Open Data (LOD) principles and semantic portal design.

In WenDAng, the dataset is created and curated collaboratively with domain experts. In line with the Sampo model's emphasis on Semantic Web–based content creation (Hyvönen, 2023), it is linked to external authority and vocabulary resources. Where applicable, class instances are aligned with Wikidata entities, while some concepts are aligned with terms from Getty Art & Architecture Thesaurus (AAT)[10], Chinese Iconography Thesaurus (CIT)[11] and others. Future work includes planned alignments with the Virtual International Authority File (VIAF)[12] and the Getty Union List of Artist Names (ULAN)[13] for people, to further support interoperability.

Although the Sampo model encourages reuse of shared ontology infrastructures (Hyvönen, 2023), no existing ontology met WenDAng's needs, given the specificity of contemporary Chinese calligraphy as a domain. A dedicated ontology was therefore developed[14]; however, where possible, classes and properties from Wikidata were reused to provide contextual information, thereby ensuring partial alignment with established standards. As the Sampo model suggests, the WRITE dataset is maintained independently of the UI layer. However, as WenDAng is still a work in progress, it relies on a local Apache Jena Fuseki[15] triplestore rather than a publicly exposed SPARQL endpoint. Nevertheless, this setup reflects the conceptual separation proposed by the Sampo model. For data entry and curation, WenDAng employs Omeka S,[16] chosen for its ease of use for non-expert users (domain experts). Data is then exported from Omeka S in JSON-LD and Turtle (TTL), cleaned, harmonised, enriched and prepared to be visualised and explored in a custom Sampo-UI semantic portal. The Sampo-UI application perspectives align perfectly with the main conceptual classes of the WRITE dataset, including four WRITE collections – Visual Arts, Performing Arts, Graffiti Art, and Decorative and Applied Arts – as well as Calli-Writing Units,[17] and any other relevant classes.[18] Faceted search of the Sampo UI has proven intuitive and useful for narrowing result sets and supporting exploratory workflows. Spatial visualisations provided by the framework can help situate artists, artworks, and events geographically, while temporal views can support the exploration of stylistic and material changes over time (e.g., shifts in the use of traditional black ink on white paper using a brush). Statistical visualisations, such as pie charts, can also provide quantitative overviews, useful both for public exploration and for research-oriented analysis within the WRITE project.

WenDAng also deals with Sampo-UI customisation to improve usability and data exploration. Given the visual nature of the WRITE data, an additional component was developed for WenDAng: a browse-by-image view implemented as a masonry image gallery (see Figure 1), allowing users to search through visual material alongside standard Sampo-UI visualisations. Furthermore, to prevent information overload and enhance readability for detailed descriptions in tabular views, data are fielded into conceptual sections, for example: *identification* (title in English, title in other languages, type), *creation* (creator, date, location), *formal description* (material/support, tool, dimension, colours), and, for Calli-Writing Units, *textual/linguistic description* (writing system, character form, script style, transcription, translation). Customisation is still in progress, including structural and visual improvements, along with enhancements to the mobile version of the portal.

The choice to provide access to WenDAng data through Sampo-UI confirms its viability as a reusable framework in a domain-specific project. However, the implementation showed that reuse is not always

[10] https://www.getty.edu/research/tools/vocabularies/aat/ (cons. 19/04/2026).
[11] https://chineseiconography.org/ (cons. 19/04/2026).
[12] https://viaf.org/en (cons. 08/02/2026).
[13] https://www.getty.edu/research/tools/vocabularies/ulan/ (cons. 08/02/2026).
[14] The WRITE ontology can be consulted here: https://w3id.org/write/ontology (cons. 19/04/2026).
[15] https://jena.apache.org/documentation/fuseki2/ (cons. 08/02/2026).
[16] https://omeka.org/s/docs/user-manual/ (cons. 08/02/2026).
[17] In an artwork, a content unit featuring elements of calligraphy and/or writing, in a form of meaningful writing, set of signs and/or set of lines.
[18] For detailed description of the classes and the project's documentation, consult the official documentation page: https://wendang-project.github.io/documentation/ (cons. 19/04/2026).

straightforward. Non-technical users may find configuration challenging when project requirements extend beyond documented scenarios. In such cases, the documentation may be insufficient, requiring consultation of existing Sampo repositories as practical references.

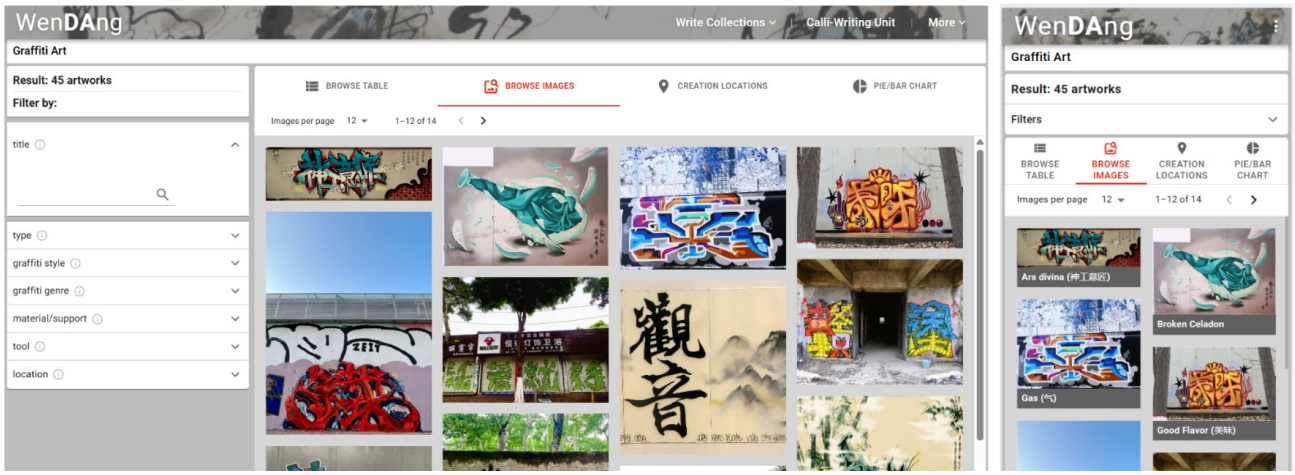


Figure 1: Masonry-style image gallery in WenDAng Sampo portal for desktop and mobile.

In the context of the BnF Early Modern Polished Dataset case study, over the last 15 years, quantitative analysis of early modern books has enabled new approaches to studying large-scale cultural, economic, and intellectual change in early modern Europe (Dittmar, 2011; Tolonen et al., 2021; Marjanen et al., 2025). Because this research requires harmonisation and enrichment of large historical datasets that are not computationally usable without preprocessing, enhancement of bibliographic information has become a topic of research in its own right. Historical catalogues such as the English Short Title Catalogue and the Finnish Fennica have therefore been subjected to "bibliographic data science" (Lahti et al., 2019a). However, many collections remain unsuitable for quantitative analysis, and limited integration between bibliographic resources constrains historical research. Many research questions, such as those concerning cultural influence, are transnational, yet most datasets have been gathered by national organisations with limited scope. Thus, studying Franco-British cultural influence in the eighteenth century (Eagles, 2000; Grieder, 1985) requires harmonising collections from both countries and identifying overlapping actors to compare cultural trends (e.g., whether an author became popular in both countries). This motivates our case study, which focuses on extracting, creating and harmonising a dataset on early modern books published in France and integrating it with British data that has already been extensively harmonised and enriched (Lahti et al., 2019b; Ijaz, 2019; Hill, 2019).

For data retrieval, BnF data was obtained through a script-based[19] workflow (Moretti et al., 2025) that mined early modern bibliographic information from the French National Library collections via its SPARQL interface. Queries were run for records published between 1454 and 1800 and for relevant metadata fields; additional queries targeted involved entities (authors, editors, translators, etc.), and the resulting biographical metadata was used to construct an actors dataset linked to bibliographic production. Operationally, SPARQL queries against the BnF endpoint were executed in batches (by year for editions and by iterating over role URIs for agents) and materialised into tabular outputs (CSV) for analysis. The dataset then underwent iterative polishing to enhance research reusability (problematic values, normalisation, deduplication, sampling-based controls, consistency checks), producing computationally usable data for comparison and potentially for mapping to external resources (e.g., ESTC), while preserving traceability through reports and intermediate outputs. Tables were rearranged according to research needs agreed with domain experts, producing curated intermediate datasets, performance metrics and quality reports. Finally, curated tables were reconverted into RDF – mapping the original BnF data model to the Cultural Heritage Acquisition and Digitisation Application Profile, CHAD-AP (Barzaghi et al., 2025) – and loaded into a triplestore (Apache Jena Fuseki), on which a custom Sampo-UI portal is configured as a web access layer. Configuration follows Sampo-UI's separation between: portal and perspective settings (JSON), perspective-specific SPARQL query files (JavaScript modules), and

[19] https://github.com/ariannamorettj/Early_Modern_BnF_Harmonisation (cons. 19/04/2026).

localisation files for labels and UI text. Portal setup was guided by established Sampo patterns. Existing GitHub instances were consulted, including the BookSampo web application repository,[20] to validate configuration choices and align perspective design with a comparable bibliographic domain. Sampo-UI worked effectively as the dissemination layer of a larger harmonisation workflow (as shown in Figure 2), keeping data engineering and user-facing exploration separate through its modularity. Portal configuration still requires technical mediation, particularly for custom views and advanced interaction scenarios, indicating strong reusability but a non-trivial implementation effort.

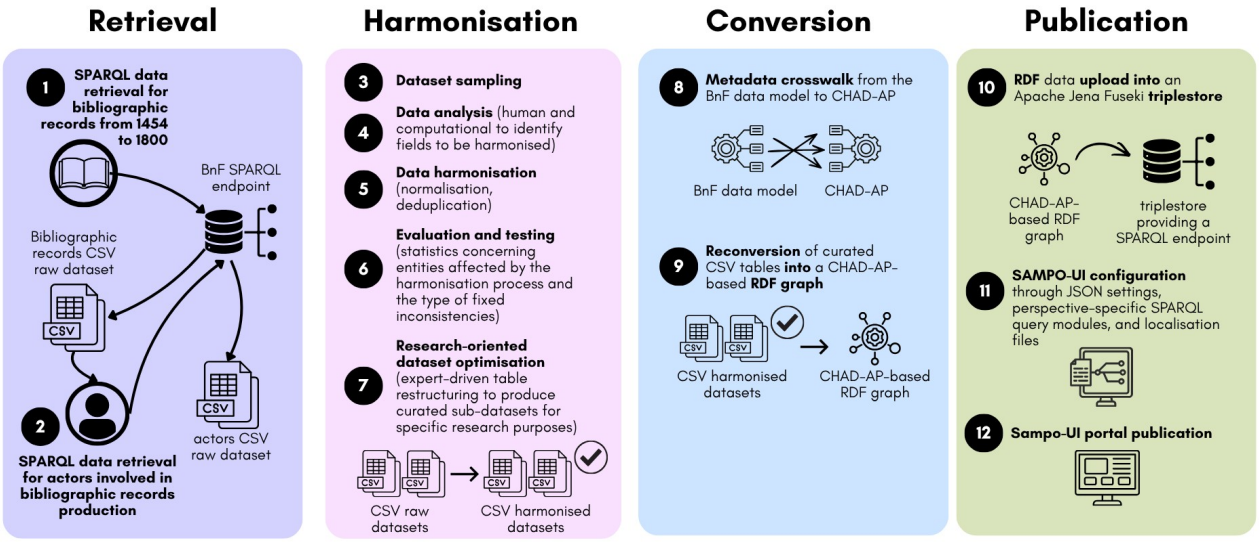


Figure 2: Data management workflow (data retrieval, harmonisation, conversion, and publication).

6. CONCLUSIONS

After a benchmark literature review on tools for web publication and exposition of RDF datasets, we selected Sampo-UI for implementation and documented two case studies following the suggested workflow: WenDAng, where Sampo-UI is assessed as a potential long-term UI layer for a domain-specific semantic archive on Chinese contemporary calligraphy and arts, and a BnF workflow where Sampo-UI is used as the dissemination step for extracted, harmonised and enriched early modern bibliographic data from the BnF SPARQL endpoint. Across both cases, Sampo proved robust for publishing and exploring RDF datasets in heterogeneous contexts. Strengths include adaptability to different information structures and a mature ecosystem of reusable references. Conversely, configuration can be an obstacle for inexperienced users when scenarios deviate from documented cases; some features require implicit knowledge or checks on other repositories. As both projects are still evolving, the conclusions concern active implementation settings rather than production-state performance. Future work will focus on stable public deployments and improved implementation documentation to support replicability and reduce adoption barriers.

In conclusion, RQ1 is addressed by demonstrating that Sampo-UI is a viable and methodologically coherent framework for LOD publication and exploration in small-to medium-sized research projects with heterogeneous semantic structures. RQ2 is addressed by identifying conditions for effective reuse: clear separation between data and UI, configuration-driven extensibility, and stronger operational documentation to bridge the gap between conceptual reusability and practical implementation effort.

ACKNOWLEDGEMENTS

According to the CEUR-WS Policy on AI-Assisting Tools, AI was ethically adopted to support Text Translation and Rephrasing (Chat-GPT-5.4 Thinking, DeepL Translate, Grammarly). The WenDAng – WRITE Digital Archive project is funded through Bando Fare (CINECA code: R20K7ZB983), financed by the Italian Ministry of University and Research (MUR). Host Institution: Department of Interpreting and Translation – University of Bologna. PI: Prof. Adriana Iezzi. The research for this article was also partially

[20] https://github.com/SemanticComputing/booksampo-web-app (cons. 08/02/2026).

funded by: (1) the Finnish Cultural Foundation; (2) Project PE 0000020 CHANGES - CUP B53C22003780006, CUP J33C22002850006, NRP Mission 4 Component 2 Investment 1.3, Funded by the European Union - NextGenerationEU; (3) the European Union's Horizon Europe Research and Innovation Programme through the Marie Sklodowska-Curie Actions Doctoral Network MECANO, Grant Agreement No. 101120349. Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

REFERENCES

Ahola, A., Peura, T., & Hyvönen, E. (2025). Using linked data for data analytic literary research: Case BookSampo—Finnish fiction literature on the semantic web. Journal of the Association for Information Science and Technology, 76(7), 937–958. https://doi.org/10.1002/asi.24984.

Arndt, N., Zänker, S., Sejdiu, G., & Tramp, S. (2019). Jekyll RDF: Template-Based Linked Data Publication with Minimized Effort and Maximum Scalability. Web Engineering: 19th International Conference, ICWE 2019, Daejeon, South Korea, June 11–14, 2019, Proceedings, 331–346. https://doi.org/10.1007/978-3-030-19274-7_24.

Barzaghi, S. (2025). Simple PUblisher of Knowledge (SPUK) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.15389756.

Barzaghi, S., Heibi, I., Moretti, A., & Peroni, S. (2025). Developing Application Profiles for Enhancing Data and Workflows in Cultural Heritage Digitisation Processes. In G. Demartini, K. Hose, M. Acosta, M. Palmonari, G. Cheng, H. Skaf-Molli, N. Ferranti, D. Hernández, & A. Hogan (Eds), The Semantic Web – ISWC 2024 (pp. 197–217). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-77847-6_11.

Candela, G. (2025). Browsing Linked Open Data in Cultural Heritage: A Shareable Visual Configuration Approach. Journal on Computing and Cultural Heritage, 18(1), 1–15. https://doi.org/10.1145/3707647.

Coimbra, D. B., Negrão, J. O. M., & Durão, F. A. (2019). LODGVis: An interactive visualization for linked open data navigation. Proceedings of the 25th Brazilian Symposium on Multimedia and the Web, 433–440. https://doi.org/10.1145/3323503.3360300.

Cudre-Mauroux, P., Daga, E., d'Aquin, M., Adamou, A., & Brown, S. (2016). The Open University Linked Data – data.open.ac.uk. Semantic Web, 7(2), 183–191. https://doi.org/10.3233/SW-150182.

Daquino, M., Wigham, M., Daga, E., Giagnolini, L., & Tomasi, F. (2023). CLEF. A Linked Open Data Native System for Crowdsourcing. J. Comput. Cult. Herit., 16(3). https://doi.org/10.1145/3594721.

De Bastiani, C., & Fabbris, G. (2025). Making Germanic Cultural Heritage accessible to students: A proposal for a case study. Diversity, Equity, and Inclusion: Challenges and Opportunities for Digital Humanities in the Age of Artificial Intelligence. Proceedings of the 14th Annual AIUCD Conference, Verona, 11–13 June 2025, 485–491.

Dittmar, J. (2011). Information Technology and Economic Change: The Impact of the Printing Press. The Quarterly Journal of Economics, 126(3), 1133–1172. https://doi.org/10.1093/qje/qjr002.

Eagles, R. (2000). Francophilia in English Society, 1748–1815. Macmillan Press and St. Martin's Press.

Ermilov, I., Lehmann, J., Martin, M., & Auer, S. (2016). LODStats: The Data Web Census Dataset. In P. Groth, E. Simperl, A. Gray, M. Sabou, M. Krötzsch, F. Lecue, F. Flöck, & Y. Gil (Eds), The Semantic Web – ISWC 2016 (pp. 38–46). Springer International Publishing. https://doi.org/10.1007/978-3-319-46547-0_5.

Frischmuth, P., Martin, M., Tramp, S., Riechert, T., & Auer, S. (2015). OntoWiki – An authoring, publication and visualization interface for the Data Web. Semantic Web, 6(3), 215–240. https://doi.org/10.3233/SW-140145.

Glaser, H., & Millard, I. C. (2007). RKB explorer: Application and infrastructure. Proceedings of the 2007 International Conference on Semantic Web Challenge - Volume 295, 97–104. https://dl.acm.org/doi/10.5555/2889885.2889898.

Grieder, J. (1985). Anglomania in France, 1740–1789: Fact, Fiction, and Political Discourse. Librairie Droz.

Heath, T., & Bizer, C. (2011). Linked Data: Evolving the Web into a global data space. Springer Cham. https://doi.org/10.1007/978-3-031-79432-2.

Hill, M. J., Vaara, V., Säily, T., Lahti, L., & Tolonen, M. (2019). Reconstructing Intellectual Networks: From the ESTC's bibliographic metadata to historical material. Digital Humanities in the Nordic and Baltic Countries Publications, 2(1), 201–219. https://doi.org/10.5617/dhnbpub.11096.

Hyvönen, E. (2012). Publishing and using cultural heritage Linked Data on the Semantic Web. Springer Cham. https://doi.org/10.1007/978-3-031-79438-4.

Hyvönen, E. (2020). Semantic "Sampo" Portals for Digital Humanities Based on a National Linked Open Data Infrastructure. https://seco.cs.aalto.fi/publications/2020/hyvonen-sampos-white-paper-2020.pdf.

Hyvönen, E. (2023). Digital humanities on the Semantic Web: Sampo model and portal series. Semantic Web, 14(4), 729–744. https://doi.org/10.3233/SW-223034.

Hyvönen, E., Leskinen, P., Tamper, M., Rantala, H., Ikkala, E., Tuominen, J., & Keravuori, K. (2019). BiographySampo—Publishing and Enriching Biographies on the Semantic Web for Digital Humanities Research. In P. Hitzler, M. Fernández, K. Janowicz, A. Zaveri, A. J. G. Gray, V. Lopez, A. Haller, & K. Hammar (Eds), The Semantic Web (pp. 574–589). Springer International Publishing. https://doi.org/10.1007/978-3-030-21348-0_37.

Hyvönen, E., Leskinen, P., & Tuominen, J. (2023). LetterSampo—Historical Letters on the Semantic Web: A Framework and Its Application to Publishing and Using Epistolary Data. ACM Journal on Computing and Cultural Heritage, 16(1), 21:1-21:23. https://doi.org/10.1145/3569372.

Hyvönen, E., Mäkelä, E., Kauppinen, T., Alm, O., Kurki, J., Ruotsalo, T., Seppälä, K., Takala, J., Puputti, K., Kuittinen, H., Viljanen, K., Tuominen, J., Palonen, T., Frosterus, M., Sinkkilä, R., Paakkarinen, P., Laitio, J., & Nyberg, K. (2009). CultureSampo: A National Publication System of Cultural Heritage on the Semantic Web 2.0. In L. Aroyo, P. Traverso, F. Ciravegna, P. Cimiano, T. Heath, E. Hyvönen, R. Mizoguchi, E. Oren, M. Sabou, & E. Simperl (Eds), The Semantic Web: Research and Applications (Vol. 5554, pp. 851–856). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-02121-3_69.

Hyvönen, E., Tuominen, J., Makela, E., Dutruit, J., Apajalahti, K., Heino, E., Leskinen, P., & Ikkala, E. (2015). Second World War on the Semantic Web: The WarSampo Project and Semantic Portal. Proceedings of the ISWC 2015 Posters & Demonstrations Track (Co-Located with the 14th International Semantic Web Conference), 1486, 4. https://urn.fi/URN:NBN:fi:aalto-201812106153.

Ijaz, A. Z., Tolonen, M., Lahti, L., & Tiihonen, I. (2019). Analytical Determination of Editions from Bibliographic Metadata. In H. Jantunen, S. Brunni, N. Kunnas, S. Palviainen, & K. Västi (Eds), Proceedings of the Research Data and Humanities (RDHUM) 2019 Conference (pp. 9–19). University of Oulu.

Ikkala, E., Hyvönen, E., Rantala, H., & Koho, M. (2021). Sampo-UI: A full stack JavaScript framework for developing semantic portal user interfaces. Semantic Web, 13(1), 69–84. https://doi.org/10.3233/SW-210428.

Kirillovich, A., & Nikolaev, K. (2022). Adapting the LodView RDF Browser for Navigation over the Multilingual Linguistic Linked Open Data Cloud. 2022 IEEE 9th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT), 143–149. https://doi.org/10.1109/SETIT54465.2022.9875628.

Lahti, L., Marjanen, J., Roivainen, H., & Tolonen, M. (2019). Bibliographic Data Science and the History of the Book (c. 1500–1800). Cataloging & Classification Quarterly, 57(1), 5–23. https://doi.org/10.1080/01639374.2018.1543747.

Lahti, L., Vaara, V., Marjanen, J., & Tolonen, M. (2019). Studying Digital Humanities with Research Data: Perspectives from Finland. In J. H. Jantunen, S. Brunni, N. Kunnas, S. Palviainen, & K. Västi (Eds), Proceedings of the Research Data and Humanities (RDHUM) 2019 Conference: Data, Methods and Tools (Vol. 17, pp. 57–65). University of Oulu. http://urn.fi/urn:isbn:9789526223216.

Lukovnikov, D., Stadler, C., & Lehmann, J. (2014). LD viewer—Linked data presentation framework. Proceedings of the 10th International Conference on Semantic Systems, 124–131. https://doi.org/10.1145/2660517.2660539.

Marden, J., Li-Madeo, C., Whysel, N., & Edelstein, J. (2013). Linked open data for cultural heritage: Evolution of an information technology. Proceedings of the 31st ACM International Conference on Design of Communication, 107–112. https://doi.org/10.1145/2507065.2507103.

Marjanen, J., Tahko, T., Lahti, L., & Tolonen, M. (2025). Book Printing in Latin and Vernacular Languages in Northern Europe, 1500–1800. In J.-M. Hanssen & S. Furuseth (Eds), The Hermeneutics of Bibliographic Data and Cultural Metadata (pp. 27–66). National Library of Norway.

Moretti, A., Tiihonen, I. L. I., & Fischer, J. P. (2025). Early Modern COMHIS BnF-Derived Dataset Production Workflow v1. https://doi.org/10.17504/protocols.io.kqdg31n5zl25/v1.

Oldman, D., & Tanase, D. (2018). Reshaping the Knowledge Graph by Connecting Researchers, Data and Practices in ResearchSpace. In D. Vrandečić, K. Bontcheva, M. C. Suárez-Figueroa, V. Presutti, I. Celino, M. Sabou, L.-A. Kaffee, & E. Simperl (Eds), The Semantic Web – ISWC 2018 (Vol. 11137, pp. 325–340). Springer International Publishing. https://doi.org/10.1007/978-3-030-00668-6_20.

Petrakis, K., Minadakis, N., Doerr, K., Theodoridou, M., & Doerr, M. (2019). RDF Visualizer: A Tool for Displaying, Browsing and Exploring High Density RDF Data. Proceedings of the ISWC 2019 Satellite Tracks (Posters & Demonstrations, Industry, and Outrageous Ideas) Co-Located with 18th International Semantic Web Conference (ISWC 2019), 2456, 197–200. https://ceur-ws.org/Vol-2456/paper51.pdf.

Rantala, H., Ahola, A., Ikkala, E., & Hyvönen, E. (2023). How to create easily a data analytic semantic portal on top of a SPARQL endpoint: Introducing the configurable Sampo-UI framework. CEUR Workshop Proceedings, 3508, 28–39. https://ceur-ws.org/Vol-3508/paper3.pdf.

Renda, G., Daquino, M., & Presutti, V. (2023). Melody: A Platform for Linked Open Data Visualisation and Curated Storytelling. Proceedings of the 34th ACM Conference on Hypertext and Social Media, 1–8. https://doi.org/10.1145/3603163.3609035.

Scholz, M., & Goerz, G. (2012). WissKI: A Virtual Research Environment for Cultural Heritage. In Frontiers in Artificial Intelligence and Applications. IOS Press. https://doi.org/10.3233/978-1-61499-098-7-1017.

Simon, A., Wenz, R., Michel, V., & Di Mascio, A. (2013). Publishing Bibliographic Records on the Web of Data: Opportunities for the BnF (French National Library). In P. Cimiano, O. Corcho, V. Presutti, L. Hollink, & S. Rudolph (Eds), The Semantic Web: Semantics and Big Data (pp. 563–577). Springer. https://doi.org/10.1007/978-3-642-38288-8_38.

Tolonen, M., Hill, M. J., Ijaz, A. Z., Vaara, V., & Lahti, L. (2021). Examining the Early Modern Canon: The English Short Title Catalogue and Large-Scale Patterns of Cultural Production. In I. Baird (Ed.), Data Visualization in Enlightenment Literature and Culture (pp. 63–119). Palgrave Macmillan. https://doi.org/10.1007/978-3-030-54913-8_3.

Vega-Gorgojo, G., Giese, M., & Slaughter, L. (2017). Exploring semantic datasets with RDF surveyor. CEUR Workshop Proceedings, 1963.

Vega-Gorgojo, G., Slaughter, L., Von Zernichow, B. M., Nikolov, N., & Roman, D. (2019). Linked Data Exploration With RDF Surveyor. IEEE Access, 7, 172199–172213. https://doi.org/10.1109/ACCESS.2019.2956345.

Risorse Terminologiche e Linked Open Data: il *Glossario delle Infrastrutture di Ricerca* (GIR) come Caso di Studio

LUCIA FRANCALANCI[1], ALESSIA SCOGNAMIGLIO[2], GIULIA PEDONESE[3], MICHELE MALLIA[4], IRENE FALINI[5], FAHAD KHAN[6]

[1] OPERA DEL VOCABOLARIO ITALIANO (CNR-OVI), ITALIA, LUCIA.FRANCALANCI@CNR.IT
[2] ISTITUTO PER LA STORIA DEL PENSIERO FILOSOFICO E SCIENTIFICO MODERNO (CNR-ISPF), ITALIA, ALESSIA.SCOGNAMIGLIO@CNR.IT
[3] ISTITUTO DI LINGUISTICA COMPUTAZIONALE "ANTONIO ZAMPOLLI" (CNR-ILC), ITALIA, GIULIA.PEDONESE@CNR.IT
[4] ISTITUTO DI LINGUISTICA COMPUTAZIONALE "ANTONIO ZAMPOLLI" (CNR-ILC), ITALIA, MICHELE.MALLIA@CNR.IT
[5] UNIVERSITÀ TELEMATICA PEGASO, ITALIA, IRENE.FALINI@UNIPEGASO.IT
[5] ISTITUTO DI LINGUISTICA COMPUTAZIONALE "ANTONIO ZAMPOLLI" (CNR-ILC), ITALIA, ANASFAHAD.KHAN@CNR.IT

ABSTRACT (ITALIANO)

Negli ultimi anni, la crescente attenzione verso l'interoperabilità dei dati, la Scienza Aperta e le Infrastrutture di Ricerca ha reso sempre più centrale la disponibilità di risorse terminologiche strutturate in grado di supportare la comunicazione scientifica, il dialogo interdisciplinare e il riuso dei prodotti della ricerca. In tale contesto, glossari e repertori terminologici non svolgono più soltanto una funzione descrittiva o divulgativa, ma si configurano come strumenti centrali per la rappresentazione della conoscenza e per l'allineamento semantico tra domini, comunità e sistemi informativi. La progettazione di tali risorse secondo il paradigma dei Linked Open Data solleva tuttavia questioni metodologiche ancora aperte, soprattutto nel caso di risorse in lingua italiana caratterizzate da elevata specializzazione e polisemia. Il contributo presenta e discute criticamente le scelte metodologiche adottate per la modellizzazione LOD del *Glossario delle Infrastrutture di Ricerca* (GIR), assumendolo come caso di studio. Il modello proposto distingue esplicitamente tra entrata lessicale, uso contestuale e concetto terminologico, articolandosi su tre livelli che integrano OntoLex-Lemon e SKOS. Particolare attenzione è dedicata al livello dell'uso, inteso come unità centrale per la rappresentazione della polisemia e dei processi di specializzazione e risemantizzazione tipici del lessico tecnico-scientifico. Si intende così offrire una riflessione metodologica replicabile per la progettazione di risorse terminologiche in ambiente Linked Open Data. Il modello proposto deriva da un processo di mediazione tra esigenze redazionali, complessità linguistica e interoperabilità semantica, che si traduce in un equilibrio negoziato tra aderenza agli standard del Web Semantico e usabilità effettiva della risorsa.
Parole chiave: Risorse terminologiche; Linked Open Data; OntoLex-Lemon; SKOS; Infrastrutture di Ricerca

ABSTRACT (ENGLISH)

Terminology Resources and Linked Open Data: the Glossary of Research Infrastructures (GIR) as a Case Study.
In recent years, the growing attention to data interoperability, Open Science, and Research Infrastructures has made the availability of structured terminology resources capable of supporting scientific communication, interdisciplinary dialogue, and the reuse of research products increasingly central. In this context, glossaries and terminology repositories serve not only a descriptive or informative function, but also as central tools for representing knowledge and aligning semantics between domains, communities, and information systems. However, designing such resources according to the Linked Open Data paradigm raises open methodological questions, particularly in the case of highly specialised, polysemous Italian
-language resources. This paper presents and critically discusses the methodological choices adopted for the LOD modelling of the *Glossary of Research Infrastructures* (GIR), using it as a case study. The proposed model distinguishes explicitly between lexical entry, contextual use, and terminological concept, and is structured into three levels that integrate OntoLex-Lemon and SKOS. Particular attention is paid to the use level, which is considered the central unit for representing polysemy and the processes of specialisation and re-semantisation typical of the technical-scientific lexicon. The aim is to offer a replicable methodological reflection for the design of terminological resources in a Linked Open Data environment. The proposed model based on a process of mediation between editorial requirements,

linguistic complexity, and semantic interoperability, which translates into a negotiated balance between adherence to Semantic Web standards and the effective usability of the resource.
Keywords: Terminology resources; Linked Open Data; OntoLex-Lemon; SKOS; Research Infrastructures

1. INTRODUZIONE

Negli ultimi anni, la crescente attenzione verso l'interoperabilità dei dati, la Scienza Aperta e le Infrastrutture di Ricerca (IR) ha reso sempre più centrale la disponibilità di risorse terminologiche strutturate, in grado di supportare la comunicazione scientifica, la cooperazione interdisciplinare e il riuso dei risultati della ricerca. In questo contesto, glossari e repertori terminologici non svolgono più soltanto una funzione descrittiva o divulgativa, ma assumono un ruolo strategico nella rappresentazione della conoscenza e nell'allineamento semantico tra sistemi, comunità e domini.
Il *Glossario delle Infrastrutture di Ricerca* (GIR) è una risorsa terminologica nata per promuovere la condivisione della conoscenza e la collaborazione tra le Infrastrutture di Ricerca in Italia, con particolare riferimento al settore delle scienze umane e sociali.[1] Sviluppato in un contesto applicativo e istituzionale, il GIR presenta specifiche caratteristiche che lo rendono un caso di studio particolarmente adatto per riflettere sui problemi di modellizzazione semantica delle risorse terminologiche: integrazione di più lessici specialistici, ampia presenza di prestiti e calchi dall'inglese, polisemia legata a una pluralità di ambiti d'uso, nonché una forte esigenza di chiarezza nelle definizioni e interoperabilità. Inoltre, fenomeni quali la polisemia, la specializzazione semantica e la variazione d'uso rendono difficile l'adozione di modelli di rappresentazione che non appiattiscano la complessità linguistica e concettuale sottostante.
La progettazione di risorse terminologiche secondo il paradigma Linked (Open) Data[2] pone problemi metodologici ancora aperti, soprattutto quando tali risorse sono costruite in lingua italiana e in contesti caratterizzati da una forte presenza di linguaggi specialistici e prestiti dall'inglese. Emerge, in particolare, una questione più generale, relativa al rapporto tra modelli *concept-oriented*, propri della tradizione terminologica, e approcci *sense-centered* caratteristici delle rappresentazioni lessicali basate su OntoLex-Lemon (McCrae *et al.*, 2017, Bosque-Gil & Gracia, 2019). La letteratura recente ha evidenziato come l'integrazione tra questi due paradigmi non costituisca un semplice problema tecnico, ma implichi un cambiamento di prospettiva nella rappresentazione del significato e nel rapporto tra termini e concetti (Bellandi *et al.*, 2023a, 2023b; Bellandi *et al.*, 2024; McCrae *et al.*, 2017).
In questo quadro, il modello proposto per il GIR si configura come una mediazione consapevole tra dimensione lessicale e dimensione terminologica: pur adottando una struttura centrata sull'uso, esso resta orientato alla rappresentazione di concetti specialistici.
Questo contributo discute criticamente le scelte metodologiche adottate per la modellizzazione del GIR come risorsa Linked Open Data (LOD). L'attenzione si concentra in particolare sulla distinzione tra forma linguistica, uso contestuale e concetto terminologico, proponendo un modello articolato su tre livelli, che combina OntoLex-Lemon e SKOS (Miles & Bechhofer, 2009).[3]
L'obiettivo che si vuole perseguire è duplice: da un lato, mostrare come un modello che attribuisce centralità all'uso consenta di rappresentare in modo esplicito la polisemia e i processi di specializzazione semantica tipici del lessico tecnico-scientifico in italiano; dall'altro, offrire una riflessione metodologica replicabile per la progettazione di risorse terminologiche in ambiente LOD, evidenziando i compromessi tra aderenza agli standard del Web Semantico e usabilità della risorsa.

2. METODOLOGIA E WORKFLOW

La costruzione del dataset LOD del GIR si basa su un workflow di trasformazione progressiva che integra strumenti di redazione, fogli di lavoro strutturati e tecnologie del Web Semantico (Fig. 1). Il processo è stato progettato per garantire, da un lato, la continuità con il lavoro redazionale originario e, dall'altro, la conformità agli standard LOD.

[1] Sul *Glossario delle Infrastrutture di Ricerca* si veda Francalanci *et al.*, 2025; Francalanci, Scognamiglio & Falini, 2025; Falini & Francalanci 2025.
[2] Sulla rappresentazione di risorse terminologiche come dati collegati cfr. Bellandi, 2022; Bellandi, 2025; Bosque-Gil *et al.*, 2015; Cimiano *et al.*, 2015; Cimiano *et al.*, 2016; Cimiano *et al.*, 2020; Diez-Ibarbia *et al.*, 2025; Martín-Chozas *et al.*,. 2025. Per applicazioni in ambito lessicografico si vedano, ad esempio, Litta *et al.*, 2024, 2025; Mambrini *et al.*, 2020; Passarotti *et al.*, 2020.
[3] Simple Knowledge Organisation System (SKOS), https://www.w3.org/2004/02/skos/ (cons. 5/2/2026).

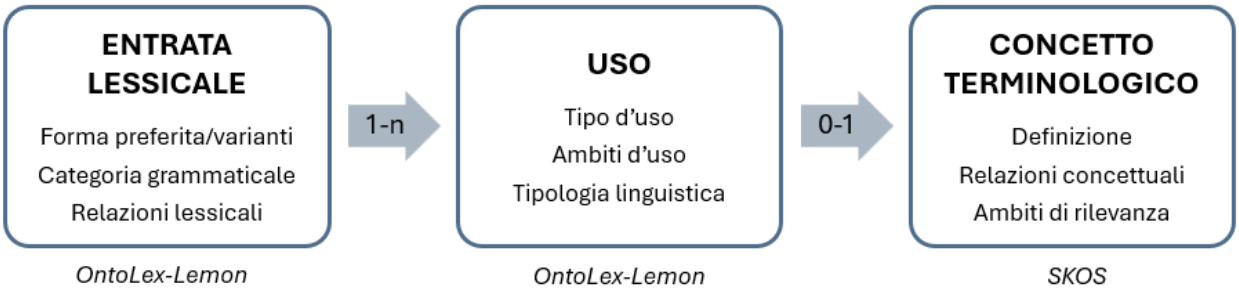


Figura 1. Workflow di costruzione del dataset LOD del GIR

Il punto di partenza è costituito dalle voci redazionali del *Glossario*, inizialmente prodotte in formato testuale. Tali contenuti sono stati progressivamente trasferiti in una tabella di lavoro strutturata (Excel), che ha fatto da base per la trasformazione dei dati. Questa fase non ha rappresentato una semplice trasposizione dei contenuti, ma una prima operazione di modellizzazione, in cui i campi e le relazioni vengono esplicitati, normalizzati e organizzati secondo i livelli concettuali previsti dal modello.
La fase successiva è quella della normalizzazione e del mapping sullo standard Resource Description Framework (RDF)[4] realizzati tramite OpenRefine, utilizzato come ambiente intermedio per la pulizia dei dati, il controllo dei valori appartenenti a vocabolari chiusi e la generazione delle strutture RDF di base. L'esportazione avviene secondo il modello OntoLex per i livelli lessicali, producendo file in formato Turtle.[5] I file RDF sono successivamente importati in un triplestore semantico (GraphDB),[6] che svolge un ruolo centrale non solo come ambiente di interrogazione secondo il protocollo di query SPARQL,[7] ma anche come spazio di verifica, integrazione e rifinitura del grafo. In questa fase vengono gestiti manualmente alcuni collegamenti strutturali che non possono essere generati automaticamente, in particolare le relazioni tra entrate lessicali e usi. Questo limite è dovuto a vincoli noti degli strumenti di mapping tabellare, che rendono complessa la gestione di relazioni uno-a-molti semanticamente dipendenti. Tali relazioni sono pertanto aggiunte in una fase successiva di post-processing direttamente sul grafo RDF, all'interno del triplestore. Questa scelta consente di preservare la correttezza semantica delle relazioni, mantenendo al contempo la flessibilità del modello.
Il livello concettuale, modellato in SKOS, viene inizialmente popolato tramite inserimento manuale, poiché le definizioni e le relazioni concettuali richiedono un controllo editoriale pieno e non sono derivate automaticamente dalla tabella di partenza.
Nel complesso, il workflow adottato può essere definito semi-automatico: l'automazione è utilizzata per la produzione delle strutture di base, mentre le decisioni semanticamente rilevanti restano affidate a un controllo umano. Questa impostazione riflette una scelta metodologica consapevole, che privilegia l'affidabilità semantica e terminologica rispetto a una completa automazione del processo.

3. SCELTE DI MODELLIZZAZIONE LOD

Il modello concettuale LOD del GIR, sviluppato a partire dalla struttura delle voci del *Glossario*, è progettato per distinguere in modo esplicito tra la forma linguistica, il contesto d'uso e il significato della forma. A questo scopo, viene adottata una struttura a tre livelli (Fig. 2) che separa sistematicamente entrate lessicali, usi e concetti terminologici, evitando sovrapposizioni nella rappresentazione di risorse terminologiche.

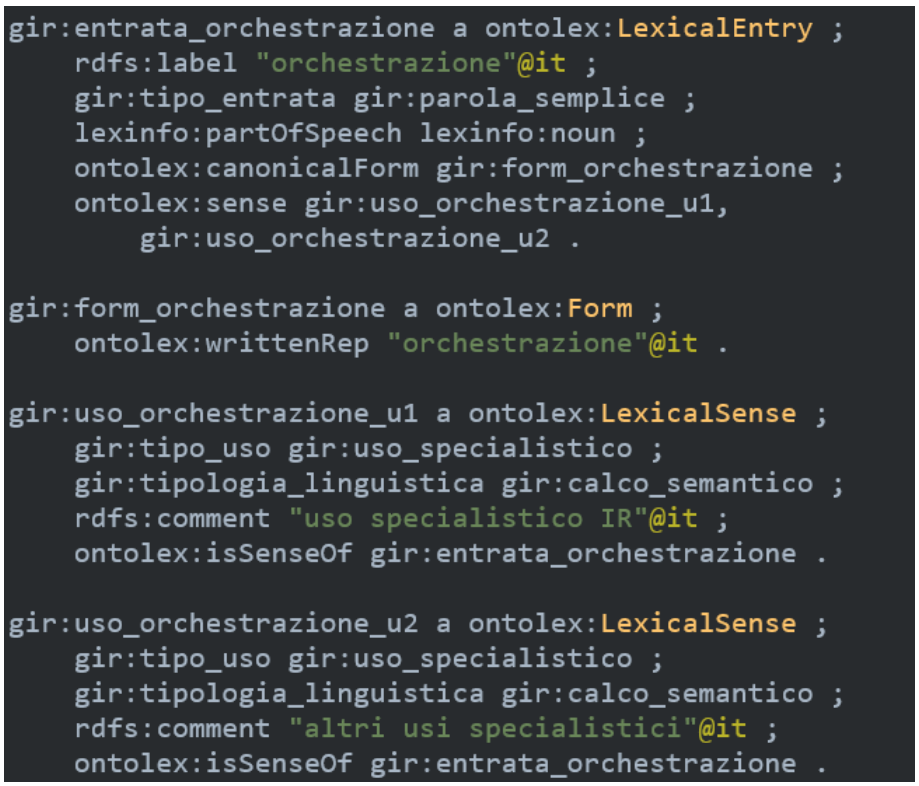


[4] https://www.w3.org/RDF/ (cons. 5/2/2026).
[5] https://www.w3.org/TR/turtle/ (cons. 5/2/2026).
[6] https://graphdb.ontotext.com/ (cons. 5/2/2026).
[7] https://www.w3.org/TR/sparql11-query/ (cons. 5/2/2026).

Figura 2. Modello concettuale LOD del GIR

Il livello dell'entrata lessicale è legato alla forma linguistica ed è modellato tramite OntoLex-Lemon. Ogni entrata lessicale (ontolex:*LexicalEntry*) corrisponde a una voce del GIR ed è associata a un lemma, assunto come forma canonica (ontolex:*canonicalForm*) dell'entrata, scelta secondo criteri predefiniti,[8] e a eventuali varianti, che non generano nuove entrate ma sono considerate realizzazioni alternative della stessa unità lessicale (varianti grafiche, forme in altra lingua).
Ciascuna entrata è collegata a una categoria grammaticale. Il vocabolario *LexInfo*[9] del modello OntoLex consente di mappare sostantivi, aggettivi e verbi; tuttavia, non prevede valori di *lexinfo:partOfSpeech* per acronimi e locuzioni. Per evitare di forzare la codifica grammaticale di unità lessicali, si ricorre all'uso di classi e/o proprietà dedicate, mantenendo distinta la descrizione della natura formale dell'entrata dalla sua categorizzazione morfosintattica. Nel caso degli acronimi, che costituiscono una tipologia particolarmente rappresentata nel lessico delle IR, la sigla e la corrispondente forma estesa sono modellate come entrate lessicali distinte (*ontolex:LexicalEntry*), ciascuna dotata della propria forma canonica. Le due entrate sono collegate tra loro tramite una relazione reciproca espressa attraverso Lexinfo:abbreviationFor. Ad esempio, l'entrata lessicale FAIR e l'entrata lessicale Findable_Accessible_Interoperable_Reusable sono entrambe modellizzate come *ontolex:LexicalEntry* e hanno ciascuna la rispettiva *ontolex:CanonicalForm.* Il vocabolario *LexInfo* le connette attraverso la proprietà *lexinfo:abbreviationFor*.[10]
Il terzo campo previsto in questo livello è quello delle voci correlate, che fa riferimento ad altre entrate del GIR collegate all'entrata principale da relazioni di natura derivazionale, riconducibili a una stessa base lessicale.
Il livello dell'uso costituisce il fulcro del modello ed è rappresentato tramite *ontolex:LexicalSense*. L'uso è definito come l'unità che descrive il valore semantico che una forma assume in un determinato contesto, ed è il livello in cui si manifestano in modo esplicito la polisemia e i processi di specializzazione.
Ogni entrata può avere uno o più usi. Ciascun uso è associato a un tipo (che distingue tra uso specialistico oppure uso della lingua comune), a uno o più ambiti di attestazione[11] e a una tipologia linguistica, che descrive il processo di formazione o risemantizzazione del termine (formazione italiana, prestito, calco, adattamento).
Per l'identificazione degli usi associati alle diverse entrate si è fatto ricorso a dizionari e repertori terminologici e lessicali; nei casi in cui il lemma non risultasse registrato in tali fonti, gli usi sono stati individuati attraverso l'analisi diretta dei contesti testuali in cui il vocabolo è attestato (ad esempio deliverable di progetto, schede tecniche, tutorial e descrizioni dei servizi).
La tipologia linguistica non fa riferimento al lemma in generale, ma è associata all'accezione che la voce assume nel contesto delle IR. La scelta di collocare la tipologia linguistica a livello dell'uso, e non dell'entrata o del concetto, risponde a un'esigenza precisa: i processi di calco semantico, estensione di significato o specializzazione riguardano il modo in cui un termine assume un valore in un determinato dominio. Un esempio significativo è offerto dal termine *orchestrazione* (insieme con il verbo *orchestrare* e il sostantivo *orchestratore*), già attestato in italiano in contesti quali la musica e la critica letteraria, ma che, come calco dell'inglese *orchestration,* acquisisce in ambito informatico un significato estensivo legato alla configurazione, al coordinamento e alla gestione di sistemi e flussi di lavoro automatizzati, accezione impiegata anche nel dominio delle IR. La Fig. 3 mostra come le scelte di modellizzazione descritte sopra siano rese operativamente in RDF, attraverso un estratto relativo alla voce *orchestrazione*.

[8] Di norma, la scelta dell'entrata ricade sulla forma più attestata nel corpus; qualora una voce sia attestata sia in italiano sia in inglese, si è scelto di promuovere la variante italiana.
[9] https://lexinfo.net/ (cons. 5/2/2026).
[10] https://www.w3.org/2016/05/ontolex/ (cons. 5/2/2026), cfr. sezione 3.2, l'esempio di "NASA".
[11] Gli ambiti d'uso corrispondono, nella versione editoriale del GIR, alle "marche d'uso".



```
gir:entrata_orchestrazione a ontolex:LexicalEntry ;
    rdfs:label "orchestrazione"@it ;
    gir:tipo_entrata gir:parola_semplice ;
    lexinfo:partOfSpeech lexinfo:noun ;
    ontolex:canonicalForm gir:form_orchestrazione ;
    ontolex:sense gir:uso_orchestrazione_u1,
        gir:uso_orchestrazione_u2 .

gir:form_orchestrazione a ontolex:Form ;
    ontolex:writtenRep "orchestrazione"@it .

gir:uso_orchestrazione_u1 a ontolex:LexicalSense ;
    gir:tipo_uso gir:uso_specialistico ;
    gir:tipologia_linguistica gir:calco_semantico ;
    rdfs:comment "uso specialistico IR"@it ;
    ontolex:isSenseOf gir:entrata_orchestrazione .

gir:uso_orchestrazione_u2 a ontolex:LexicalSense ;
    gir:tipo_uso gir:uso_specialistico ;
    gir:tipologia_linguistica gir:calco_semantico ;
    rdfs:comment "altri usi specialistici"@it ;
    ontolex:isSenseOf gir:entrata_orchestrazione .
```

Figura 3. Estratto del lessico RDF del GIR

Per la codifica di informazioni quali tipo di uso e tipologia linguistica è stato evitato l'uso di *dcterms:type*, in quanto proprietà troppo generica. Tali informazioni sono invece modellate tramite proprietà dedicate del vocabolario GIR, al fine di rendere il grafo più leggibile, meno ambiguo e più facilmente interrogabile. Solo l'uso rilevante per il dominio delle IR è collegato a un concetto terminologico e dotato di definizione; gli altri usi sono mantenuti ai fini documentari, per rendere conto della polisemia e del contesto d'uso, senza forzare una rappresentazione concettuale non pertinente.
Il livello concettuale è modellato in SKOS e descrive i concetti terminologici, intesi come oggetti di conoscenza astratti e non linguistici. Nel modello adottato, i concetti non sono collegati direttamente alle entrate lessicali, ma agli usi semanticamente rilevanti.
In questo terzo livello sono rappresentati le definizioni concettuali e gli ambiti di rilevanza (ambiti disciplinari o istituzionali in cui il concetto è rilevante), che non devono essere confusi con gli ambiti d'uso (ambiti in cui il termine è usato e attestato).

4. CONCLUSIONI

Questo contributo ha presentato le scelte metodologiche adottate per la modellizzazione LOD del *Glossario delle Infrastrutture di Ricerca*, assunto come caso di studio per riflettere sulle criticità che emergono nella rappresentazione di risorse terminologiche in lingua italiana. La distinzione tra entrata lessicale, uso e concetto terminologico ha consentito di affrontare in modo esplicito fenomeni quali la polisemia, la specializzazione semantica e la variazione d'uso, che nel dominio delle IR risultano particolarmente critici, senza sacrificare la stabilità concettuale.
Il modello proposto non nasce come applicazione diretta degli standard OntoLex e SKOS, ma come risultato di un processo di mediazione tra esigenze redazionali, complessità linguistica e interoperabilità semantica. In questo senso, le scelte di modellizzazione adottate riflettono un equilibrio negoziato tra aderenza agli standard del Web Semantico e usabilità effettiva della risorsa.
Pur essendo radicato in un contesto specifico, l'approccio illustrato intende offrire spunti metodologici replicabili per la progettazione di altre risorse terminologiche, in particolare in ambiti caratterizzati da forte interferenza tra lingua comune e linguaggi specialistici.
In prospettiva, il modello potrà essere esteso attraverso collegamenti con altre risorse del Web Semantico e con repertori terminologici istituzionali (es. IATE), nonché con knowledge graph generalisti come Wikidata. In particolare, l'integrazione potrà riguardare vocabolari affini (ad esempio nell'ambito

della Scienza Aperta) e risorse lessicali e terminologiche già pubblicate in ambiente LOD, favorendo l'interoperabilità e il riuso dei dati.
Il modello si presta inoltre a diversi scenari di utilizzo. La possibilità di consultare la terminologia attraverso interrogazioni sul grafo consentirà di accedere in modo strutturato ai dati del GIR, offrendo uno strumento di supporto alla comunità scientifica per la redazione e la comprensione dei documenti nell'ambito delle IR: sarà infatti possibile verificare il significato dei tecnicismi, disambiguare i termini e individuare le forme da impiegare per garantire una maggiore uniformità del lessico. L'interrogazione avanzata del *Glossario* consentirà inoltre di estrarre informazioni di varia natura, ad esempio conoscere tutti i termini associati a un dato ambito o individuare specifiche relazioni tra concetti.

BIBLIOGRAFIA

Bellandi, A. (2022). Le Risorse Linguistiche nell'era del Web Semantico: Un insieme di servizi informatici per la gestione di lessici e terminologie. AIDAinformazioni, 1-2, 31-50.

Bellandi, A. (2025). Lexicographic Resources in the Semantic Web: Models, Tools, and Case Studies. Umanistica Digitale, 9, 21, 107-40.

Bellandi, A., Di Nunzio, G. M., Piccini, S., Vezzani, F. (2023a). From TBX to OntoLex-Lemon: Issues and Desiderata. In Proceedings of the 2nd International Conference on Multilingual Digital Terminology Today (MDTT 2023), CEUR Workshop Proceedings, Lisbon, Portugal. CEUR. ISSN: 1613-0073.

Bellandi, A., Di Nunzio, G. M., Piccini, S., Vezzani, F. (2023b). The Importance of Being Interoperable: Theoretical and Practical Implications in Converting TBX to OntoLex-Lemon. Proceedings of the 4th Conference on Language, Data and Knowledge (LDK 2023), 646-51.

Bellandi, A., Di Nunzio, G. M., Piccini, S., Vezzani, F. (2024). LemonizeTBX: Design and Implementation of a New Converter from TBX to OntoLex-Lemon. Digital Humanities Quarterly, 18, 2.

Bosque-Gil, J., Gracia, J. (eds.) (2019). The OntoLex Lemon Lexicography Module. W3C Community Group Report.

Cimiano, P., McCrae, J. P., Rodríguez-Doncel, V., Gornostay, T., Gómez-Pérez, A., Siemoneit, B., & Lagzdins, A. (2015). Linked terminologies: applying linked data principles to terminological resources. In *Proceedings of the eLex 2015 Conference,* 504-17.

Cimiano, P., McCrae, J. P., Buitelaar, P. (2016). Lexicon Model for Ontologies: Community Report. W3C Community Group Final Report.

Cimiano, P., Chiarcos, C., McCrae, J. P., e Gracia, J. (2020). Linguistic Linked Data: Representation, Generation and Applications. Linguística: Revista De Estudos Linguísticos Da Universidade Do Porto, 16, 111–117. https://ojs.letras.up.pt/index.php/EL/article/view/11023.

Di Buono, M. P., Cimiano, P., Elahi, M. F., & Grimm, F. (2020). Terme-à-LLOD: Simplifying the Conversion and Hosting of Terminological Resources as Linked Data. In Proceedings of the 12th Language Resources.

Bosque-Gil, J., Gracia, J., Aguado-de-Cea, G., & Montiel-Ponsoda, E. (2015). Applying the ontolex model to a multilingual terminological resource. In European Semantic Web Conference, Cham, Springer International Publishing, 283-94.

Diez-Ibarbia, P., Martín-Chozas, P., & Montiel-Ponsoda, E. (2025). The Art of Modelling Terminology Resources with Ontolex-lemon and Semantic Web Standards: the TERMCAT Usecase. In Journal on Digital Terminology and Lexicography, 1 (JDTL-Volume 1, issue 1), 1-24.

Falini, I., & Francalanci, L. (2025). Dal cantiere del Glossario delle Infrastrutture di ricerca (GIR): indagini linguistiche su un primo nucleo di voci, Bollettino dell'Opera Del Vocabolario Italiano, XXX, 361-90.

Francalanci, L., Scognamiglio, A., & Falini, I. (2025). Per una terminologia univoca, condivisa e non ambigua per le Infrastrutture di Ricerca: il Glossario delle Infrastrutture di Ricerca (GIR), TermCD, TERMinologie, Communication et Discours, III, in corso di stampa.

Francalanci, L., Scognamiglio, A., Falini, I., Restaneo, P., & Pedonese, G. (2025). Il Glossario delle Infrastrutture di Ricerca (GIR), in Rebora, S., Rospocher, M., & Bazzaco, S. (a cura di), Diversità, Equità e Inclusione: Sfide e Opportunità per l'Informatica Umanistica nell'Era dell'Intelligenza Artificiale. Atti del XIV Convegno Annuale dell'AIUCD, Verona, 11-13 giugno 2025, Quaderni di Umanistica Digitale, 2025, 24-28.

Litta, E., Passarotti, M., Brasolin, P., Moretti, G., Basile, V., Di Fabio, A., Bosco, C. (2024). The Lemma Bank of the LiITA Knowledge Base of Interoperable Resources for Italian. In F. Dell'Orletta et al. (eds.), Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024). CEUR Workshop Proceedings, 517-22.

Litta, E., Passarotti, M. C., Basile, V., Bosco, C., Di Fabio, A., Brasolin, P. (2025). LiITA: a Knowledge Base of Interoperable Resources for Italian. In M. Alam et al. (eds.), Proceedings of the 5th Conference on Language, Data and Knowledge. Napoli: UniorPress, 130-35.

Mambrini, F., Passarotti, M., Sporleder, C., *et al.* (2020). LiLa: Linking Latin. Risorse linguistiche per il latino nel Semantic Web. Umanistica Digitale, 8, 63–78.

Martín-Chozas, P., Declerck, T., Montiel-Ponsoda, E., & Rodríguez-Doncel, V. (2025). Representing terminological data in the Semantic Web: A proposal based on OntoLex-lemon. Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication, 31, 2, 171-207.

McCrae, J. P., Bosque-Gil, J., Gracia, J., Buitelaar, P., Cimiano, P. (2017). The OntoLex-Lemon Model: Development and Applications. Proceedings of eLex.

Miles, A., & Bechhofer, S. (2009). SKOS Simple Knowledge Organization System Reference. W3C Recommendation.

Passarotti, M., Mambrini, F., Franzini, G., Cecchini, F. M., Litta, E., Moretti, G., Ruffolo, P., & Sprugnoli, R. (2020). Interlinking through Lemmas. The Lexical Collection of the LiLa Knowledge Base of Linguistic Resources for Latin. Studi e Saggi Linguistici, LVIII, 1, 177–212.

Orchestrazione di servizi e scenari avanzati: fruizione e necessità delle comunità di ricerca nel marketplace di H2IOSC

PIETRO SICHERA[1], CRISTINA MARRAS[1], ENRICO PASINI[2], VITTORIA FABIANI[1], PAOLO ONGARO[3], MICHELE SCAPICCHI[3], CHIARA DI PIETRO[3]

[1]CNR-ILIESI
[2]UNIVERSITÀ DI TORINO
[3]DEDA NEXT S.R.L.

ABSTRACT (ITALIANO)

A partire dal marzo 2023, il nodo italiano dell'infrastruttura di ricerca OPERAS (CNR-ILIESI) ha curato lo studio di fattibilità, la specifica tecnica e il modello di orchestrazione dei servizi per il Marketplace del progetto H2IOSC – Humanities and Heritage Italian Open Science Cloud. Il Marketplace è progettato per consentire di integrare risorse eterogenee provenienti dalle infrastrutture di ricerca federate (OPERAS, E-RIHS, CLARIN, DARIAH), favorendo da un lato la cooperazione e la condivisione tra ecosistemi digitali, dall'altro la collaborazione e il coinvolgimento attivo degli utenti.
Il contributo descrive la soluzione API-driven della piattaforma, con un'architettura a tre livelli dove i workflow applicativi e descrittivi diventano elementi qualificanti del modello. In particolare, viene formalizzata la distinzione tra orchestrazione (in cui il Marketplace coordina servizi eseguiti remotamente su sistemi esterni) e composizione (in cui i servizi operano localmente all'interno di una singola piattaforma). I workflow applicativi e descrittivi diventano elementi qualificanti del modello, rappresentando non solo strumenti tecnici, ma servizi capaci di estendere le funzionalità del Marketplace e di abilitare scenari avanzati di elaborazione e integrazione dei dati e funzionalità collaborative preservando la responsabilità dei fornitori e la sostenibilità del sistema.
Viene descritto il ruolo dei protocolli e dei flussi di harvesting basati su OAI-PMH, la visualizzazione delle risorse, le modalità di ricerca, la creazione e l'esecuzione di workflow applicativi tramite WSO2 Micro Integrator, e gli strumenti di analisi di utilizzo.
Il Marketplace risponde così alle necessità di progetto e ai bisogni degli utenti, configurandosi come infrastruttura socio-tecnica capace di sostenere pratiche di ricerca e non come semplice repertorio descrittivo.
Parole chiave: marketplace; API; orchestrazione; workflow; sostenibilità.

ABSTRACT (ENGLISH)

Since March 2023, the Italian node of the OPERAS research infrastructure (CNR-ILIESI) has been responsible for the feasibility study, technical specification, and service orchestration model for the H2IOSC – Humanities and Heritage Italian Open Science Cloud – Marketplace project. The Marketplace is designed to integrate heterogeneous resources from federated research infrastructures (OPERAS, E-RIHS, CLARIN, DARIAH), fostering cooperation and sharing between digital ecosystems on the one hand, and collaboration and active user engagement on the other.
The paper describes the platform's API-driven solution, with a three-tier architecture where application and descriptive workflows become key elements of the model. Specifically, the distinction between orchestration (in which the Marketplace coordinates services executed remotely on external systems) and composition (in which services operate locally within a single platform) is formalized. Application and descriptive workflows become key elements of the model, representing not only technical tools but also services capable of extending the Marketplace's functionality and enabling advanced data processing and integration scenarios and collaborative features while preserving supplier accountability and system sustainability.
The role of OAI-PMH-based harvesting protocols and flows, resource visualization, search methods, the creation and execution of application workflows using WSO2 Micro Integrator, and usage analysis tools are described.
The Marketplace thus responds to project and user needs, becoming a socio-technical infrastructure capable of supporting research practices rather than a simple descriptive repository.

Keywords: marketplace; API; orchestration; workflow; sustainability.

1. INTRODUZIONE: RIDEFINIZIONE DEGLI SPAZI EUROPEI DELLA RICERCA COLLABORATIVA

Accessibilità ed *engagement* (nel senso più ampio di partecipazione attiva) sono due principi ormai fondamentali per una cittadinanza digitale, per favorire la democratizzazione dell'accesso ai contenuti e al patrimonio culturale, per coinvolgere attivamente i cittadini nei processi culturali, per incentivare l'alfabetizzazione digitale, e la ricerca collaborativa e aperta: "open science, including mainstreaming open access to publications and research data, has a crucial role in boosting impact, quality, efficiency, transparency and integrity of research and innovation, and brings science and society closer together whilst taking into account legitimate legal, security and privacy aspects" (Council of the European Union, 2023).
Nell'ambito delle Digital Humanities (DH) questi principi sono al centro dell'attenzione da più di vent'anni impegnando le comunità scientifiche nello sfruttare le potenzialità delle tecnologie a tutti i livelli, facendo del coinvolgimento partecipativo una componente essenziale anche del design di progetti: collaborazione tra ricercatori, identificazione degli stakeholders e degli utenti, e corresponsabilizzazione delle comunità scientifiche di riferimento (Tomasi, Vitali 2013).[1]
Molta attenzione viene data all'inclusività e accessibilità delle interfacce, ai frontend, e alla fairificazione dei dati (FAIR - Findeable, Accessible, Interoperable, Reusable), mentre meno frequente è l'attenzione relativa ai backend/backoffice dei sistemi, ai protocolli e alle API (*Application Programming Interface*) che garantiscono la robustezza e la sostenibilità del sistema, l'accessibilità e l'apertura alla dimensione collaborativa, come anche la fiducia degli utenti.
In questa proposta si prende in considerazione il Marketplace sviluppato nell'ambito del progetto H2IOSC,[2] un *discovery portal*, dal design di frontend e backend robusto per rappresentare e soddisfare al meglio le sue finalità, ovvero quello di essere uno spazio digitale unificato dove servizi, strumenti, dataset, risorse di apprendimento, progetti, workflow, pubblicazioni sono resi disponibili per supportare le molteplici esigenze della ricerca e seguire l'intero ciclo di vita dei dati. Un aspetto di particolare interesse è che il MarketPlace costituisce il punto di convergenza dei servizi della federazione H2IOSC, fungendo da elemento unificante tra il mondo della ricerca e l'esterno. Si inserisce, così, nel circuito di ecosistemi di ricerca digitale aperti e interoperabili, promuovendo l'accessibilità, l'approccio FAIR ai dati e l'integrazione di nuovi servizi, facilitando l'ampliamento progressivo dell'ecosistema stesso.
L'architettura del Marketplace e il suo modello di orchestrazione dei servizi sono stati ideati e formalizzati dal nodo italiano dell'infrastruttura OPERAS (CNR-ILIESI). La Figura 1 presenta la cronologia del lavoro: dalla definizione prototipale iniziale con lo studio di fattibilità (aprile 2023) dettagliata nella specifica tecnica per la gara di affidamento (giugno 2023), alla definizione dell'architettura di orchestrazione (novembre 2024) e delle specifiche operative (febbraio 2025) fino all'implementazione operativa (marzo 2026) (Sichera, Marras, Pasini 2024 e 2026b)[3].

[1] Cfr. Progetto Discovery: http://www.discovery-project.eu/home.html) una federazione di collezioni ed edizioni digitali di testi filosofici e scientifici, interconnessi in un network collaborativo per annotazioni, commenti e ricerche.
[2] H2IOSC - Humanities and Heritage Italian Open Science Cloud (IR0000029), Missione 4 "Istruzione e Ricerca" – Componente 2 "Dalla ricerca all'impresa" – Linea di Investimento 3.1 "Fondo per la realizzazione di un sistema integrato di infrastrutture di ricerca e innovazione", Azione 3.1.1 "Creazione di nuove IR o rafforzamento di quelle esistenti che contribuiscono agli obiettivi di Eccellenza Scientifica di Horizon Europe e costituzione di reti"
[3] L'intero percorso progettuale e implementativo è documentato in una serie di pubblicazioni tecniche che costituiscono il riferimento per le attività di integrazione dei servizi nell'ambito del progetto H2IOSC (Sichera, Marras, Pasini 2026a; 2026c; 2026d; 2026e).

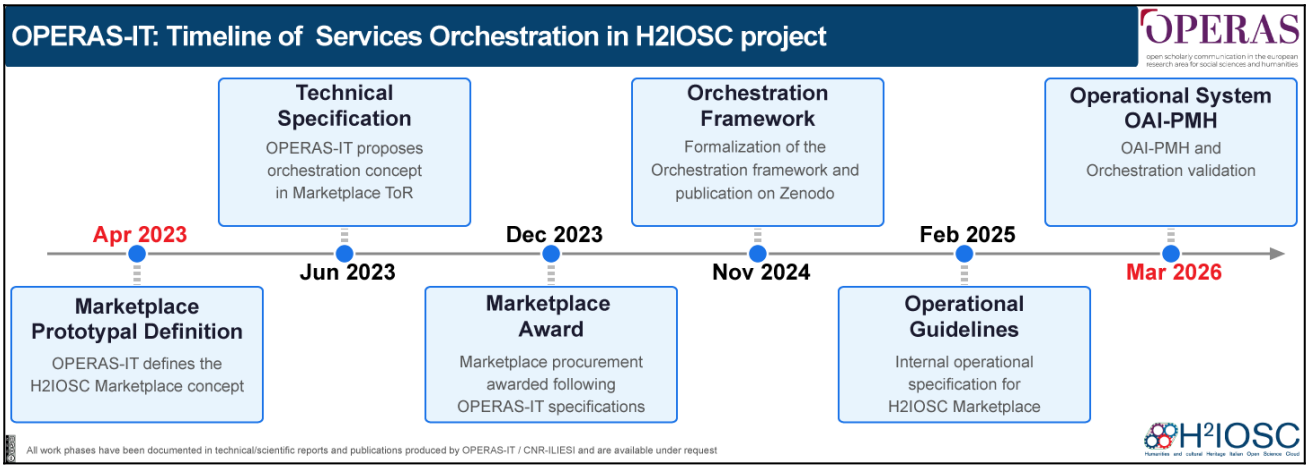


Figura 1 - Orchestrazione dei servizi definita da OPERAS-IT / CNR-ILIESI nell'ambito del progetto H2IOSC.

Ma in che modo un marketplace DH può trasformarsi da repertorio descrittivo a infrastruttura in grado di sostenere pratiche d'uso e percorsi di *engagement*, mantenendo trasparenza e controllabilità dei processi? Come progettare un livello di interoperabilità che renda integrabili risorse eterogenee senza imporre uniformità "a monte"? Come abilitare workflow applicativi senza centralizzare l'esecuzione, preservando la responsabilità dei fornitori e la sostenibilità del sistema? La risposta si articola in una soluzione API-driven con architettura a tre livelli e con una distinzione tra workflow descrittivi e workflow applicativi orchestrati, considerati come un elemento qualificante del modello H2IOSC.
Il Marketplace di H2IOSC è progettato, per interoperare con piattaforme analoghe a livello europeo, abilitando lo scambio sicuro di informazioni e metadati e, quindi, l'integrazione di risorse eterogenee. Descriveremo i protocolli e i flussi di harvesting (¶2), la visualizzazione delle risorse – contestualizzazione e serendipity ¶3); le modalità di ricerca multilineare (¶4); la creazione di workflow applicativi (¶5); proporremo alcune brevi considerazioni su strumenti e analisi per venire incontro alle necessità delle comunità di ricerca (¶6); in conclusione (¶7) lo scopo è mostrare la dialettica che intercorre tra istanze socio-culturali e scelte infrastrutturali, tra frontend orientato alla discovery e backend orientato alla governance, tra apertura/riuso e sostenibilità operativa. Il Marketplace, nell'implementazione di backend, backoffice e frontend risponde così alle necessità di progetto: come essere un'infrastruttura modulare, interoperabile e guidata da API che permette l'orchestrazione servizi e risponde alle necessità degli utenti in generale e della comunità scientifica in particolare.

2. PROTOCOLLI E FLUSSI DI HARVESTING

Il popolamento del catalogo delle risorse di Marketplace avviene attraverso un processo strutturato di "harvesting dei dati", che consente di raccogliere e aggiornare automaticamente dati e metadati da fonti esterne qualificate, raccoglie infatti in modo automatizzato e interoperabile le informazioni provenienti dai diversi fornitori di contenuti – infrastrutture di ricerca, istituzioni e repository esterni – garantendo uniformità, aggiornamento e scalabilità del sistema. Il processo è basato sul protocollo standard OAI-PMH,[4] formato preferito in contesti accademici o istituzionali.
Il flusso si attiva abilitando gli Harvesting Provider, ovvero le sorgenti esterne abilitate all'esposizione dei dati tramite endpoint OAI-PMH. Ciascun provider rende disponibili i propri record descrittivi (item, risorse, prodotti, dataset, ecc.) secondo schemi condivisi. Il Marketplace, in base a diverse configurazioni possibili, interroga periodicamente queste sorgenti, eseguendo processi di raccolta incrementale basati su finestre temporali, così da acquisire sia nuove pubblicazioni sia eventuali aggiornamenti di record già esistenti.
Durante i *job* di harvesting, il sistema esegue una fase di validazione tecnica, in cui vengono intercettate eventuali anomalie – errori di struttura, campi mancanti, duplicazioni – registrate nei log e rese visibili nelle dashboard statistiche, così da supportare attività di monitoraggio e correzione. I record validati vengono quindi importati nel Marketplace e resi disponibili alla consultazione pubblica secondo le logiche di ricerca e visualizzazione che mette a disposizione il frontend. Questo processo consente di mantenere il catalogo costantemente aggiornato, riducendo al minimo gli interventi manuali e assicurando tracciabilità

[4] OAI-PMH *Open Archives Initiative Protocol for Metadata Harvesting* https://www.openarchives.org/pmh/

delle operazioni. Dal punto di vista strategico, l'adozione del protocollo OAI-PMH offre diversi vantaggi: interoperabilità, standardizzazione, automazione, possibilità di estendere facilmente il numero di fornitori integrati. Nell'ambito del progetto H2IOSC, il gruppo di lavoro di OPERAS ha reso disponibile un'applicazione HTML *single-page* per l'inserimento di metadati delle risorse organizzate secondo il modello dati H2IOSC e compatibile col protocollo OAI-PMH (Sichera, Marras, Pasini 2026a; 2026f).

3. VISUALIZZAZIONE DELLE RISORSE DEL CATALOGO

Una volta acquisite, normalizzate e validate attraverso il flusso di harvesting, le risorse vengono rese disponibili alla consultazione pubblica tramite il frontend del Marketplace. La loro visualizzazione è progettata per garantire accesso strutturato, leggibilità delle informazioni e facilità di esplorazione, valorizzando al tempo stesso la ricchezza dei metadati raccolti.
Ogni risorsa è rappresentata attraverso una scheda descrittiva che aggrega e organizza i principali elementi informativi provenienti dalle sorgenti originarie. La struttura di visualizzazione segue il modello dati H2IOSC Marketplace, tra cui titolo e descrizione sintetica della risorsa, tipologia di appartenenza, fornitore e istituto di riferimento, parole chiave e ambiti disciplinari, informazioni tecniche e di accesso, collegamenti alla sorgente originale o al repository esterno, varie funzioni specifiche per gli utenti che si registrano al Marketplace.

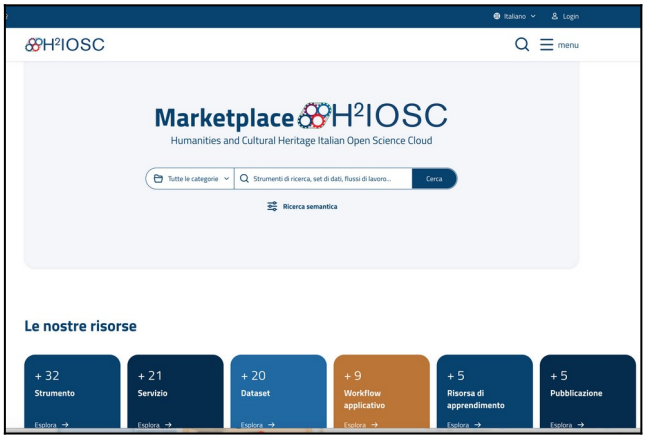


Figura 2 – Home page del Marketplace di H2IOSC

L'interfaccia del Marketplace è progettata secondo criteri di gerarchia visiva e accessibilità, con sezioni modulari che facilitano la lettura dei contenuti. Particolare attenzione è dedicata alla contestualizzazione della risorsa all'interno dell'ecosistema: ogni scheda consente infatti di navigare verso risorse correlate, progetti affini, istituzioni coinvolte o ulteriori contenuti dello stesso fornitore.
Dal punto di vista UX, la visualizzazione delle risorse risponde a tre obiettivi principali: comprensione immediata del valore della risorsa, tracciabilità della provenienza scientifica e istituzionale, facilitazione dell'accesso alla risorsa o ai suoi dati associati.
In questo modo, il frontend non si limita a esporre record descrittivi, ma costruisce un ambiente informativo navigabile che supporta attività di ricerca, analisi e riuso.

4. RICERCA MODULARE E MULTILINEARE

Per consentire un'esplorazione efficace del patrimonio informativo del Marketplace, sono previste diverse metodologie di ricerca, progettate per rispondere a bisogni conoscitivi eterogenei, dalla consultazione esplorativa alla ricerca specialistica.
La modalità più immediata è la ricerca libera, che consente di interrogare l'intero corpus delle risorse tramite inserimento di parole chiave. Il motore full-text i contenuti restituendo risultati ordinati per pertinenza.
Accanto alla ricerca libera, il sistema offre una navigazione filtrata basata su "faccette". Gli utenti possono restringere progressivamente il perimetro dei risultati applicando filtri quali categoria dell'item, tipologia di risorsa, fornitore, istituzione, ambito disciplinare, periodo di pubblicazione. Questo approccio consente una ricerca guidata, particolarmente utile in contesti ad alta densità informativa. Il Marketplace consente

anche all'utente di salvare le ricerche effettuate al fine di rieseguirle successivamente in modo più rapido, senza dover reimpostare tutti i parametri e filtri.
Per esigenze analitiche più complesse è prevista una modalità di ricerca avanzata, che permette di combinare più parametri contemporaneamente. Gli utenti possono costruire query strutturate incrociando campi specifici, affinando così la precisione dei risultati. Questa funzionalità risponde in particolare ai bisogni di ricercatori, analisti e stakeholder istituzionali. Inoltre, grazie ad un sistema di notifiche in-app, il Marketplace avvisa l'utente ogni volta in cui vengono effettuati aggiornamenti alle risorse per i quali ha espresso un interesse sia salvate nei preferiti e sia nei risultati di ricerca.

5. CREAZIONE ED ESECUZIONE DEI WORKFLOW APPLICATIVI

All'interno del Marketplace è prevista una funzionalità dedicata alla gestione dei workflow applicativi, concepita per consentire la creazione, pubblicazione ed esecuzione di pipeline operative integrate con l'infrastruttura dei servizi (Sichera, Marras, Pasini 2026c; 2026d; 2026h; 2026i). Il processo prende avvio dall'interfaccia di backoffice, attraverso la configurazione iniziale che consente il caricamento di un file JSON Manifest, documento descrittivo contenente metadati e specifiche tecniche del workflow.
Il caricamento del manifest non è vincolante: l'utente può procedere anche in assenza del file, configurando manualmente i campi richiesti. Qualora il manifest venga fornito, il sistema provvede a estrarre automaticamente i metadati disponibili, mappandoli, per quanto possibile, sul modello dati del Marketplace, semplificando così la fase di configurazione.
Una volta creato il workflow, l'utente accede alla relativa sezione di modifica, dove sono disponibili i campi per la gestione e l'editing dei metadati e un comando dedicato al caricamento del file ".car" che descrive il workflow (o del file XML equipollente).
Questo file rappresenta la definizione esecutiva del workflow e costituisce un prerequisito tecnico per le fasi successive di pubblicazione, orchestrazione ed esecuzione. Il caricamento del file di orchestrazione abilita il pulsante "Pubblica". Attraverso questa azione, il sistema invoca un servizio dedicato che gestisce la procedura di pubblicazione e il deploy del workflow sull'infrastruttura WSO2, rendendolo disponibile all'esecuzione. Solo a seguito di una pubblicazione avvenuta con esito positivo, si attiva il comando "Esegui", che consente di avviare operativamente la pipeline.
L'esecuzione del workflow avviene tramite una finestra modale che permette l'inserimento dei dati di input necessari e la successiva attivazione tramite il pulsante "Avvia". Qualora il JSON Manifest caricato includa un riferimento a specifiche OpenAPI, il sistema procede automaticamente al parsing della documentazione, generando dinamicamente la maschera di input strutturata. In caso di errore nel parsing, o su preferenza dell'utente, è comunque possibile inserire manualmente i dati in formato JSON all'interno di un campo textarea dedicato, garantendo flessibilità operativa.
Una volta inseriti i dati di input, se necessari, l'utente può avviare il workflow. L'esecuzione avviene tramite chiamata a un endpoint dedicato, al quale vengono trasmessi, come payload, i dati inseriti.
Il processo è gestito in modalità asincrona. Ciò significa che l'utente riceve una notifica immediata di avvio e che l'esecuzione prosegue lato infrastruttura senza necessità di presidio attivo.
Un sistema di monitoraggio schedulato tramite "cron" verifica periodicamente lo stato dell'esecuzione, aggiornando le notifiche fino al completamento del processo.
Dal punto di vista implementativo, l'esecuzione asincrona è supportata dal layer di orchestrazione WSO2: la richiesta dell'utente viene persistita in un *message store* e instradata verso i servizi remoti tramite una sequenza di mediatori; al termine dell'elaborazione, il completamento può essere notificato al Marketplace tramite callback oppure reso interrogabile attraverso un endpoint di stato per polling. In questo modo si separano chiaramente tempo di risposta e tempo di elaborazione, si evitano timeout lato client e si mantengono tracciabilità e auditability dell'intero ciclo di esecuzione.
Una sezione dedicata consente la consultazione storica di tutte le esecuzioni avviate, con stato di avanzamento, esito finale, eventuali messaggi di errore, dati di input utilizzati, risultati prodotti dal workflow. Questa area di monitoraggio garantisce tracciabilità completa delle operazioni, supportando attività di auditing, debugging e riutilizzo delle configurazioni.
Il flusso descritto consente di integrare in Marketplace pipeline applicative complesse garantendo l'automazione delle esecuzioni, il controllo sugli stati di lavorazione e la riusabilità dei workflow. In questa prospettiva, i workflow applicativi non rappresentano solo strumenti tecnici, ma veri e propri servizi orchestrati, capaci di estendere le funzionalità del Marketplace e di abilitare scenari avanzati di elaborazione e integrazione dei dati.

Un aspetto qualificante del modello architetturale del Marketplace H2IOSC è la distinzione formale tra orchestrazione e composizione di servizi (Sichera, Marras, Pasini 2026b). Si parla di orchestrazione quando il Marketplace coordina servizi che vengono eseguiti remotamente, su sistemi informatici esterni alla piattaforma, gestendo le interazioni attraverso protocolli di rete, meccanismi di polling e callback asincroni. Si parla invece di composizione quando tutti i servizi necessari all'esecuzione di un workflow sono interni a un'unica piattaforma e possono essere invocati localmente in modo sincrono (Degl'Innocenti et al. 2025). Questa distinzione riflette la compresenza, nell'architettura della federazione H2IOSC, di livelli differenti di integrazione: un workflow composto all'interno di una singola infrastruttura può essere esposto nel Marketplace come metaservizio e partecipare a orchestrazioni più ampie che coinvolgono servizi di infrastrutture diverse. Un esempio concreto di orchestrazione è la pipeline automatizzata che collega un server IIIF, una piattaforma di HTR (eScriptorium[5]) e un sistema di pubblicazione digitale (TEI Publisher[6]): tre servizi ospitati su infrastrutture distinte, coordinati dal Marketplace senza che quest'ultimo esegua direttamente alcuna computazione (Sichera, Marras, Pasini 2026c).
Questo modello si distingue da approcci basati sulla composizione, in cui i servizi devono essere registrati e distribuiti all'interno della stessa istanza della piattaforma di integrazione.

6. LETTURA INTEGRATA E DINAMICA DEI DATI PER LE NECESSITÀ DELLE COMUNITÀ DI RICERCA

All'interno del backoffice del Marketplace è presente una dashboard dedicata alla visualizzazione e all'analisi delle statistiche della piattaforma (insight), concepita come strumento operativo a supporto delle attività di monitoraggio, governance e miglioramento continuo del servizio.
Questo è reso possibile dall'architettura modulare a tre livelli e dal disaccoppiamento (decoupling) tra persistenza dei dati, servizi applicativi e interfacce: la piattaforma può combinare servizi in workflow "su misura" senza imporre migrazioni dei contenuti, mantenendo separazione delle responsabilità e favorendo interoperabilità e manutenibilità.
La dashboard è accessibile da backoffice ai profili abilitati e consente di consultare in forma sintetica e analitica i principali indicatori relativi sia alle pubblicazioni degli item sia alle interazioni degli utenti con l'ecosistema del Marketplace. Una prima area della dashboard è dedicata all'analisi delle pubblicazioni delle risorse, permettendo di osservare l'andamento del popolamento del Marketplace e di individuare eventuali criticità nei processi di ingestion (Sichera et al. 2025). Accanto alle statistiche di pubblicazione, la dashboard integra un'area dedicata all'analisi dei comportamenti degli utenti, offrendo una lettura delle modalità di fruizione del Marketplace. Questi dati permettono di analizzare non solo cosa viene pubblicato, ma come viene utilizzato, fornendo insight sui comportamenti di ricerca e sulle necessità emergenti delle comunità scientifiche.
L'obiettivo è offrire un ambiente di lettura integrata dei dati, in grado di restituire non solo misurazioni quantitative, ma anche evidenze utili all'interpretazione dei comportamenti d'uso e delle dinamiche di alimentazione dei contenuti.

7. CONCLUSIONI

Il caso del Marketplace H2IOSC mostra come un marketplace DH possa evolvere da repertorio descrittivo a infrastruttura socio-tecnica capace di sostenere pratiche di ricerca: la qualità dei metadati è garantita da processi di harvesting standardizzati e validati; la discovery è potenziata da strategie di visualizzazione e ricerca multilivello; soprattutto, la possibilità di pubblicare ed eseguire workflow applicativi come servizi orchestrati consente di integrare strumenti e dati eterogenei senza imporre centralizzazione dell'esecuzione o migrazione dei contenuti. Il layer di orchestrazione, basato su API gateway e pattern asincroni, introduce governance (autenticazione, policy, audit) e sostenibilità (quote, monitoraggio) come requisiti infrastrutturali abilitanti per *engagement* e fiducia.
La distinzione tra orchestrazione e composizione, formalizzata nel modello OPERAS-IT (Sichera, Marras, Pasini 2026b), fornisce un quadro concettuale per l'integrazione dei servizi a più livelli nella federazione H2IOSC: piattaforme specifiche delle singole infrastrutture operano come ambienti di composizione,

[5] e-editiones, "TEI Publisher Documentation", 2024. Disponibile al link: https://teipublisher.com/exist/apps/tei-publisher/doc/documentation.xml
[6] TEI Publisher. Disponibile al link: https://teipublisher.com/index.html.

mentre il Marketplace agisce come orchestratore trasversale. L'implementazione tecnica del modello, basata su WSO2 Micro Integrator per l'esecuzione dei workflow e su endpoint OAI-PMH 2.0 per l'interoperabilità dei metadati (Sichera, Marras, Pasini 2026d), è documentata in una serie di relazione tecnica su Zenodo (Sichera, Marras, Pasini 2026a; 2026e). Nella fase conclusiva del progetto si intende consolidare la descrizione formale dei workflow (manifest e specifiche OpenAPI) con meccanismi di tracciamento della provenienza e ri-eseguibilità; estendere l'onboarding di provider e la federazione con marketplace esterni; valutare l'impatto delle funzionalità di orchestrazione e analytics sulle pratiche reali delle comunità. In questo senso, l'orchestrazione delle API non è solo un dettaglio implementativo, ma un punto di leva per rendere scalabili e riusabili servizi DH in ambienti federati sul modello di European Open Science Cloud.

RINGRAZIAMENTI

Il lavoro presentato è stato svolto nell'ambito del progetto H2IOSC – Humanities and Cultural Heritage Italian Open Science Cloud (PNRR, Missione 4 "Istruzione e ricerca", Componente 2 "Dalla ricerca all'impresa", Investimento 3.1 "Fondo per la realizzazione di un sistema integrato di infrastrutture di ricerca e innovazione", CUP B63C22000730005).

BIBLIOGRAFIA

Council of the European Union, General Secretariat, (2023). Draft Council conclusions on high-quality, transparent, open, trustworthy and equitable scholarly publishing, May 2023: https://data.consilium.europa.eu/doc/document/ST-8827-2023-INIT/en/pdf (cons. 21/04/2026).

Cristofaro, S., Fabiani, V., Marras, C., E. Pasini, Sichera P., M. Yu. (2024). "Infrastrutture di ricerca come strumenti di "interculturalità. Digitale". In: Me.Te. Digitali. Mediterraneo in rete tra testi e contesti, Proceedings XIII Convegno Annuale AIUCD, Catania 28- 30 maggio 2024, Università di Catania, a cura di Di Silvestro A., Spampinato, D., Quaderni di Umanistica Digitale, pp. 572-576. DOI 10.6092/unibo/amsacta/7927 (cons. 21/04/2026).

Degl'Innocenti, E., Pinna, F., Spadi A., Spinelli, F., (2025). "Supporting Executable Scientific Workflows in a Clustered Infrastructure: DARIAH-IT and H2IOSC", Transformations, A DARIAH Journal, vol. 1, pp. 1–19, 2025.

Kondam, A. (2024). Api Gateway Orchestration In Multi-Cloud Environments: Challenges And Solutions. International Journal of Creative Research Thoughts, vo. 12, 10 Oct. 2024, pp. 596-602. https://www.researchgate.net/publication/385345794_Api_Gateway_Orchestration_In_Multi-Cloud_Environments_Challenges_And_Solutions (cons. 21/04/2026)

Marras, C., Fabiani, V., Pasini, E., Reggiani, L., Sichera, P., Ongaro, P., & Rossi, M. (2025). "Modelli e tecnologie integrate e innovative per una cittadinanza digitale equa e sostenibile." In: S. Rebora, M. Rospocher, & S. Bazzaco (Eds.), Diversity, Equity, and Inclusion: Challenges and Opportunities for Digital Humanities in the Age of Artificial Intelligence. Proceedings del XIV Convegno Annuale AIUCD, Verona 11–13 giugno 2025, Università di Verona (pp. 538–543). DOI 10.6092/unibo/amsacta/8380 (cons. 21/04/2026).

Mittal, D., Mease, R., Kuner, T., Flor, H., Kuner, R., Andoh, J., (2023). "Data management strategy for a collaborative research center", *GigaScience*, Volume 12, 2023, giad049, https://doi.org/10.1093/gigascience/giad049 (cons. 21/04/2026).

Putta, N., Dave, A., Balasubramaniam, V., Prasad, M., Kumar, S. (2024). "Optimizing Enterprise API Development for Scalable Cloud Environments". Journal of Quantum Science and Technology (JQST) Vol.1, 3, Special Issue July-Sept 2024, pp. 229-246. https://www.researchgate.net/publication/387523300_Optimizing_Enterprise_API_Development_for_Scalable_Cloud_Environments (cons. 21/04/2026).

Rebora, S., Rospocher, M., Bazzaco, S., (a cura di) (2025) *Diversità, Equità e Inclusione: Sfide e Opportunità per l'Informatica Umanistica nell'Era dell'Intelligenza Artificiale*, Proceedings del XIV Convegno Annuale AIUCD2025. Verona: AIUCD, p. 663. In: Quaderni di Umanistica Digitale, DOI 10.6092/unibo/amsacta/8380 (cons. 21/04/2026).

Sichera, P., Marras, C., Pasini, E. (2024). Orchestrazione API per workflow applicativi nell'ambito delle Digital Humanities, Scientific-technical Report, 19 nov. 2024, Zenodo. https://doi.org/10.5281/zenodo.14187534 (cons. 21/04/2026).

Sichera, P., Marras, C., Pasini, E. (2026a). "A Multi-Infrastructure OAI-PMH Endpoint for Research Infrastructure Orchestration in the H2IOSC Project". Zenodo. https://doi.org/10.5281/zenodo.19353859 (cons. 21/04/2026).

Sichera, P., Marras, C., Pasini, E. (2026b). "Orchestrating Research Infrastructure Services through OAI-PMH: The OPERAS-IT Approach within H2IOSC". Zenodo. https://doi.org/10.5281/zenodo.19596981 (cons. 21/04/2026).

Sichera, P., Marras, C., Pasini, E. (2026c). "An Automated Workflow for Historical Manuscript Transcription and Publication: Orchestrating IIIF, eScriptorium, and TEI Publisher through WSO2 Micro Integrator". Zenodo. https://doi.org/10.5281/zenodo.19625779 (cons. 21/04/2026).

Sichera, P., Marras, C., Pasini, E. (2026d). "Building Executable Research Workflows for SSH Infrastructures: A WSO2-Based Orchestration Methodology within the H2IOSC Marketplace". Zenodo. https://doi.org/10.5281/zenodo.19625645 (cons. 21/04/2026).

Sichera, P., Marras, C., Pasini, E. (2026e). "From Data Model to Interoperable Metadata: Implementing the H2IOSC Resource Description Schema across a Multi-Infrastructure Pipeline". Zenodo. https://doi.org/10.5281/zenodo.19409795 (cons. 21/04/2026).

Sichera, P., Marras, C., Pasini, E. (2026f). "Un endpoint OAI-PMH condiviso per l'orchestrazione dei servizi delle infrastrutture di ricerca nel progetto H2IOSC". Zenodo. https://doi.org/10.5281/zenodo.19541925 (cons. 21/04/2026).

Sichera, P., Marras, C., Pasini, E. (2026g). "Orchestrazione dei servizi delle infrastrutture di ricerca tramite OAI-PMH: l'approccio OPERAS-IT nell'ambito di H2IOSC". Zenodo. https://doi.org/10.5281/zenodo.19597060 (cons. 21/04/2026).

Sichera, P., Marras, C., Pasini, E. (2026h). "Un workflow automatizzato per la trascrizione e la pubblicazione di manoscritti storici: orchestrazione di IIIF, eScriptorium e TEI Publisher tramite WSO2 Micro Integrator". Zenodo. https://doi.org/10.5281/zenodo.19625781 (cons. 21/04/2026).

Sichera, P., Marras, C., Pasini, E. (2026i). "Costruire workflow di ricerca eseguibili per le infrastrutture SSH: una metodologia di orchestrazione basata su WSO2 nell'ambito del Marketplace H2IOSC". Zenodo. https://doi.org/10.5281/zenodo.19625718 (cons. 21/04/2026).

Sichera, P., Marras, C., Pasini, E. (2026j). "Dal modello dati ai metadati interoperabili: implementazione dello schema di descrizione delle risorse H2IOSC in un workflow multi-infrastruttura". Zenodo. https://doi.org/10.5281/zenodo.19546279 (cons. 21/04/2026).

Sichera, P., Monachini, M., Quochi, V., Giampietro, N., Fabiani, V., Ottaviani, R., & Luzietti, R. B. (2025). "Synergies between CLARIN-IT and OPERAS-IT within H2IOSC: Monitoring Communities and Orchestrating Digital Service." In: CLARIN Annual Conference Proceedings 2025 (pp. 26–30). DOI 10.5281/zenodo.17357825 (cons. 21/04/2026).

Tomasi, F., F. Vitali, 2013, DH-CASE '13: Proceedings of the 1st International Workshop on Collaborative Annotations in Shared Environment: metadata, vocabularies and techniques in the Digital Humanities, NYC, ACM

[5/B] Patrimonio culturale: sistemi informativi integrati

Architettura e *Digital Humanities*. Conoscenza e valorizzazione del Patrimonio attraverso la costruzione di sistemi informativi integrati

DONATELLA RITA FIORINO[1], CATERINA GIANNATTASIO[1], ELISA PILIA[1], VALENTINA PINTUS[1] ,ANDREA PIRINU[1], MARCELLO SCHIRRU[1]

[1]UNIVERSITÀ DEGLI STUDI DI CAGLIARI

ABSTRACT (ITALIANO)

Sulla scorta di collaborazioni scientifiche interdisciplinari condotte negli ultimi anni presso la Scuola di Architettura di Cagliari, in particolare dalle Cattedre di Disegno e rilievo, di Storia dell'Architettura e di Restauro architettonico, il contributo illustra gli esiti di alcune esperienze di ricerca, svolte nell'ambito di progetti condivisi. Si tratta di percorsi di conoscenza, ricognizione, rappresentazione e definizione di proposte operative per la tutela, la valorizzazione e la conservazione del patrimonio costruito storico, alle diverse scale, di paesaggio, urbana e architettonica, che hanno sperimentato a diversi livelli l'utilizzo di sistemi informativi digitali come strumento operativo a supporto delle fasi di registrazione, di interpretazione critica e di correlazione dei dati storici, materiali e spaziali. L'ambiente digitale ha infatti facilitato il processo di intersezione transdisciplinare dei saperi umanistici e delle discipline tecniche, che condividono un comune impianto epistemologico, fondato sulla centralità della conoscenza come processo critico, interpretativo e storicamente centrato. Il contributo si inserisce in questo quadro teorico e metodologico, proponendo una riflessione sulle esperienze di ricerca sviluppate nel contesto mediterraneo, nazionale e sardo, caratterizzato da specifiche stratificazioni storiche e insediative. Si ritiene, infatti, che ripercorrendo brevemente tali esperienze si creino le condizioni per interrogare criticamente il ruolo del digitale nella costruzione della conoscenza del patrimonio, aprendo al contempo a prospettive di confronto e comparazione su scala nazionale e internazionale.
Parole chiave: Protocolli per la conoscenza, Catalogazione, classificazione tipologica, riconoscimento del patrimonio culturale, ontologie.

ABSTRACT (ENGLISH)

Architecture and Digital Humanities. Knowledge and Heritage Enhancement through the Development of Integrated Information Systems

Drawing on interdisciplinary scientific collaborations conducted in recent years at the Cagliari School of Architecture, particularly by the Chairs of Drawing and Surveying, History of Architecture, and Architectural Restoration, the paper illustrates the results of several research experiences conducted as part of shared projects. These include knowledge, reconnaissance, representation, and definition of operational proposals for the protection, enhancement, and conservation of historical built heritage, at various scales, landscape, urban, and architectural. These have experimented at multiple levels with the use of digital information systems as an operational tool to support the recording, critical interpretation, and correlation phases of historical, material, and spatial data. The digital environment has indeed facilitated the transdisciplinary intersection of humanistic knowledge and technical disciplines, which share a common epistemological framework, grounded on the centrality of knowledge as a critical, interpretative, and historically centered process. The contribution fits into this theoretical and methodological framework by proposing a reflection on research experiences conducted in the Mediterranean, national, and Sardinian contexts, characterized by specific historical and settlement stratifications. Indeed, it is believed that briefly retracing these experiences creates the conditions for critically questioning the role of digital in building heritage knowledge, while also opening perspectives for comparison on a national and international scale.
Keywords: Protocols of knowledge, Catalogue, Typological classification, Recognition of cultural heritage, Ontologies.

1. INTRODUZIONE

La rilevanza dei sistemi informativi integrati per le discipline del Disegno, della Storia dell'Architettura e del Restauro dell'Architettura è ormai un dato consolidato nella letteratura di settore. Nel perseguire il

comune obiettivo dello studio del patrimonio storico tangibile e intangibile, per la sua rappresentazione, tutela e salvaguardia, le tre discipline si avvalgono di fonti archivistiche, fotografiche e grafiche comuni, ricorrendo a banche dati informatiche come indispensabile strumento di lavoro (Metin, Rognoni 2024; Guidarelli, Giordano 2025; Guidarelli 2025). Tuttavia, la trasformazione digitale in atto, non si limita all'adozione di nuovi strumenti tecnologici, ma coinvolge in profondità le modalità di produzione, organizzazione e trasmissione della conoscenza umanistica. In questo quadro, si può affermare che le discipline sopra indicate rappresentino ambiti privilegiati di sperimentazione delle *Digital Humanities* (DH), in virtù della loro intrinseca vocazione interdisciplinare e della centralità che esse attribuiscono alla ricerca e interpretazione delle fonti come fondamento della conoscenza critica dell'architettura e più in generale del patrimonio costruito.

Le esperienze scientifiche condotte negli ultimi anni presso la Scuola di Architettura di Cagliari, nell'ambito delle corrispondenti Cattedre, offrono un osservatorio significativo per riflettere sul contributo che tali discipline possono apportare alle Digital Humanities, mettendo in dialogo saperi umanistici, tecnologie digitali e pratiche operative di analisi, rappresentazione e tutela del patrimonio alle diverse scale del paesaggio, della città e dell'architettura. Esse condividono infatti un comune impianto epistemologico fondato sulla centralità della conoscenza come processo critico, interpretativo e storicamente situato. Il Disegno, inteso come dispositivo conoscitivo e non come mera tecnica rappresentativa, svolge un ruolo fondamentale nella lettura analitica dell'opera architettonica, che si pone in continuità tanto con l'indagine storico-critica della Storia dell'Architettura, orientata alla ricostruzione dei processi di formazione e trasformazione del costruito, quanto con la comprensione profonda e documentata del bene che il Restauro presuppone come condizione preliminare e imprescindibile per ogni intervento di tutela e conservazione.

L'integrazione tra rilievo, rappresentazione, analisi storica e progetto di conservazione trova oggi nelle Digital Humanities un ambito di sviluppo particolarmente fertile. Le tecniche di rilievo digitale, la modellazione tridimensionale, i sistemi informativi territoriali, le piattaforme HBIM e gli archivi digitali consentono di organizzare e interrogare dati eterogenei – storici, cartografici, iconografici, materici e spaziali – restituendo una visione complessa e stratificata dell'architettura e del contesto urbano. In tale prospettiva, la rappresentazione digitale, non limitandosi alla visualizzazione delle informazioni, diviene parte integrante del processo conoscitivo, rendendo esplicite ipotesi interpretative, relazioni spaziali e temporalità spesso non immediatamente percepibili attraverso strumenti tradizionali.

All'interno delle Digital Humanities, la nozione stessa di conoscenza si ridefinisce come processo aperto, dinamico e relazionale. La formalizzazione digitale implica scelte teoriche e metodologiche che incidono sulla struttura del sapere, rendendo interrogabili non solo i dati, ma anche i criteri attraverso cui essi sono selezionati, organizzati e messi in relazione. In questo senso, le discipline suddette contribuiscono a riaffermare la centralità dell'interpretazione critica, evitando il rischio di una riduzione del digitale a mera accumulazione di informazioni. Il patrimonio architettonico viene così restituito come oggetto culturale complesso, nel quale dimensione materiale e immateriale, memoria storica e trasformazioni nel tempo, valori tangibili e intangibili concorrono a definire sistemi di significato articolati.

Un ulteriore aspetto rilevante riguarda il rapporto tra DH, tutela e accessibilità alla conoscenza. Le tecnologie digitali consentono, infatti, di costruire memorie strutturate del patrimonio, capaci di integrare documentazione storica, dati diagnostici e tracce degli interventi, in coerenza con i principi della conservazione critica e della reversibilità. Al contempo, esse favoriscono nuove forme di accessibilità, non limitate alla fruizione visiva dei manufatti, ma estese alla comprensione dei processi conoscitivi, delle interpretazioni e dei valori culturali associati. In questa prospettiva, le DH prendono forma come ambiente cognitivo e operativo entro cui la conoscenza del patrimonio architettonico può essere costruita, condivisa e trasmessa in modo più consapevole, inclusivo e interdisciplinare.

Il contributo si inserisce in questo quadro teorico e metodologico, proponendo una riflessione sulle esperienze di ricerca sviluppate nel contesto mediterraneo, nazionale e sardo, caratterizzato da specifiche stratificazioni storiche e insediative. Si ritiene, infatti, che ripercorrendo brevemente tali esperienze si creino le condizioni per interrogare criticamente il ruolo del digitale nella costruzione della conoscenza del patrimonio, aprendo al contempo a prospettive di confronto e comparazione su scala nazionale e internazionale.

2. DAL DATO ARCHIVISTICO ALLA CREAZIONE DEL DATABASE

La creazione di banche dati è un ausilio fondamentale per la ricerca storico-architettonica; quest'utile strumento supporta gli studi sulle epoche produttrici di fonti scritte, così come le fasi antiche, per le quali, venendo meno la disponibilità di manoscritti e documenti grafici, ci si avvale spesso di testimonianze e competenze vicine all'Archeologia. Ci soffermeremo, nel prosieguo, sul primo ambito per evidenziare le difficoltà insite nella catalogazione e messa a sistema delle banche dati per le discipline legate allo studio, alla salvaguardia e alla rappresentazione del patrimonio architettonico storico. Gli strumenti digitali consentono, infatti, di immagazzinare quantità di informazioni infinitamente superiori ai tradizionali supporti cartacei. L'assunto è valido per la quantità di spazio riservato alla trascrizione, ma anche per la velocità con cui il dato d'interesse può essere acquisito: in una seduta d'archivio, il quantitativo di informazioni acquisibili supera, di gran lunga, le potenzialità affidate al tradizionale appunto cartaceo. Emerge, tuttavia, subito il problema di incrociare i dati registrati. Una maestranza, ad esempio, potrebbe aver operato in più fabbriche o avere legami di parentela con altri affermati costruttori. In una architettura oggetto di studio potremmo ritrovare un particolare costruttivo, un decoro, un arredo sacro presente anche in altri edifici. L'unità archivistica contenente un elevato numero di carte potrebbe racchiudere documenti riguardanti altre opere o maestranze. Come consentire la visione e la fruizione simultanee di queste informazioni senza adeguati sistemi di catalogazione e di ricerca dati?
Strumenti digitali ordinari, come i software Excell o Access - Pacchetto Microsoft Office -, offrono efficaci funzionalità di registrazione e reperimento delle informazioni, anche se con evidenti limiti per la condivisione delle informazioni: le uniche modalità di lavoro comune passano, inevitabilmente, per la duplicazione del file o per la gestione promiscua attraverso la rete Internet (Google Drive, OneDrive e cloud computing similari). Nessun coordinamento è possibile tra studiosi ed utenti in merito all'ordinamento, alla codifica, alla natura materiale del dato; anche elaborando notazioni e griglie condivise all'interno del medesimo gruppo di lavoro, le informazioni tendono a disperdersi con il passaggio ad altri operatori.
Il problema è emerso in tutta la complessità in occasione della conferenza internazionale di Dublin, nel marzo del 1995, in cui è stato costruito uno standard internazionale noto come "Dublin Core Metadata Element Set", di descrizione e codifica dei dati. Allo stato attuale, il software più adatto alla creazione di banche dati secondo la notazione del Dublin Core è Omeka, la cui versione S, destinata alle istituzioni pubbliche, consente l'integrazione e il collegamento tra i contenuti informativi e incrementa le possibilità di dialogo e condivisione tra gli utenti. Riguardo alla Storia dell'Architettura e alle discipline affini, si segnala inoltre l'importanza georeferenziazione dei dati in ambiente GIS - Geographic Information Systems**.**

3. PATRIMONIO E TERRITORIO. DALLA STRATIGRAFIA MATERIALE AL LAYER DIGITALE

Nel campo del Restauro architettonico, il patrimonio culturale e il paesaggio sono tradizionalmente interpretati come sistemi complessi e stratificati, nei quali la materia costruita conserva le tracce dei processi (storici, tecnici e culturali, istituzionali, ecc.) che li hanno determinati, influenzandone la formazione e implicando continue trasformazioni. La nozione di stratigrafia costituisce, in questo senso, uno dei fondamenti epistemologici della disciplina, poiché consente di leggere l'architettura come risultato di una sedimentazione di fasi costruttive e di significati, riconoscendo la compresenza di permanenze e discontinuità. Tale approccio, consolidato nella scala del manufatto, si estende oggi in modo sempre più consapevole alla dimensione territoriale, nella quale edifici, infrastrutture e paesaggi costruiti si configurano come espressione di sistemi culturali articolati e interrelati.
Gli studi conoscitivi volti alla definizione di strategie di conservazione, restauro e tutela si fondano, dunque, imprescindibilmente su metodologie nelle quali i dati derivanti dalle analisi dirette (rilievo, analisi materica e del degrado, analisi delle tecniche costruttive e stratigrafia,) devono integrarsi con quanto emerge dallo studio delle fonti indirette (documenti d'archivio, iconografie, disegni e progetti, cartografie, ...). Seppure tale approccio possa avvalersi di protocolli consolidati, la crescente disponibilità di dati eterogenei e la necessità di metterli in relazione tra loro determinano nuove sfide metodologiche. La frammentazione delle informazioni, la difficoltà di confrontare scale diverse e la complessità della dimensione diacronica rendono evidente il limite degli strumenti tradizionali di documentazione e rappresentazione. In questo quadro, le Digital Humanities offrono un supporto decisivo alla ricerca nel

Restauro che, grazie all'insieme di tecnologie applicative messe a disposizione, offre un ambiente metodologico capace di strutturare la conoscenza in forma critica, relazionale e dinamica.
Il passaggio dalla stratigrafia materiale al *layer* digitale rappresenta, dunque, un momento chiave. Ogni *layer*, incorporando dati, fonti e interpretazioni, è uno spazio di verifica delle ipotesi di ricerca, nel quale la conoscenza non è data una volta per tutte, ma si costruisce attraverso l'interazione continua tra dati e interpretazione. In questo senso, la modellazione digitale e la costruzione di database configurano un sistema aperto, interrogabile e aggiornabile, coerente con l'idea di conoscenza come processo.
In questo quadro, l'esperienza maturata nel contesto sardo, e in particolare le attività di ricerca condotte dalla Cattedra di Restauro dell'Università degli Studi di Cagliari, rappresentano un caso significativo. La costruzione di database dedicati al patrimonio architettonico e territoriale regionale e nazionale, ha consentito di organizzare in modo sistematico informazioni provenienti da fonti d'archivio, rilievi, analisi diagnostiche e documentazione degli interventi, restituendo una lettura stratificata e comparabile dei manufatti e dei contesti.
In questo contesto di sistematizzazione delle ricerche, la ricognizione delle fonti bibliografiche e storico-archivistiche conservate presso i vari archivi presenti sul territorio regionale e nazionale, ha permesso di ricomporre un significativo quadro, spingendosi fino alla contemporaneità e alla ricomposizione del quadro degli interventi di restauro e di rifunzionalizzazione posti in atto e di rileggere criticamente le scelte funzionali e tecnico-operative compiute nella riconversione. Dall'analisi dei punti di forza e delle criticità di tali esperienze scaturiscono ulteriori riflessioni su possibili nuovi scenari di riuso sostenibile.
Tali esperienze dimostrano l'utilità di sistemi informativi territoriali quali dispositivi conoscitivi dinamici, nei quali la strutturazione dei dati riflette scelte teoriche e interpretative. La possibilità di interrogare le informazioni secondo criteri spaziali, temporali, formali e materico-costruttivi consente di mettere in relazione architetture e territori, evidenziando ricorrenze, differenze e processi di lunga durata.
La strutturazione digitale delle informazioni consente, inoltre, di conservare la memoria delle trasformazioni e degli interventi, supportando decisioni progettuali fondate su una comprensione approfondita del bene e del suo contesto. Al tempo stesso, la dimensione digitale favorisce una maggiore accessibilità e condivisibilità della conoscenza, ampliando il dialogo tra comunità scientifica, istituzioni e società. In questa prospettiva, il passaggio dalla stratigrafia materiale al layer digitale rappresenta una evoluzione coerente degli strumenti metodologici nel supporto alla ricerca scientifica sul patrimonio e sul paesaggio.

4. DALLA CONOSCENZA ALLA COMUNICAZIONE. STRUMENTI DI RILIEVO E DIVULGAZIONE DIGITALE

In aggiunta alle specificità già introdotte, le *Digital Humanities*, o Umanistica Digitale, vedono tra i campi di applicazione anche la gestione, l'analisi e la comunicazione della cultura tradizionale mediante l'integrazione di procedure computazionali, sistemi informatici e ambienti multimediali.
In tale prospettiva, il settore delle Scienze della Rappresentazione assume un ruolo centrale quale disciplina trasversale, in grado di contribuire in modo significativo all'applicazione delle tecnologie digitali per la conoscenza, la valorizzazione e la comunicazione del patrimonio culturale. La condivisione di tale patrimonio richiede infatti un linguaggio grafico capace di coniugare chiarezza e accessibilità per un pubblico ampio, anche non specialistico, con un'elevata densità informativa e valenza interpretativa.
Il Disegno, inteso come atto interpretativo, svolge nelle *Digital Humanities* un ruolo decisivo nel processo di trasformazione di dati e fonti in conoscenza visuale. Esso contribuisce alla formalizzazione visiva dell'informazione, alla traduzione di concetti complessi in modelli leggibili e accessibili e alla riflessione teorica sui processi di codifica, astrazione, convenzione e simbolizzazione. In quanto disciplina trasversale, il Disegno dialoga con l'informatica nella definizione di modelli, strutture e banche dati; interagisce con storici, archeologi e filologi nell'interpretazione critica delle fonti; media, infine, tra rigore scientifico e accessibilità comunicativa.
Il contributo delle Scienze della Rappresentazione si esplica nella selezione e nella costruzione di differenti forme di modellazione digitale del patrimonio culturale, articolate secondo diverse scale di lettura e specifici obiettivi conoscitivi e comunicativi. Rientrano in questo ambito il rilievo digitale, la modellazione bidimensionale e tridimensionale di architetture, città storiche e paesaggi, le ricostruzioni virtuali — non solo realistiche, ma critiche e documentate — e i sistemi HBIM, intesi come strumenti di conoscenza e non esclusivamente di gestione. In tale contesto, il modello digitale non è concepito come

semplice copia conforme del reale, ma come ipotesi interpretativa esplicita, verificabile e scientificamente fondata.
L'apporto disciplinare non si limita alla produzione di modelli e database digitali, ma si estende in modo significativo alla visualizzazione dei dati umanistici. Le competenze proprie del Disegno consentono di gestire gerarchie visive, scale, livelli di lettura, nonché aspetti legati a usabilità, accessibilità e interpretabilità, con l'obiettivo di rendere il dato non solo visibile, ma soprattutto comprensibile, interrogabile e criticabile.
L'accessibilità della conoscenza è ulteriormente rafforzata dallo sviluppo di forme ibride di rappresentazione, che integrano la precisione dei dati digitali con le capacità comunicative del disegno analogico, anche attraverso strategie di narrazione visiva (storytelling). A queste si aggiungono le mappe storiche interattive, atlanti digitali, storytelling visuale per musei, mostre e piattaforme web, nonché ambienti immersivi (VR/AR) fondati su un solido impianto scientifico. Le conoscenze e le modalità comunicative prodotte attraverso procedure informatiche e rielaborate mediante forme ibride di rappresentazione possono costituire il prodotto finale di allestimenti digitali, interfacce per musei, archivi e biblioteche, esperienze user-centered e strumenti di comunicazione visiva rivolti anche a fruitori non specialisti. In ambito archeologico, infine, i modelli digitali consentono di trasformare dati e fonti in conoscenza visuale, favorendo significativi avanzamenti nella ricerca. Ciò è reso possibile dal miglioramento delle tecniche di analisi del sito — attraverso il rilievo sistematico di grandi quantità di dati relativi sia ai reperti sia al territorio, anche nelle fasi precedenti alle operazioni di scavo — dall'impiego dei Sistemi Informativi Geografici per la produzione di cartografie tematiche e analisi complesse, nonché dallo sviluppo di applicazioni di modellazione tridimensionale utilizzate sia in ambito scientifico sia divulgativo.

5. SPERIMENTAZIONI LOCALI E PROSPETTIVE INTERISTITUZIONALI

La Scuola di Architettura di Cagliari, in collaborazione con le Soprintendenze, la Regione Autonoma della Sardegna, gli Enti locali e il Ministero della Difesa, porta avanti da oltre un ventennio diversi progetti di ricerca dedicati a mettere a fuoco la consistenza del patrimonio religioso, civile, difensivo e militare della Sardegna: luoghi, architetture e contesti molto diversi tra loro, ma legati da reciproche relazioni storiche e contemporanee. Le ricerche hanno messo in evidenza la ricca stratificazione del patrimonio sul territorio isolano e hanno permesso di costruire complessi sistemi informativi in cui sono localizzati e descritti conventi, chiese campestri, cattedrali, torri costiere, fortini, batterie, caserme, nosocomi, idroscali e aeroporti, etc.
La raccolta delle informazioni, organizzate su basi di dati georiferite, ha compreso le fonti documentarie, le tipologie e le tecniche costruttive, i caratteri morfologici e cronotipologici, lo stato di conservazione, la condizione giuridica e il livello di tutela ai sensi degli artt. 10 e 12 del Codice dei Beni Culturali e del Paesaggio, nonché la condizione d'uso storica e attuale. Infine, la ricognizione delle fonti storico-archivistiche conservate presso gli archivi del Ministero della Cultura, ha permesso di ricomporre il quadro degli interventi posti in atto sulle architetture e di rileggere criticamente le scelte funzionali e tecnico-operative compiute nel tempo. Dall'analisi dei punti di forza e delle criticità di tali esperienze scaturisce la riflessione su possibili nuovi scenari di intervento e di potenziale rifunzionalizzazione sostenibile delle architetture non più in uso.
Date la numerosità e l'eterogeneità delle informazioni raccolte, di notevole importanza è stata la gestione dei dati. Al fine di consentire un elevato grado di accessibilità alle informazioni acquisite, i dati raccolti in due progetti di ricerca (Tecniche murarie tradizionali: conoscenza per la conservazione e il miglioramento prestazionale. Progetto di ricerca di base finanziato dalla R.A.S. L.R. 7/2007, annualità 2013, e progetto ISOS – ISOle Sostenibili, Programma Italia-Francia Marittimo 2014-2020) sono confluiti all'interno del Sistema Informativo Territoriale Regionale (SITR), e più recentemente nel geoportale della Regione, raggiungendo così un importante obiettivo di diffusione e condivisione non solo del dato finale, ma dell'intero protocollo di indagine adottato.
Nell'ambito dell'accordo di collaborazione tra l'Università degli Studi di Cagliari e il Ministero della Difesa, è utile richiamare in questa sede anche la costruzione del "Portale delle infrastrutture per il volo in Italia", che contiene il censimento e la schedatura di tutti gli idroscali, aeroporti e campi volo storici italiani, costruiti sul territorio nazionale dai primi anni del Novecento al primo Dopoguerra. Il progetto, sviluppato nel 2023 in occasione del Centenario di fondazione della Regia Aeronautica, ha prodotto un articolato sistema informativo che rende accessibile, unitamente al dato geografico, anche interessanti e inediti

materiali storico-archivistici provenienti dall'Archivio dell'Ufficio Storico dello Stato Maggiore Aeronautica di Roma. Dalla medesima collaborazione e con il contributo dell'Istituto Italiano dei Castelli, è invece scaturito il contributo dell'Università di Cagliari all'Atlante Nazionale dei Castelli e il progetto Costellazione XX, finanziato dalla Fondazione di Sardegna e sostenuto dalle Soprintendenze di Cagliari e Sassari, nell'ambito del quale sono stati schedati e resi disponibili sulla piattaforma SIGEC/web del MIC e in un sito web dedicato, i dati relativi ai bunker, le batterie e gli altri sistemi di difesa della II Guerra Mondiale in Sardegna.
Anche il progetto di ricerca PRIN 2022, dal titolo Vulnerabilità e rischio dei centri storici fra spopolamento, abbandono e riuso intensivo si inserisce entro tale linea di ricerca più ampia. Contenuti e strumenti per il monitoraggio e la gestione dell'edilizia storica diffusa, vede il coinvolgimento delle sedi universitarie di Roma, Cagliari, Milano, Parma, Catania, con l'obiettivo di indagare le trasformazioni in corso nel patrimonio storico diffuso delle città italiane, sempre più segnato da spopolamento, abbandono e interventi di rifunzionalizzazione intensiva, e dunque di definire contenuti e strumenti innovativi per il monitoraggio, la gestione e la valorizzazione di questo patrimonio, in una prospettiva che tenga insieme tutela, sicurezza, sostenibilità e ricadute sociali.
Un'ulteriore sperimentazione ha infine riguardato il censimento dei luoghi di produzione della calce in Sardegna, condotto nell'ambito del progetto biennale finanziato dalla Fondazione di Sardegna "Officinae et
Tabernae" - annualità 2022, in collaborazione con il Dipartimento di Lettere, Lingue, Beni Culturali, che ha permesso, attraverso le schede digitali, di sistematizzare le conoscenze sul tema, che si presentavano piuttosto eterogenee sia sotto il profilo della natura che della consistenza delle informazioni. Per ciascuna emergenza archeologica, è stato prodotto un tracciato schedografico digitale che raccoglie informazioni geografiche, morfologiche, tipologiche, cronologiche e sulle tecniche costruttive. Una sezione è inoltre dedicata allo stato di conservazione e al livello di accessibilità e fruizione dei manufatti, nell'ottica dello sviluppo di un potenziale sistema di monitoraggio, gestione e valorizzazione di questo particolare patrimonio culturale.
A queste sperimentazioni, incardinate in programmi di ricerca strutturati, si devono aggiungere i numerosi censimenti tematici georeferenziati prodotti nell'ambito di ricerche dottorali, tesi di laurea e laboratori didattici.
L'applicazione delle Digital Humanities alle ricerche richiamate ha consentito di mettere in evidenza la varietà e la vulnerabilità del patrimonio culturale dell'Isola, mostrando contestualmente l'elevato grado di stratificazione nei territori investigati, la limitata disponibilità di studi specialistici, il diffuso stato di abbandono di particolari categorie di beni e la mancanza di programmi coordinati di restauro conservativo e di riconversione funzionale.
Per questo motivo, si ritiene che tali strumenti debbano essere ulteriormente implementati e condivisi con i portatori di interesse al fine di pervenire a quadri conoscitivi approfonditi e condivisi, orientati non solo alla tutela dei singoli manufatti, ma anche alla promozione di progetti di ricerca 'esplorativi', atti a individuare e testare vocazioni territoriali e scenari di riqualificazione basati sui molteplici valori storici - materiali e immateriali - di cui tali beni sono testimonianza e sul complesso sistema di relazioni paesaggistiche e territoriali di cui sono parte integrante.

BIBLIOGRAFIA

Acierno, M., & Fiorani, D. (2025). Conservation process model: An ontology for conservation in architecture. Sapienza University Press. (Studi e Ricerche).

Biagetti, M.T. (2010). Le ontologie come strumenti per l'organizzazione della conoscenza in rete. AIDA Informazioni, 12(2), 9–32.

Colavitti, A.M., Fiorino. D.R., Serra, S., Usai, A. (2019). Sustainable development of small Mediterranean islands: bioregional strategies for the valorisation of military heritage and the islands territory, in Proceedings of the International Conference on Changing Cities IV Spatial, Design, Landscape & Socio-economic Dimensions, 2019, pp.1500-1511.

Damiani, G., & Fiorino, D.R. (2018). Military landscapes: A future for military heritage. Skira.

Ercolino, M.G. (2020). Archeologia, architettura e restauro tra XIX e XXI secolo: Dallo scavo alle digital humanities. In M.I. Pascariello & A. Veropalumbo (Eds.), La città palinsesto. Tracce, sguardi e narrazioni sulla complessità dei contesti urbani storici (Vol. 2, pp. 61–69). FedOA – Federico II University Press.

Fiorani, D. (2019). Il futuro dei centri storici: Digitalizzazione e strategia conservativa. Quasar.

Fiorani, D., Acierno, M., Cutarelli, S., & Donatelli, A. (2020). Transformation of tools and conservation of architecture. Építés – Építészettudomány, 49(1–2), 97–131.

Fiorino, D.R., & Pintus, M. (2015). Verso un atlante dei sistemi difensivi della Sardegna. Giannini Editore.

Fiorino, D.R., & Pintus, V. (2015). Il sistema informativo per la conoscenza integrata e la costruzione dell'atlante tematico. In D.R. Fiorino & M. Pintus (Eds.), Verso un atlante dei sistemi difensivi della Sardegna (pp. 423–439). Giannini Editore.

Fiorino, D.R., Giannattasio, C., Grillo, S.M., & Pintus, V. (2015). Fortificazioni e cronologie: Protocolli conoscitivi per la conservazione. In D.R. Fiorino & M. Pintus (Eds.), Verso un atlante dei sistemi difensivi della Sardegna (pp. 128–172). Giannini Editore.

Fiorino, D.R., Giannattasio, C., Pilia, E., & Pintus, V. (Eds.). (2026). Vulnerabilità e rischio dei centri storici fra spopolamento, abbandono e riuso intensivo: Contenuti e strumenti per il monitoraggio e la gestione dell'edilizia storica diffusa. Unica Press. (In corso di stampa; Collana "Continuità. Tutela, conservazione, restauro, valorizzazione").

Fiorino, D.R., Vargiu, M. (2023). Tra Terra e cielo. Le infrastrutture aeronautiche in Italia, dalle prime esperienze al Secondo Dopoguerra. Collana Centenario Aeronautica Militare. Edizioni Rivista Aeronautica.

Giannattasio, C. (Ed.). (2019). Arte muraria tradizionale in Sardegna: Conoscenza, conservazione, miglioramento. Gangemi Editore. (Paesaggi in trasformazione).

Giannattasio, C., Grillo, S.M., & Murru, S. (2017). Il sistema di torri costiere della Sardegna: Forma, materiali, tecniche murarie / The Sardinian coastal tower system: Form, materials and masonry techniques. L'Erma di Bretschneider.

Giannattasio, C., Grillo, S.M., Pintus, V., & Pirisino, M. S. (2022). Castelli medievali in Sardegna (XI–XV secolo) / Medieval castles in Sardinia (11th–15th cent.). L'Erma di Bretschneider.

Guidarelli, G., & Giordano, A. (2025). Storia dell'architettura e della città, digital humanities e rappresentazione digitale: Il progetto di ricerca sull'isola di San Fantin. Ateneo Veneto, 23, 57–62.

Masetti, C., & Spadafora, G. (Eds.). (2023). Digital humanities, patrimonio culturale e applicazioni geostoriche: Atti del convegno (Roma, 5–7 dicembre 2022). Labgeo Caraci.

Metin, A., & Rognoni, F. (2024). Storia dell'architettura e digital humanities per la conoscenza e la conservazione del patrimonio costruito: Un'esperienza italiana. Magazén, 5(1), 9–36.

Pintus, V., & Pili, D. (2023). Archiviazione, gestione e fruizione: Un sistema informativo per la riscoperta del patrimonio carcerario. In C. Giannattasio & G. B. Cocco (Eds.), Historical prisons. Studi e proposte per il riuso del patrimonio carcerario dismesso in Sardegna (ArcHistoR Extra, 11/2023, pp. 178–193).

Pirinu, A., Paba, N., & Sanna, G. (2024). Tecniche integrate per la preservazione e la comunicazione del patrimonio materiale e immateriale: La Sagra di San Sisinnio a Villacidro (Sardegna, Italia). In F. Bergamo, A. Calandriello, M. Ciammaichella, I. Friso, F. Gay, G. Liva, & C. Monteleone (Eds.), Misura/Dismisura. Measure/Out of Measure. Transitions: Proceedings of the 45th International Conference of Representation Disciplines Teachers (pp. 3533–3542). FrancoAngeli.

Sanna, G., & Pirinu, A. (2025). Lettura diacronica tramite il fumetto: La Torre di Carcangiolas nel Poetto di Quartu, Sardegna. In GRAF_I. La rappresentazione per la narrazione della cultura immateriale (pp. 46-47). Publica.

Vacca, G., Pili, D., Fiorino, D.R., & Pintus, V. (2017). A WebGIS for the knowledge and conservation of the historical wall structures of the 13th–18th centuries. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLII-5/W1, 551–556.

Tra testo e architettura digitali per lo studio delle trasformazioni storico-artistiche dei monasteri benedettini

GIANMARIO GUIDARELLI[1], PAOLO BORIN[2], SONIA CAVICCHIOLI[3], CHANTAL PIVETTA[4], RENATO CANEARO[5],

[1]UNIVERSITÀ DI PADOVA, GIANMARIO.GUIDARELLI@UNIPD.IT
[2]UNIVERSITÀ DI BRESCIA, PAOLO.BORIN@UNIBS.IT
[3]UNIVERSITÀ DI BOLOGNA, SONIA.CAVICCHIOLI@UNIBO.IT
[4]LUND UNIVERSITY, CHANTAL.PIVETTA@ROM.LU.SE
[5]SILENTWAVE, RENATO.CAENARO@SILENTWAVE.EU

ABSTRACT (ITALIANO)

Il contributo presenta gli esiti di una ricerca focalizzata sull'analisi della cultura architettonica e artistica della Congregazione Benedettina Cassinese tra il XV e il XVIII secolo. L'obiettivo centrale è la verifica dell'ipotesi storiografica di una cultura progettuale e artistica condivisa, favorita dalla riforma della Regola di San Benedetto e dalle normative imposte dai Capitoli Generali. La metodologia proposta si fonda su un approccio interdisciplinare che integra il Building Information Modeling (BIM) con la codifica semantica dei testi (TEI) attraverso un'infrastruttura digitale coordinata.
Il sistema si articola su tre pilastri: un database relazionale (RDB), modelli informativi BIM derivati da rilievi laser scanner e fotogrammetrici, e trascrizioni TEI di fonti documentarie. L'integrazione avviene tramite l'implementazione di dizionari controllati e standard IFC per la semantizzazione degli elementi architettonici storici, mentre un modello ispirato al CIDOC-CRM permette di strutturare i dati archivistici come reti di relazioni tra persone, luoghi e processi costruttivi.
L'interoperabilità, garantita da API REST e GraphQL, consente di interrogare il sistema per rispondere a quesiti complessi, come l'analisi della densità dei cantieri o il ruolo dei committenti nelle trasformazioni abbaziali. I risultati dimostrano come la creazione di una Single Source of Information (SSoI) favorisca nuove prospettive di indagine comparativa, rendendo i dati navigabili e processabili per la storia dell'arte e dell'architettura.
Parole chiave: GraphQL, TEI, BIM, architettura, benedettina

ABSTRACT (ENGLISH)

The paper presents the results of a research project focused on the analysis of the architectural and artistic culture of the Cassinese Benedictine Congregation between the 15th and 18th centuries. The central objective is to verify the historiographical hypothesis of a shared design and artistic culture, fostered by the reform of the Rule of Saint Benedict and the regulations imposed by the General Chapters. The proposed methodology is based on an interdisciplinary approach that integrates Building Information Modeling (BIM) with the semantic encoding of texts (TEI) through a coordinated digital infrastructure.
The system is built upon three pillars: a relational database (RDB), BIM information models derived from laser scanning and photogrammetric surveys, and TEI transcriptions of documentary sources. Integration is achieved through the implementation of controlled dictionaries and IFC standards for the semantization of historical architectural elements, while a model inspired by CIDOC-CRM allows for the structuring of archival data as networks of relationships between people, places, and construction processes.
Interoperability, guaranteed by REST and GraphQL APIs, enables the system to be queried to address complex questions, such as the analysis of construction site density or the role of patrons in abbatial transformations. The results demonstrate how the creation of a Single Source of Information (SSoI) fosters new perspectives for comparative investigation, making data navigable and processable for the fields of art and architectural history.
Bridging text and architecture: a methodological digital approach to the artistic transformations of Benedictine monasteries.
Keywords: GraphQL, TEI, BIM, architecture, benedictine

1. INTRODUZIONE

La proposta di paper si inserisce nell'ambito del progetto di ricerca PRIN 2022 "CoenoBI(u)M. Art and architecture of the Cassinese Benedictine Congregation (XV-XVIII centuries): digital and spatial analysis strategies through BIM models." L'obiettivo del progetto è lo studio della cultura architettonica e artistica della Congregazione Benedettina Cassinese dalla sua fondazione fino alla fine del XVIII secolo, in un'ottica comparativa e grazie all'uso innovativo e sperimentale della metodologia del Building Information Modeling (BIM). La Riforma della Regola di S. Benedetto, istituita da Ludovico Barbo nel 1419, ha introdotto un tipo nuovo di monastero ed una cultura artistica e architettonica condivisa in molte abbazie della Congregazione, grazie anche a una normativa imposta dai Capitoli Generali. Questa ipotesi storiografica è stata formulata in decenni di studi (Cavicchioli, 2017; Cavicchioli & Vandelli, 2017; Guidarelli, 2022).
Per poter verificare questa ipotesi storiografica un gruppo di ricerca interdisciplinare ha predisposto una serie di strumenti, coordinati in una piattaforma digitale, che permetta di integrare in modo strutturato un corpus documentario vasto ed eterogeneo con la consistenza fisica dei manufatti (edifici ed oggetti artistici) rielaborati in modelli informativi BIM (Building Information Modeling) (Giordano et al., 2018; Murphy et al., 2013).
La necessità di condurre un'analisi di singoli casi-studio messi a sistema in un'ottica comparativa, infatti, ha portato alla costruzione di un ambiente di lavoro versatile e collaborativo che permetta un'indagine coordinata e multidisciplinare, a partire da tre casi studio: le abbazie di San Pietro a Modena, Santa Giustina a Padova e Praglia. Infatti, la metodologia BIM, che prevede la realizzazione del modello digitale di un edificio nel suo stato di fatto, è implementabile con tutti i dati materici, costruttivi e geometrici della sua architettura ma anche dei manufatti in essa contenuti, soprattutto opere d'arte come affreschi, pale d'altare e statue. Per potere integrare la catalogazione dei dati inerenti architetture, opere d'arte (nelle rispettive campagne di realizzazione) e persone coinvolte (artisti e committenti) con le fonti primarie è stato adottato in sistema integrato di strumenti: un database relazionale (RDB) si pone quale struttura di raccordo tra il modello BIM di ciascun monastero (realizzato in seguito a un rilievo strumentale) e la trascrizione TEI dei documenti a supporto dello studio architettonico e artistico. BIM e TEI si presentano infatti come le principali fonti informative per un progetto di questo tipo: il primo per la parte geometrica viene utilizzato come un database spazialmente orientato, capace di gestire la stratificazione temporale e semantica dell'architettura storica, il secondo per supportare i documenti testuali.
In questo paper ci proponiamo di verificare come questo sistema, interrogato attraverso delle interrogazioni strutturate, permetta di rispondere alla domanda storiografica di partenza

2. STRUMENTI DIGITALI: RILIEVO E MODELLAZIONE BIM

La modellazione degli spazi architettonici dei monasteri è eseguita attraverso un processo che parte dal rilievo digitale dell'intera estensione, allo stato attuale, per poi ricostruire le fasi precedenti andando a ritroso nel tempo, all'interno di un unico modello.
Un nodo centrale della trattazione riguarda quindi le strategie di acquisizione metrica, che nel progetto non seguono un'intensità uniforme, bensì una logica differenziata basata sull'importanza storico-artistica degli spazi. Per gli ambienti di massimo rilievo, quali chiese e capitoli, è stato impiegato il laser scanning statico, garantendo nuvole di punti con precisione millimetrica indispensabili per l'analisi delle irregolarità costruttive e delle proporzioni (Figura 1). Al contrario, per le zone connettive e di servizio, la tecnologia SLAM (Simultaneous Localization and Mapping) ha permesso una documentazione rapida e continua, ottimizzando i tempi di rilievo senza sacrificare la comprensione della distribuzione planimetrica. Questo apparato è stato ulteriormente integrato da rilievi fotogrammetrici da drone per le coperture — spesso trascurate dalla storiografia tradizionale — e da fotogrammetria terrestre ad alta risoluzione per gli apparati decorativi e scultorei, permettendo la creazione di modelli 3D testurizzati che conservano la fedeltà cromatica e materica necessaria per l'analisi del restauro e della percezione liturgica originale.
Il passaggio dalla nuvola di punti al modello BIM strutturato richiede una scomposizione intellettuale dell'architettura in componenti tipologici funzionali. La modellazione non è intesa come ricalco geometrico, ma come ricostruzione della logica costruttiva storica: le volte, ad esempio, sono state modellate attraverso famiglie parametriche basate sui trattati rinascimentali, rendendo il modello un documento filologicamente attendibile e confrontabile su larga scala.

L'innovazione più significativa risiede però nella semantizzazione del modello attraverso l'implementazione di dizionari di dati che permettono di collegare semanticamente i dati tra fonti di informazioni differenti. Poiché le logiche dell'industria delle costruzioni risultano inadeguate a descrivere l'architettura storica e monastica, la ricerca ha sviluppato dizionari controllati specifici. Attraverso il bSDD (buildingSmart Data Dictionary), ogni entità geometrica è stata arricchita con attributi storici, riferimenti archivistici e classificazioni tipologiche (come "chiostro", "refettorio" o elementi specifici come "affresco") (Alexiev et al., 2023). Questo processo garantisce che il dato non sia solo leggibile dall'uomo, ma algoritmicamente processabile e interoperabile. L'integrazione del bSDD all'interno dello standard IFC (Industry Foundation Classes) permette di mappare concetti storici precisi su entità geometriche, trasformando un semplice muro in un elemento informativo che porta con sé la propria cronologia e le fonti documentarie associate (Bernardello & Bernd, 2023).

3. STRUMENTI DIGITALI: TEI

L'Encoding Model sviluppato per CoenoBIuM nasce dall'esigenza di trasformare le fonti documentarie dei complessi cassinesi in dati strutturati, interoperabili e integrabili con gli altri sistemi della piattaforma (Ciotti & Tomasi, 2016). In questo contesto, la scelta della TEI P5 non risponde a finalità editoriali tradizionali, ma alla necessità di adottare un linguaggio capace di funzionare come strumento di modellazione concettuale. La TEI, grazie alla sua modularità, alla diffusione internazionale e alla possibilità di definire personalizzazioni controllate, consente di rappresentare entità, ruoli, processi e relazioni in modo formalizzato, mantenendo al contempo la piena aderenza agli standard della comunità scientifica (Burnard, 2014).
L'approccio adottato si colloca nel quadro della digital history, dove la digitalizzazione non è una mera trascrizione, ma un'operazione interpretativa che estrae dal testo strutture informative, le normalizza e le rende interrogabili. In CoenoBIuM, ciò avviene attraverso un modello event-based, ispirato al paradigma del CIDOC-CRM: gli eventi diventano nodi semantici che collegano persone, materiali, luoghi, oggetti e sequenze temporali. Questa prospettiva permette di passare dalla narrazione documentaria a una rete di relazioni computazionalmente analizzabili, facilitando l'integrazione con il database relazionale, con EVT (Edition Visualization Technology), l'Atlante e i modelli BIM.
La ricerca filologica assume così una nuova funzione: non solo interpretare i testi e prepararne l'edizione digitale, ma produrre dati che alimentano l'intera infrastruttura di ricerca, arricchendo e collegando ogni disciplina coinvolta nel progetto. Le liste TEI sono generate in maniera dinamica dal database e collegate alla piattaforma utilizzata per la trascrizione dei testi in XML/TEI, in questo caso è stata scelta Digital Philology for Dummies (DPhD). Le liste identificate nel processo di codifica da parte dei trascrittori all'interno di DPhD vengono automaticamente aggiornate dal DBR attraverso l'uso di API (Figura 2). Questo processo garantisce coerenza semantica, unicità degli identificatori e allineamento continuo tra corpus e RDB. La piattaforma DPhD rende operativa questa metodologia, permettendo a con limitate competenze XML/TEI di applicare il modello attraverso un ambiente controllato, con markup assistito, validazione continua, integrazione automatica con i sistemi del progetto e possibilità di revisione e tracciabilità di ogni operazione.
In questo modo, il processo di encoding diventa un processo collettivo e replicabile che trasforma le fonti in un patrimonio di dati interoperabili, destinato a sostenere analisi storiche, prosopografiche, spaziali e materiali su scala crescente.

4. STRUMENTI DIGITALI: DATABASE RELAZIONALE

L'infrastruttura digitale del progetto CoenoBIuM è stata concepita come un ecosistema integrato capace di sostenere la complessità dei dati storico-documentari, architettonici e filologici relativi al patrimonio benedettino. Il nucleo dell'intero sistema è il modello dati relazionale, progettato per rappresentare in modo coerente le entità fondamentali – monasteri, cronologie, persone, ruoli, fonti documentarie e iconografie – e per collegarle attraverso relazioni che integrano dimensioni spaziali, temporali e agentive. Il monastero funge da nodo centrale, mentre l'adozione di principi ispirati al CIDOC CRM garantisce interoperabilità semantica e sostenibilità a lungo termine (Chryssoula Bekiari et al., 2025).
Il modello è stato implementato in SQL Server tramite Entity Framework Core con metodologia Code-First, scelta che consente un'evoluzione controllata dello schema attraverso migrazioni versionate e ambienti separati di sviluppo e produzione (AndriySvyryd, 2026). Particolare attenzione è stata dedicata all'integrità referenziale, all'uso di chiavi composite per le relazioni molti-a-molti e all'inserimento

sistematico di metadati (CreatorId, RevisorId, ecc.), che assicurano tracciabilità e responsabilità del processo di costruzione dei dati.
Per rendere il database un'infrastruttura realmente interoperabile, sono stati sviluppati servizi di consultazione basati su REST e soprattutto su GraphQL. Mentre le API REST forniscono accessi strutturati e prevedibili, GraphQL permette interrogazioni flessibili e profondamente relazionali, consentendo ai client di selezionare esattamente i campi necessari e di attraversare in un'unica richiesta catene complesse di entità (ad esempio monastero > cronologie > persone > fonti). Questa capacità di modellare query ad alta granularità trasforma il database in un hub informativo capace di alimentare strumenti eterogenei senza duplicare logiche applicative, lasciando libertà di azione ai partecipanti.
L'interoperabilità è dimostrata dalle integrazioni operative: la piattaforma di filologia digitale DPhD utilizza le API per inserire identificativi univoci del database nelle trascrizioni TEI, creando collegamenti semantici tra testi e dati strutturati; i modelli BIM in Autodesk Revit interrogano GraphQL per associare elementi architettonici a cronologie, persone e fonti; infine, un Ambiente di Condivisione dei Dati (ACDat) openBIM coordina il lavoro collaborativo sugli oggetti tridimensionali.
Nel loro insieme, il modello relazionale e i servizi di interoperabilità costituiscono la spina dorsale metodologica di CoenoBIuM: un'infrastruttura che non si limita a conservare dati, ma li rende navigabili, verificabili e riutilizzabili, favorendo nuove forme di ricerca interdisciplinare sul patrimonio benedettino.

5. RISULTATI

Questa ricerca ha portato alla sperimentazione operativa di una serie di strumenti e soprattutto di flussi di lavoro collaborativi, testando il ruolo degli strumenti digitali per lo studio delle trasformazioni dei monasteri benedettini. Si tratta di un flusso di lavoro spesso complesso, che vede da un lato la digitalizzazione dell'ambiente costruito, con riferimento alle pratiche BIM e in particolare allo standard aperto IFC, dall'altro la digitalizzazione dei documenti storici, con riferimento al processo che coinvolge la TEI. In entrambi i casi è necessario descrivere specifiche personalizzazioni per gli scopi del progetto: nel primo caso la modellazione delle componenti storiche precise, ad esempio statue e volte, che difficilmente rientrano nella pratica contemporanea; nel secondo caso sono risultate necessarie delle modellazioni specifiche per rendere possibile il collegamento tra cantiere, persone ed elementi costruttivi.
Il database relazionale è il punto di unione tra tutti gli elementi, e consente in prima istanza la redazione di schede riepilogative del monastero, con un risultato solo apparentemente pari ad una schedatura tradizionale. In realtà permette il collegamento e di seguito il trasferimento informativo per molteplici utilizzi legati al progetto.
Un primo esempio è costituito dal collegamento tra le zone spaziali di interesse storico e artistico per il gruppo di lavoro con le loro copie virtuali nel modello BIM. Un apposito script nell'ambiente di modellazione BIM permette, attraverso una query di GraphQL di ricavare tutte le informazioni ad esse legate e di connetterle in modo univoco, scrivendo gli ID nel database all'interno di parametri e opportune relazioni nel file IFC. Lo stesso esempio può essere fatto con i cantieri individuati all'interno del RDB e successivamente create all'interno dello strumento di modellazione, che così permette ai modellatori tridimensionali una maggiore accuratezza nella nomenclatura ma anche un approccio votato alla Single Source of Information (SSoI).
Un secondo esempio riguarda il collegamento tra RDB e ambiente di trascrizione DPhD: in questo caso opportune API permettono il collegamento con cantieri, persone e i ruoli che assumono all'interno del cantiere, oltre che agli oggetti artistici.
Un terzo esempio riguarda invece la possibilità, attraverso query GraphQL, di importare i dati all'interno di strumenti di Business intelligence (BI), in modo che questi possano essere esplorati anche da non addetti ai lavori, e controllati in modo massivo. I dati presenti nel RDB possono quindi essere opportunamente trasformativi per permettere a tutti i partecipanti la loro visualizzazione, e quindi l'indagine. A questo riferimento sono state scelte due questioni di ricerca specifiche, proposte dagli esperti di storia dell'arte e dell'architettura. La prima domanda riguarda la densità delle zone spaziali legate ai cantieri per ogni monastero. A questo fine sono stati usati alcuni grafici tipici della visualizzazione dei dati (treemap e gantt), che mostrano una distribuzione uniforme delle localizzazioni spaziali dei cantieri nei diversi monasteri, senza produrre specifici picchi informativi (Figura 3).
La seconda domanda riguarda invece il ruolo degli abati nelle trasformazioni dei singoli cantieri. In questo caso è possibile filtrare i dati per monastero e intervenire con la visualizzazione per mostrare quanti cantieri sono iniziati con un determinato abate nel ruolo di committente. Il regesto dei cantieri, così

organizzato, permette di rintracciare pattern tra i singoli monasteri e la congregazione (Figura 4). Questo secondo esempio è paradigmatico a supportare altre domande di ricerca specifiche, quali gli artisti coinvolti o i materiali utilizzati nella produzione architettonica.



Figura 1. Nuvola di punti del monastero di Praglia

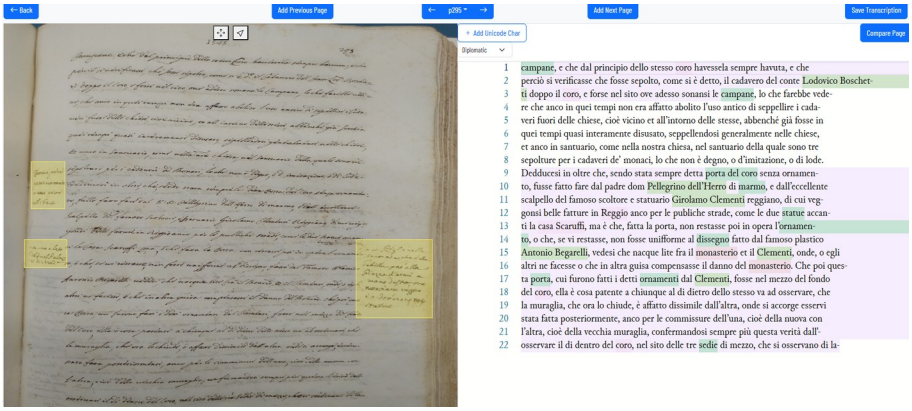


Figura 2. Trascrizione TEI del volume *Informatione dell'Archivio del Monistero di San Pietro di* M. A. Lazarelli in ambiente DPhD (S. Gottardi e C. Pivetta)

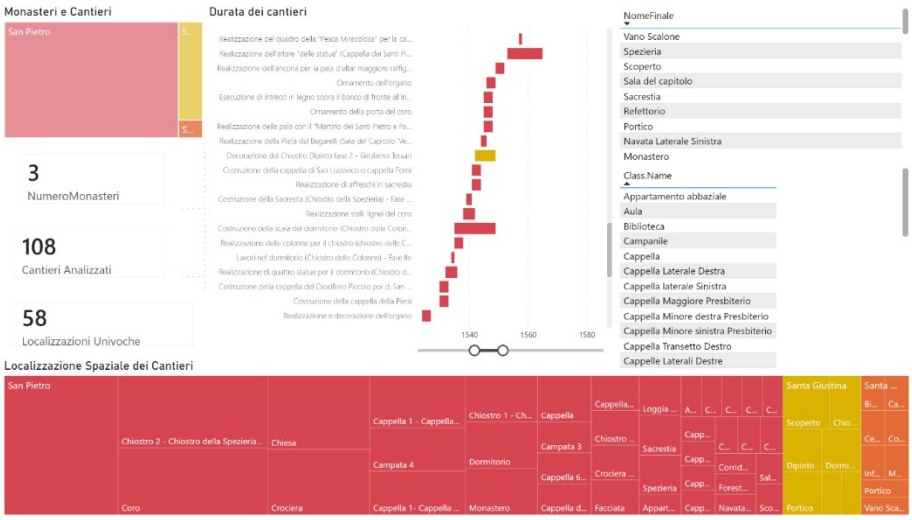


Figura 3. Dashboard di visualizzazione della relazione tra cantieri e zone spaziali (P. Borin)

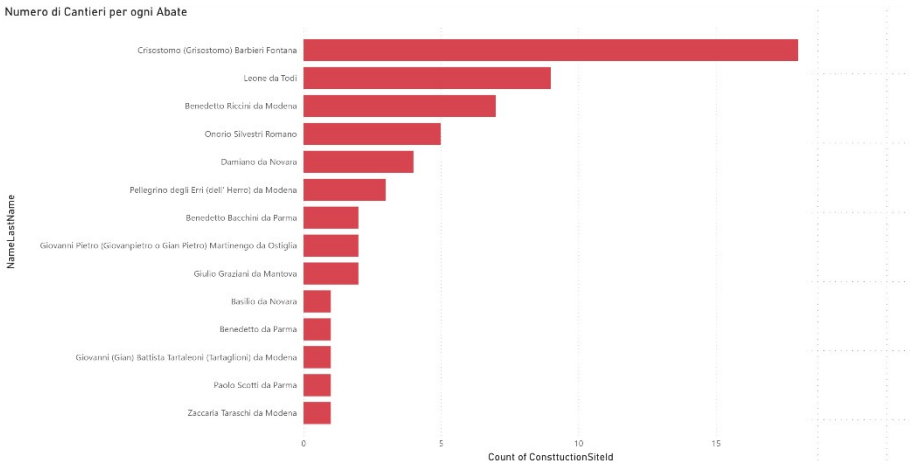


Figura 4. Analisi dei committenti nei cantieri di San Pietro a Modena (P.Borin)

RICONOSCIMENTI

Questo studio è stato realizzato nell'ambito del progetto PRIN 2022 "Fondo per il Programma Nazionale di Ricerca e Progetti di Rilevante Interesse Nazionale (PRIN)", Progetto: CoenoBI(u)M. Arte e architettura della Congregazione Benedettina Cassinese (XV-XVIII secolo): strategie di analisi digitale e spaziale attraverso modelli BIM. Codice: 20228Z4LX3. Codice CUP C53D23006740006.
1 – Guidarelli-Cavicchioli, 2- Borin, 3 – Pivetta, 4 – Caenaro, 5 – Borin

BIBLIOGRAFIA

Alexiev, V., Radkov, M., & Keberle, N. (2023). Semantic bSDD: Improving the GraphQL, JSON and RDF Representations of buildingSmart Data Dictionary. *CEUR Workshop Proc.*, *3633*, 85–97.

AndriySvyryd. (2026, gennaio 12). *Microsoft SQL Server Database Provider—EF Core*. https://learn.microsoft.com/en-us/ef/core/providers/sql-server/

Bernardello, R., & Bernd, D. (2023). *Interoperability an introduction to IFC and buildingSmart standards, integrating infrastructure modeling*. EPFL Press. https://www.research.unipd.it/handle/11577/3546443

Burnard, L. (2014). What is the Text Encoding Initiative? : How to add intelligent markup to digital resources. In *What is the Text Encoding Initiative? : How to add intelligent markup to digital resources*. OpenEdition Press. https://books.openedition.org/oep/426

Cavicchioli, S. (2017). The fresco decoration of the church of the Cassinese monastery of San Sisto in Piacenza, Italy: Meaning and function. *Equipamentos Monásticos e prática espiritual, coord. Maria de Lurdes Craveiro, Lisboa, Bens Culturais da Igreja.*

Cavicchioli, S., & Vandelli, V. (2017). *Benedettini in Europa. Cultura e committenze, restauri e nuove funzioni*. Franco Cosimo Panini.

Chryssoula Bekiari, George Bruseker, Erin Canning, Martin Doerr, Philippe Michon, Christian-Emil Ore, Stephen Stead, & Athanasios Velios (A c. Di). (2025). *Definition of the CIDOC Conceptual Reference Model*. https://cidoc-crm.org/Version/version-7.3.1

Ciotti, F., & Tomasi, F. (2016). Formal Ontologies, Linked Data, and TEI Semantics. *Journal of the Text Encoding Initiative*, (Issue 9). https://doi.org/10.4000/jtei.1480

Giordano, A., Friso, I., Borin, P., Monteleone, C., & Panarotto, F. (2018). Time and Space in the History of Cities. In S. Münster, K. Friedrichs, F. Niebling, & A. Seidel-Grzesińska (A c. Di), *Digital Research and Education in Architectural Heritage* (Vol. 817, pp. 47–62). Springer. https://doi.org/10.1007/978-3-319-76992-9_4

Guidarelli, G. (2022). La formazione di una cultura architettonica comune della congregazione di Santa Giustina nel XV secolo: Cantieri, normativa e soluzioni spaziali. In *Dalla riforma di Santa Giustina alla Congregazione Cassinese, genesi, evoluzione e irradiazione di un modello monastico europeo (sec. XV-XVI)- Convegno internazionale, 2019, Padova: Dalla riforma di S. Giustina alla Congregazione Cassinese*. ITA. https://www.research.unipd.it/handle/11577/3477260

Murphy, M., McGovern, E., & Pavia, S. (2013). Historic Building Information Modelling – Adding intelligence to laser and image based surveys of European classical architecture. *ISPRS Journal of Photogrammetry and Remote Sensing*, *76*, 89–102. https://doi.org/10.1016/j.isprsjprs.2012.11.006

Dalla pergamena al virtuale. Manoscritti miniati nello spazio tridimensionale

VALERIA MINISINI[1], BRUNO FANINI[2], GIORGIO GOSTI[3]

[1]SAPIENZA UNIVERSITÀ DI ROMA E CNR-ISPC, ITALIA – VALERIA.MINISINI@UNIROMA1.IT
[2]CNR-ISPC, ITALIA – BRUNO.FANINI@CNR.IT
[3]CINECA, ITALIA – G.GOSTI@CINECA.IT

ABSTRACT (ITALIANO)

I manoscritti miniati medievali sono beni culturali complessi e difficili da valorizzare presso il grande pubblico per le loro problematiche conservative e il contenuto spesso di difficile comprensione. La digitalizzazione dei volumi ha aperto nuove possibilità di fruizione, portando anche all'elaborazione di strategie che sfruttano le potenzialità della realtà virtuale per coinvolgere maggiormente il pubblico, permettendogli non solo di navigare ma di interagire con gli oggetti acquisiti.
Il contributo presenta l'interfaccia immersiva elaborata nel progetto "IlluminAI" per la valorizzazione dei manoscritti miniati che, combinando le tecnologie Web3D con l'intelligenza artificiale (AI), intende offrire una soluzione innovativa per esplorare il contenuto delle pagine in un ambiente 3D strutturato.
Inizieremo delineando il contesto in cui il progetto è stato concepito e sviluppato, presentando esperienze di valorizzazione simili che impiegano visualizzatori e interfacce sia 2D che 3D. Passeremo quindi alla sua descrizione, soffermandoci brevemente sulla rete neurale profonda appositamente creata per classificare le immagini e il dataset usato per il suo addestramento. Ampio spazio verrà poi dedicato all'applicazione web e alla sua interfaccia attraverso un'esposizione dettagliata della suddivisione in gruppi tematici e delle diverse possibilità di disposizione spaziale interna. Dopo aver riportato le interazioni previste, concluderemo presentando i vantaggi del sistema elaborato, il possibile impiego anche con collezioni diverse e gli sviluppi previsti nel prossimo futuro.
Parole chiave: realtà virtuale; intelligenza artificiale; framework ATON; Web3D; manoscritti miniati

ABSTRACT (ENGLISH)

From parchment to virtual. Illuminated manuscripts in the three-dimensional space
Medieval illuminated manuscripts are complex cultural artifacts, difficult to valorize for a larger audience due to their conservation issues and often difficult-to-understand content. The digitization of volumes has opened up new possibilities for their fruition, leading to the development of strategies that leverage the potential of virtual reality to further engage the public, allowing them not only to navigate but also to interact with the acquired objects.
The paper presents the immersive interface elaborated in the "IlluminAI" Project and designed for the illuminated manuscripts enhancement. By combining Web3D technologies with artificial intelligence (AI), it aims to offer an innovative solution for exploring the content of the pages in a structured 3D environment.
We will begin by outlining the context in which the project was conceived and developed, presenting similar enhancement experiences that adopt both 2D and 3D viewers and interfaces. We will then move on to its description, focusing briefly on the deep neural network specifically created to classify the images and the dataset used for its training. Large space will then be dedicated to the web application and its interface, with a detailed description of the thematic grouping and the internal spatial arrangement options. After explaining the planned interactions, we will conclude by presenting the advantages of the developed system, its potential use with different collections, and the progress expected for the near future.
Keywords: virtual reality; artificial intelligence; ATON Framework; Web3D; illuminated manuscripts

1. INTRODUZIONE

La smaterializzazione del libro è un fenomeno che, oltre a caratterizzare il mercato editoriale contemporaneo, ha radicalmente cambiato il nostro modo di fruire i codici antichi liberandoli dalla fragilità della materia (Ambrosio & Vitolo, 2023).

Le diverse campagne di digitalizzazione intraprese da biblioteche e archivi hanno come scopo primario la valorizzazione delle collezioni e non è più raro trovare nei cataloghi virtuali risorse visive in alta risoluzione e corredate da ricchi apparati informativi. L'acquisizione degli esemplari avviene nella maggioranza dei casi per singole pagine e può interessare l'intero codice oppure una selezione di fogli ritenuti particolarmente significativi e funzionali ai fini divulgativi.
Tramite l'adozione di interfacce appositamente sviluppate o la più semplice integrazione nei siti di visualizzatori spesso dotati di funzionalità dinamiche, sfogliare virtualmente un volume medievale è oggi un gesto potenzialmente alla portata di tutti.
I codici miniati, in particolare, sono ampiamente apprezzati dal pubblico anche se privo di conoscenze specialistiche che può trovare all'interno delle miniature figurative indizi utili a comprendere il contenuto delle pagine, involontariamente celato nel testo dalla lingua o dallo stile di scrittura impiegati.
La presenza di apparati illustrativi più o meno elaborati può essere anche il punto di partenza per creare strategie di valorizzazione le quali, coniugando conoscenze proprie dell'ambito umanistico con le nuove possibilità provenienti dal settore informatico, intendono permettere non solo l'accessibilità fisica ma anche quella culturale del materiale librario antico.
Perseguendo un simile connubio è stato avviato il progetto "IlluminAI" che, con la sua applicazione web-based, intende proporre un'interfaccia di visualizzazione accessibile online e capace di rendere le riproduzioni bidimensionali, già acquisite e presenti comunemente nei cataloghi digitali di biblioteche e archivi, fruibili in un ambiente virtuale strutturato.
Inizieremo tracciando le premesse che hanno condotto all'ideazione del progetto concentrandoci, in particolare, sui precedenti tentativi d'applicazione della realtà virtuale ai codici manoscritti per evidenziare i punti di forza come le criticità esistenti in un simile approccio.
Nella descrizione del lavoro svolto, invece, partiremo presentando brevemente le sue tre componenti principali per poi concentrare la nostra attenzione sull'applicazione omonima, ultima di queste per sviluppo e attualmente in fase di completamento. Di essa verranno fornite le principali caratteristiche tecniche, spiegate le meccaniche di funzionamento e illustrate le diverse possibilità di visualizzazione in sperimentazione, passando dalle disposizioni degli oggetti più semplici a quelle maggiormente articolate, sia dal punto di vista spaziale che concettuale.
Prima di concludere con i futuri sviluppi previsti, ci soffermeremo sulle problematiche incontrate e sui punti di forza del sistema che lo rendono impiegabile anche per collezioni diverse.

2. MANOSCRITTI NEL MONDO VIRTUALE

La transizione digitale ha trovato nei codici antichi un naturale catalizzatore che sta conducendo a un cambiamento nel nostro modo di rapportarci ai manoscritti i quali, lasciati i depositi di biblioteche e archivi, hanno trovato spazio e un'inedita visibilità sugli scaffali virtuali di portali istituzionali e cataloghi nazionali (Cappellotto, 2020). Ormai da diversi decenni (Niutta, 2010), la rete è assurta a ultima frontiera verso la quale numerosi e ambiziosi progetti stanno indirizzando i propri sforzi per intercettarne il potenzialmente illimitato pubblico di cui gli studiosi costituiscono solo una piccola percentuale.
Avendo come fine ultimo la valorizzazione, la fase di acquisizione del materiale manoscritto è sempre stata affiancata da iniziative volte alla sua restituzione ma, nel tempo, il progresso tecnologico e la sua assimilazione nel quotidiano hanno condotto alla possibilità di elaborare strategie strutturate, quanto facilmente accessibili, per comunicare messaggi complessi (van Renswoude, 2024).
Nell'eterogeneo panorama attuale diverse soluzioni sono state adottate, la maggioranza delle quali si basa ancora su visualizzatori bidimensionali come l'open-source Mirador[1] che permette la comparazione e l'annotazione di immagini dotate di Manifesto IIIF (Salarelli, 2017). Oltre che fornire un valido strumento per gli studiosi, simili interfacce consentono la creazione di percorsi virtuali originali (O'Daly et al., 2020) oppure derivati da esposizioni reali come nel caso dell'iniziativa *Picturing Prayer. Book of Hours in Houghton Library*[2] della Harvard University.
Anche se ancora poco frequente, il più recente ricorso a tecniche avanzate di acquisizione tridimensionale sta consentendo diverse sperimentazioni con il virtuale nelle sue varie forme. L'impiego della realtà estesa (XR), aumentata (AR) e mista (MR) nel settore culturale è pratica ormai affermata (Bekele et al.,

[1] Per approfondire: https://projectmirador.org/ (cons. 21/04/2026).
[2] https://library.harvard.edu/sites/default/files/static/collections/picturingprayer/index.html (cons. 20/04/2026).

2018), specialmente nella ricostruzione di siti e ambienti oppure per la digitalizzazione di manufatti artistici o archeologici.
Rispetto a questi casi, l'adozione per i manoscritti (Endres, 2019; Schreibman & Papadopoulos, 2019) è in fase di diffusione ma rimane limitata, nel caso dei volumi, dalla loro struttura articolata su più fogli che porta spesso a digitalizzare i codici complessivamente come oggetti e non insieme sequenziale di pagine singolarmente acquisite.
La complessità di realizzare copie 3D sfogliabili, sull'esempio del progetto della University of Southern California diretto da Lynn Dodd e Sabina Zonno[3], porta a concentrarsi su un limitato numero di fogli poi restituiti anche ricorrendo a narrazioni multimediali (Schettino et al., 2023).
Nonostante pochi esemplari che, per conformazione, permettono una più agevole acquisizione totale (Bordignon & Domenici, 2025), le potenzialità di copie anche parziali sono evidenti (Elkhuizen et al., 2024) e stanno portando a investimenti consistenti quanto a progetti di ampio respiro (Chirivì et al., 2026).
Oltre alla restituzione delle immagini, negli ambienti 3D assumono particolare rilevanza le modalità con cui le informazioni collegate vengono visualizzate per arricchire l'esperienza senza impattare sulla fluidità della navigazione. Per tale ragione, il settore delle immersive analytics (IA) (Marriott et al., 2018) sta acquisendo sempre maggiore rilevanza anche in ambito culturale (Koebel et al., 2020), per le promettenti possibilità nella codifica visiva 3D e nell'interazione spaziale con gli oggetti (Saffo et al., 2024).

3. IL PROGETTO ILLUMINAI

Avviato nel 2023 e di prossima conclusione, il progetto "IlluminAI" prevede lo sviluppo di tre elementi distinti ma strettamente interconnessi: un dataset di training, una rete neurale profonda e una webapp XR, fruibile direttamente tramite comuni browser.
Del modello, basato sull'architettura ResNet (He et al., 2016) e addestrato attraverso un ciclo di apprendimento incrementale in più passaggi, è già stata data notizia così come dei risultati di classificazione ottenuti (Minisini et al., 2025). Esso è capace di riconoscere le pagine dei manoscritti anche quando inserite all'interno di collezioni eterogenee di beni culturali e identificare cinque caratteristici elementi di layout: testo, miniature figurative, decorazioni astratte o fitomorfe, musica e iniziali ornamentali.
Complessivamente, è stato utilizzato per etichettare 59.859 immagini, 46.014 delle quali impiegate anche per l'addestramento e confluite nell'IAIset, un dataset per il deep learning contenente dieci classi di oggetti storico-artistici e archeologici.
Le restanti 13.845, invece, provengono dal Gabinetto Fotografico Nazionale di Roma[4] e sono state selezionate in un secondo momento con lo scopo di validare il modello impiegando materiale radicalmente diverso rispetto a quello usato nella fase di training.
Indipendentemente dalla loro provenienza, tutte le riproduzioni sono state uniformate nelle dimensioni e nel formato file prima di essere caricate sul server del CNR-ISPC di Roma, ricorrendo a NextCloud, e collegato alla piattaforma open-source Label Studio (Tkachenko et al., 2026) installata localmente e usata per verificare le predizioni.
Allo stesso modo, le annotazioni generate automaticamente ed esportate in tabelle CSV sono confluite nel database necessario al funzionamento dell'applicazione.

3.1 L'applicazione web XR

IlluminAI è un'applicazione Web3D basata sul framework open-source ATON (Fanini et al., 2021), tra i servizi schierati dall'infrastruttura E-RIHS nel progetto H2IOSC – *Humanities and cultural Heritage Italian Open Science Cloud*.
Nonostante nasca per la gestione di oggetti e scene 3D, ATON possiede caratteristiche che lo rendono vantaggioso e impiegabile anche con riproduzioni bidimensionali. In particolare, la presentazione adattiva e scalabile dei contenuti rende gli oggetti virtuali fruibili su diversi supporti come chioschi museali, dispositivi XR immersivi ma anche computer e smartphone, senza bisogno di installazioni per l'utente finale. A renderlo facilmente implementabile e adattabile a scopi specifici, concorrono la possibilità di

[3] https://dornsife.usc.edu/xrlab/neh-vr-exploration-of-illuminated-manuscripts/ (cons. 21/04/2026).
[4] La sperimentazione ha riguardato i Fondi Zeri, Cisterna-Monti e 11.854 immagini provenienti dall'Archivio del Gabinetto Fotografico Nazionale. Per la collezione completa: https://fotografia.cultura.gov.it/ (cons. 20/04/2026).

manipolare la scena, modificare i livelli oppure la selezione gerarchica quanto intervenire sulla composizione e sulle trasformazioni a cascata degli oggetti.
L'intuitiva interfaccia utente messa a disposizione dagli sviluppatori, infine, è facilmente utilizzabile anche da soggetti inesperti che possono però essere guidati nell'esperienza attraverso la funzionalità multiutente in sessioni collaborative sincrone oppure strutturando percorsi arricchiti con annotazioni semantiche dalla forma rettangolare e poligonale libera.
Il cuore della wabapp è costituito dal database relazionale al suo interno che permette il recupero delle immagini dal cloud e l'associazione con le rispettive annotazioni, le relative mappe di attivazione (Milani & Fraternali, 2021) generate a partire dall'ultimo livello della nostra ResNet e le rielaborazioni ottenute attraverso Depth Anything, un modello open-source di stima della profondità monoculare (MDE) sviluppato dall'Università di Hong Kong (Yang et al., 2024).
La sua struttura modulare si articola in semplici tabelle CSV generate in maniera automatica e, solo se necessario, integrate manualmente o riviste nella struttura per ottimizzare il funzionamento dell'applicazione.
Attualmente, il database di IlluminAI si compone di tre tipologie principali di tabelle che descrivono gli oggetti digitali da visualizzare, i soggetti raffigurati e i volumi d'appartenenza nel caso si tratti di pagine manoscritte o a stampa.
Alla prima appartengono i file generati dal modello e contenenti, oltre le già menzionate annotazioni e la posizione nel cloud, il codice identificativo delle immagini, comune anche alle versioni derivate dalla diversa estensione, e le principali informazioni catalografiche riprese dal sito istituzionale di provenienza. Per evitare che la dimensione di un unico documento con più di 59.000 entrate rallentasse il caricamento dell'applicazione, esse sono state ripartite in quattro distinte tabelle dalla medesima struttura.
Attraverso il record relativo al soggetto raffigurato, il file si collega al vocabolario che riconcilia in un'unica tabella i due sistemi di classificazione iconografica più usati in ambito culturale a livello internazionale: Iconclass, concepito da Henri van de Waal (1910-1972), e il *thesaurus* elaborato dal Getty Research Institute. Ogni personaggio o evento narrativo individuato, inoltre, è stato associato con le coordinate necessarie a collocarlo spazialmente, al momento della visualizzazione, all'interno del cluster tematico stabilito.
Il terzo CSV, invece, permette di ricondurre tutte le pagine originariamente parte del medesimo volume al codice univoco a esso assegnato e alle relative informazioni riportate, in maniera estesa e dettagliata, per consentire confronti e condurre approfondimenti per tipologia, contenuto e ambito geografico di realizzazione.
Il vantaggio principale di un simile database, semplice nel formato per quanto ricco, è la velocità di aggiornamento dei campi esistenti e la facile espansione a cui può essere sottoposto, lavorando sulle tabelle in uso o aggiungendone di nuove nello stesso formato.

3.2 La disposizione delle immagini nei cluster

Per organizzare lo spazio virtuale in cui l'utente andrà a muoversi ed interagire, le riproduzioni sono state suddivise in dieci distinti cluster dalla forma cilindrica e associati a categorie di personaggi, cicli narrativi e generi artistici individuati a partire dal materiale a disposizione. In particolare, quattro contengono raffigurazioni a carattere narrativo legate al Nuovo Testamento, due ospitano personaggi della tradizione cristiana come santi o provenienti dalla Bibbia, tre sono dedicati ai soggetti profani mentre l'ultimo riunisce le pagine prive di miniature o illustrazioni sulla base degli altri elementi grafici individuati nella fase di annotazione.
Essendo IlluminAI concepita per i manoscritti miniati medievali e rinascimentali, ampio spazio è stato dato all'iconografia sacra esplorabile, oltre che nelle miniature, anche nell'arte monumentale a consentire confronti tra le diverse tecniche pittoriche, grafiche e scultoree.
A suggerire questo approccio di "ricontestualizzazione visiva", è stato l'*Atlante della Memoria* di Aby Warburg (1866-1929), storico dell'arte e precursore dell'iconologia come disciplina accademica. In particolare, dei grandi pannelli che formano l'*Atlante* oggi conservati a Londra, è stata ripresa la non sequenzialità nella presentazione del dato visivo e il suo essere nodo nella libera ramificazione spaziale dei concetti, caratteristica che rende la struttura particolarmente adatta alla trasposizione in virtuale (Hristova, 2016).
Sfruttando il database relazionale e le informazioni fornite nelle schede interattive d'approfondimento corredate alle immagini, l'utente potrà spostarsi tra i vari cluster creando un percorso di visita

personalizzato sulla base dei suoi interessi specifici. Sarà nella condizione, inoltre, di interrogare e interagire con i singoli item, attraverso specifiche interfacce spaziali.
Dal lato sviluppatore, invece, è prevista la possibilità di monitorare la navigazione attraverso "Interlumo", una suite di servizi realizzata come pilot E-RIHS sempre nel progetto H2IOSC, appositamente progettata per catturare e analizzare grandi quantità di interazioni remote generate da installazioni fisiche o applicazioni.
Tra le interazioni previste, il fruitore avrà la possibilità di filtrare per categorie predefinite, ingrandire o ridurre le riproduzioni, selezionare gli elementi e raggrupparli in insiemi personalizzabili, quanto creare collegamenti e confrontare le informazioni caricate.
Per quanto riguarda l'esplorazione a 6 gradi di libertà tramite dispositivi XR, abbiamo progettato l'ambiente virtuale lavorando su diverse ipotesi di spazio fisico uguali o superiori a 2x2 metri, variandole in funzione del numero di immagini da visualizzare e della loro disposizione. Per quest'ultima, la nostra attenzione si è focalizzata principalmente su due tipologie di ordinamento: a spirale semplice su circonferenza singola o doppia (Figura 1) e strutturato in spicchi, associati a specifici eventi o personaggi, ulteriormente ripartiti in anelli dall'andamento radiale.
In entrambi i casi, alla profondità sono stati associati valori relativi alla cronologia delle opere, divise per secoli o quarti di secolo, mentre l'altezza nel secondo caso è stata legata alla tipologia di bene raffigurato. Tra le due disposizioni, la differenza più significativa riguarda il numero di immagini gestibili che varia da poche decine a diverse centinaia nella ripartizione più articolata.



Figura 1. Esempi di visualizzazione dei cluster da computer con disposizione interna delle immagini a spirale semplice. Le riproduzioni visibili appartengono al Gabinetto Fotografico Nazionale di Roma.

Una volta scelto il cluster d'interesse, l'applicazione procederà a caricare le miniature degli oggetti contenuti, generate automaticamente al momento della chiamata. Queste sono in scala e hanno risoluzione standard per il web, minore rispetto agli originali in alta definizione visualizzabili, selezionando il singolo item, insieme ai pulsanti di navigazione e alla scheda informativa.
Tanto la partizione in gruppi tematici quanto la disposizione spaziale delle immagini sono state studiate privilegiando la fruizione tramite visore ma adottando accorgimenti tecnici per mantenere la comunicazione efficace indipendentemente dal supporto utilizzato e dalla dimensione dello schermo a disposizione.
La principale sfida tecnica incontrata ha riguardato la gestione di grandi quantità d'immagini per le quali si possono impostare alcune soluzioni a evitare che, l'alto numero di dati da richiamare simultaneamente, rallenti la fase di caricamento del cluster.
Attualmente sono in procinto di iniziare i test utente che intendono verificare l'efficacia dell'interfaccia spaziale, valutando la sua intuitività e la fluidità della navigazione attraverso la raccolta delle opinioni dei

soggetti coinvolti e selezionati per rappresentare diversi livelli sia di conoscenza dei manoscritti che di esperienza con strumentazioni immersive.

4. CONCLUSIONI

Tra le possibilità date dal digitale, la fruizione dei manoscritti in ambienti virtuali è stata esplorata solamente in parte a causa delle complessità insite nell'acquisizione e restituzione dei codici, ma promette di aprire nuove possibilità nella loro valorizzazione.
L'applicazione IlluminAI, permettendo il riuso di materiale già acquisito, intende offrire una soluzione flessibile, facilmente implementabile e adatta a diverse tipologie di utenti, compresi coloro che non apprezzano le esperienze immersive e l'uso dei visori. Inoltre, grazie alla caratteristica fluidità di ATON, non sono necessari ambienti dedicati rendendola adottabile da istituzioni di piccole dimensioni, dotate di spazi limitati oppure impossibilitate ad acquistare costose strumentazioni.
Che la navigazione avvenga attraverso un visore per realtà virtuale, un computer oppure uno smartphone, l'applicazione consentirà di interagire con le collezioni per scoprirle o riscoprirle attraverso un nuovo sguardo. Coinvolgendo l'utente in un viaggio attivo di scoperta del materiale digitalizzato dalle tappe solo parzialmente stabilite a priori, si intende suscitare curiosità verso l'artefatto fisico che non viene sostituito dalla sua copia digitale ma reso più facilmente intellegibile impiegando un nuovo linguaggio.

RINGRAZIAMENTI

Questa ricerca è finanziata dal Progetto H2IOSC - *Humanities and cultural Heritage Italian Open Science Cloud* (https://www.h2iosc.cnr.it/, ultimo accesso 09/04/2025), finanziato dall'Unione europea NextGenerationEU - Piano Nazionale di Ripresa e Resilienza (PNRR), Mission 4 "Education and Research", Componente 2 "From research to business", Investment 3.1 "Fund for the realization of an integrated system of research and innovation infrastructures", Action 3.1.1 "Creation of new research infrastructures strengthening of existing ones and their networking for Scientific Excellence under Horizon Europe", codice progetto IR0000029-CUP-B63C22000730005. Ente di implementazione CNR.

BIBLIOGRAFIA

Ambrosio, A., & Vitolo, P. (Eds.). (2023). *Medioevo digitale. Documenti e archivi. Arte e architettura*. Viella Editrice.

Bekele, M. K., Pierdicca, R., Frontoni, E., Malinverni, E. S., & Gain, J. (2018). A Survey of Augmented, Virtual, and Mixed Reality for Cultural Heritage. *Journal of Computing and Cultural Heritage*, *11*(2), Articolo 7. https://doi.org/10.1145/3145534

Bordignon, A., & Domenici, D. (2025). Enhancing access and understanding of ancient manuscripts: the 3D Digital Edition of the Codex Cospi. *Digital Applications in Archaeology and Cultural Heritage*, *39*, Articolo e00474. https://doi.org/10.1016/j.daach.2025.e00474

Cappellotto, A. (2020). From codex to apps: the medieval manuscript in the age of its digital reproduction. *Umanistica Digitale*, *4*(9), 1–18. https://doi.org/10.6092/issn.2532-8816/11459

Chirivì, A., Pietroni, E., Greco, M., Pandurino, A., & Bucciero, A. (2026). An evolving scenario in the study of illuminated manuscripts: innovative approaches to research questions. *Digital Applications in Archaeology and Cultural Heritage*, *40*, Articolo e00496. https://doi.org/10.1016/j.daach.2026.e00496

Elkhuizen, W. S., Love, J., Parisi, S., Karana, E. (2024). On the Role of Materials Experience for Novel Interactions with Digital Representations of Historical Pop-up and Movable Books. In F. F. Mueller, P. Kyburz, J. R. Williamson, C. Sas, M. L. Wilson, P. T. Dugas, I. Shklovski (Eds.), *CHI '24: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, Association for Computing Machinery, Articolo 619. https://doi.org/10.1145/3613904.3642142

Endres, B. (2019). *Digitizing Medieval Manuscripts: The St. Chad Gospels, Materiality, Recoveries, and Representation in 2D & 3D*. Arc Humanities Press. https://doi.org/10.2307/j.ctvpb3vr0

Fanini, B., Ferdani, D., Demetrescu, E., Berto, S., & D'Annibale, E. (2021). ATON: An Open-Source Framework for Creating Immersive, Collaborative and Liquid Web-Apps for Cultural Heritage. *Applied Sciences*, *11*(22), Articolo 11062. https://doi.org/10.3390/app112211062

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770–778). IEEE. https://doi.org/10.1109/CVPR.2016.90

Hristova, S. (2016). Images as Data: Cultural Analytics and Aby Warburg's Mnemosyne. *International Journal for Digital Art History*, *2*, 117–132. https://doi.org/10.11588/dah.2016.2.23489

Koebel, K., Agotai, D., & Çöltekin, A. (2020). Exploring cultural heritage collections in immersive analytics: challenges, benefits, and a case study using virtual reality. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *XLIII-B4-2020*, 599–606. https://doi.org/10.5194/isprs-archives-XLIII-B4-2020-599-2020

Marriott, K., Schreiber, F., Dwyer, T., Klein, K., Riche, N. H., Itoh, T., Stuerzlinger, W., & Thomas, B. H. (Eds.). (2018). *Immersive Analytics*. *Lecture Notes in Computer Science, vol. 1190*, Springer. https://doi.org/10.1007/978-3-030-01388-2

Milani, F., & Fraternali, P. (2021). A Dataset and a Convolutional Model for Iconography Classification in Paintings. *Journal on Computing and Cultural Heritage*, *14*(4), Articolo 46. https://doi.org/10.1145/3458885

Minisini, V., Gosti, G., & Fanini, F. (2025). IlluminAI: un sistema di navigazione interattivo per i manoscritti miniati rinascimentali. In S. Rebora, M. Rospocher, & S. Bazzaco (Eds.), *Diversità, Equità e Inclusione: Sfide e Opportunità per l'Informatica Umanistica nell'Era dell'Intelligenza Artificiale, Proceedings del XIV Convegno Annuale AIUCD2025* (pp. 478–483). AIUCD.

Niutta, F. (2010). Manoscritti nella rete. *DigItalia, 5*(2), 9–28. https://digitalia.cultura.gov.it/article/view/232

O'Daly, I., van Renswoude, I., & Teeuwen, M. (2020, dicembre). The Art of Reasoning in Medieval Manuscripts. https://art-of-reasoning.huygens.knaw.nl/

Saffo, D., Di Bartolomeo, S., Crnovrsanin, T., South, L., Raynor, J., Yildirim, C., & Dunne, C. (2024). Unraveling the Design Space of Immersive Analytics: A Systematic Review*. IEEE Transactions on Visualization and Computer Graphics*, *30*(1), 495–506. https://doi.org/10.1109/TVCG.2023.3327368

Salarelli, A. (2017). International Image Interoperability Framework (IIIF): una panoramica. *JLIS.it, 8*(1), 50–66. https://doi.org/10.4403/jlis.it-12090

Schettino, P., Pietroni, E., & d'Annibale, E. (2023). Re-Thinking Visitor Experience with Ancient Manuscripts via the Holographic Showcase: The Case of the Codex4D Project and Its First Public Results from a Mixed-Method Evaluation In Situ. *Heritage*, *6*(9), 6035–6065. https://doi.org/10.3390/heritage6090318

Schreibman, S., & Papadopoulos, C. (2019). Textuality in 3D: three-dimensional (re)constructions as digital scholarly editions. *International Journal of Digital Humanities*, *1*, 221–233. https://doi.org/10.1007/s42803-019-00024-6

Tkachenko, M., Malyuk, M., Holmanyuk, A., & Liubimov, N. (2020-2026). Label Studio: data labeling software. Retrieved from https://github.com/HumanSignal/label-studio

van Renswoude, I. (2024). After Digitization, What More?: The Touch and Feel of Virtual Manuscripts. *Quaerendo*, *54*(2–3), 210–225. https://doi.org/10.1163/15700690-bja10026

Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., & Zhao, H. (2024). *Depth Anything V2*, arXiv. https://arxiv.org/abs/2406.09414

Modeling and visualizing palimpsests with IIIF Presentation API 3.0

PANAGIOTIS LEONTARIDIS[1], GIACOMO MARCHIORO[2], GLEN ROBSON[3]

[1] UNIVERSITY OF VERONA, ITALY – PANAGIOTIS.LEONTARIDIS@UNIVR.IT
[2] UNIVERSITY OF VERONA, ITALY – GIACOMO.MARCHIORO@UNIVR.IT
[3] IIIF CONSORTIUM, UK – GLEN.ROBSON@IIIF.IO

ABSTRACT (ENGLISH)

Palimpsests are multi-layered manuscripts made from reused parchment, where the original text has been erased and overwritten with a new one. While advanced imaging has transformed the recovery possibilities of these erased layers, navigating their digital facsimiles is often limited to linear "galleries" that do not reflect the material and structural complexity of the physical object. The community lacks a reusable reference implementation for modeling reconstructed page sequences, spectral layering, and other reading complexities in palimpsests. We propose a comprehensive framework based on the International Image Interoperability Framework (IIIF) Presentation API 3.0 to digitally restore and navigate complex manuscripts within a unified digital environment. Our approach uses IIIF Ranges with the "sequence" behavior to enable toggling between the current folio sequences and the reconstructed ones. The IIIF Choice element facilitates this navigation among spectral layers. We further employ Selectors to isolate or rotate specific undertext segments without altering the original image files.
Keywords: IIIF; palimpsests; interoperability; WEB visualization

ABSTRACT (ITALIANO)

Modellazione e visualizzazione di palinsesti con la IIIF Presentation API 3.0. I palinsesti sono manoscritti multistrato realizzati su pergamena riciclata, in cui il testo originale è stato cancellato e sovrascritto con un nuovo testo. Sebbene l'imaging avanzato abbia radicalmente migliorato la possibilità di recuperare gli strati cancellati, la navigazione delle loro riproduzioni digitali è spesso limitata a "gallerie" lineari che non riflettono la complessità materiale e strutturale dell'oggetto fisico. La comunità non dispone di un'implementazione di riferimento riutilizzabile per la modellazione di sequenze di pagine ricostruite, della stratificazione spettrale e di altre anomalie di lettura nei palinsesti. Proponiamo un framework completo basato sulla Presentation API 3.0 dell'International Image Interoperability Framework (IIIF) per restaurare digitalmente e navigare manoscritti complessi all'interno di un ambiente digitale unificato. Il nostro approccio utilizza i Range IIIF con il behavior "sequence" per consentire il passaggio dalle attuali sequenze dei fogli a quelle ricostruite. Il costrutto Choice facilita la navigazione tra le bande spettrali. Utilizziamo inoltre Selectors per isolare o ruotare segmenti specifici del sottotesto a livello di Canvas, senza alterare le immagini originali.
Parole chiave: IIIF; palinsesti; interoperabilità; visualizzazione WEB

1. INTRODUCTION

Palimpsests are manuscripts whose parchment folios were scraped or washed of their original writings to prepare the surface for reuse. The erased writing is referred to as the undertext, and the new writing as the overtext. Some palimpsests exhibit multiple such recycling cycles in their lifetime, resulting in intermediary textual layers (Rossetto, 2023). The codex in its present form is known as the *codex recentior*, while the source codices from which the parchment was reused are known as the *codices antiquiores*.

Researching palimpsests involves both recovering the erased text and reconstructing the original page composition and sequence to facilitate textual identification and analysis. Contemporary recovery is achieved through advanced imaging techniques at different energy levels, primarily using Multispectral Imaging (MSI) and, in more challenging cases, X-ray Fluorescence (XRF) (Toth et al., 2020). MSI captures multiple images of a single page under varying lighting conditions, which are subsequently processed to facilitate comparison of its features. These features, some invisible to the naked eye, such as ink traces embedded in the collagen of the reused parchment, can help enhance and distinguish overlapping textual features. Nevertheless, a reusable reference implementation for the efficient

visualization of layered image data and the complexity of reconstructed manuscript structures remains lacking, with repositories usually prioritizing the structure of the *codex recentior*, while the navigability of the *codices antiquiores* traditionally relies on catalog descriptions or collation visualizations.[1]
We present the work in progress towards a reference implementation using the International Image Interoperability Framework (IIIF) Presentation API 3.0[2] to address the challenges in palimpsest visualization that we identify. IIIF provides a set of standards for the interoperable delivery of digital multimedia content, metadata enrichment, and resource discovery. Specifically, Presentation API 3.0 can describe resources and their relationships in a JSON-LD file known as a Manifest. A Manifest incorporates descriptive metadata, image annotations, and rendering instructions for displaying resources in rich, interactive environments. Unlike alternative data models for structuring the digitization of library materials, such as MAG of Italy's Central Institute for the Union Catalog (ICCU) or Encoded Archival Description (EAD), IIIF has, since its first implementation,[3] long considered the modeling of complex digital objects, such as palimpsests.

2. STATE OF THE ART IN PALIMPSEST VISUALIZATION WITH IIIF

Several projects have approached the visualization of complex books using IIIF. The Sinai Palimpsests Project[4] represents the most extensive systematic application of MSI to date. Beginning in 2011, the Early Manuscripts Electronic Library (EMEL) imaged a substantial part of the known palimpsest codices at Saint Catherine's Monastery, totaling 6,800 palimpsest pages (Phelps, 2023). To publish this material, UCLA developed a customized implementation of the Mirador[5] viewer tailored to the project-specific needs.[6] Among its innovations is an interface that allows users to layer multiple spectral images of a single page with adjustable opacity for feature comparison. For a similar purpose, the implementation supports synchronized windows, which enable mirrored panning, zooming, and rotation across views. Similarly, the collaboration between the Vatican Library and Stanford University Libraries sought to enhance IIIF protocols for palimpsest visualization through experimentation with Mirador and the implementation of tools for the parallel display of image variants within a single Manifest, including a mirroring functionality to address accidental ink transfers during parchment reuse (Németh, 2023).
Other projects have addressed structural issues arising from damaged, fragmented, or dispersed materials. Fragmentarium[7] uses IIIF to enable the reuse of images and metadata from diverse sources to virtually reassemble original codices from surviving fragments (Duba, 2019). Biblissima[8] employed the IIIF Presentation API 2.0[9] to virtually reconstruct the *Florus dispersus*, reuniting its 117 folios now preserved across three collections (Koroleva, 2015). In this context, separated folios are organized within the Manifest's sequences, and empty Canvases are inserted where needed to represent *lacunae*. Both are practices relevant to the reconstruction of usually fragmented palimpsest texts.

3. REQUIREMENTS IDENTIFIED FOR PALIMPSEST VISUALISATION

To better contextualize the structural challenges that this study seeks to address, we first provide a representative case study.

3.1 Case Study

Manuscript XL of the Biblioteca Capitolare di Verona is a 7th–8th-century copy of Gregory the Great's *Moralia in Iob*. It is a partial palimpsest that incorporates parchment from four distinct *codices antiquiores*. Among these is a significant late 5th-century fragmentary witness to Virgil's works and *codex unicus* to the scholia surrounding the Virgilian text (De Paolis, 2025).

[1] A notable example is *VisColl* (https://viscoll.org/, cons. 20/04/2026), a digital tool for modeling and visualizing the physical collation of manuscripts. With a data model at its core, it encodes quires, bifolia, and folio relationships to generate collation diagrams and formulas (Porter et al., 2017).
[2] (cons. 05/02/2026)
[3] https://iiif.io/api/metadata/1.0/ (cons. 05/02/2026)
[4] https://sinaied.library.ucla.edu/ (cons. 05/02/2026)
[5] https://projectmirador.org/ (cons. 05/02/2026)
[6] For all project-specific features, see https://sinai.library.ucla.edu/user-guide/navigating-site (cons. 05/02/2026)
[7] https://fragmentarium.ms/ (cons. 05/02/2026)
[8] https://projet.biblissima.fr/ (cons. 05/02/2026)
[9] Demos are available at https://demos.biblissima.fr/ (cons. 05/02/2026)

The layout of the undertext closely resembles that of the overtext (Figure 1A). Each bifolio page contains a single column of text, though the undertext is often rotated 180° relative to the overtext. Following conservation, the manuscript was cut open; the palimpsest section now survives as loose bifolios, whereas the non-palimpsest ones were rebound.
Baschera (1999) reconstructs the original structure of the Virgilian codex and identifies several *lacunae*. Figure 1B shows the 34th gathering (a quinion), of which only bifolios 212–215 and 252–257 survive unbound. Each extant bifolio image thus captures two non-sequential pages from both the *codex antiquior* and the *codex recentior*.

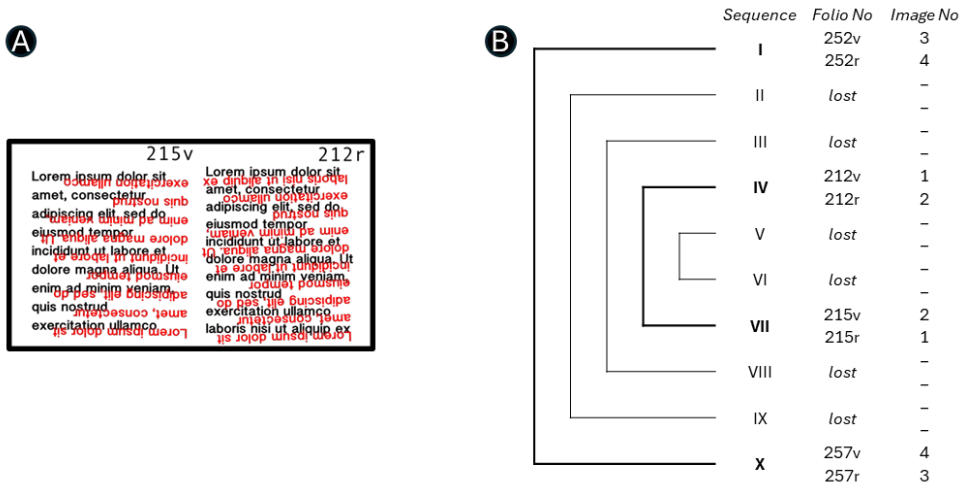


Figure 1. A: Mockup of a loose palimpsest bifolio. Both the overtext (in black) and the undertext (in red) occupy a single column on each page, but the undertext is rotated 180° relative to the overtext. B: Reconstruction of the 34th quinion of *Vergilius Veronensis* (Baschera, 1999).

3.2 Requirements

The features of manuscript XL point to a set of requirements that guide the modeling proposed for palimpsest visualization.
Layer visualization. MSI is among the standard methods for non-invasive analysis of palimpsests. It produces image stacks across a range of wavelengths (from ultraviolet to infrared), which may be further processed using statistical methods, such as Principal Components Analysis (PCA), to produce derivative image layers. To support sophisticated research, a visualization framework must allow users to layer and manipulate these stacks in order to compare spectral data and isolate features useful for deciphering the undertext.
Page reconstruction. Scribes often repurposed parchment without adhering to its original sequence or orientation. Consequently, the undertext may be oriented at 0°, 90°, or 180° relative to the overtext. A visualization framework must therefore account for page-level rotation for reading the undertext.
In some cases, a single large folio of the *codex antiquior* could have been rotated 90° to span an entire bifolio of the *codex recentior* (Figure 2, Case 1). Reconstructing such a folio requires combining disparate images drawn from different parts of the book into a unified view. Conversely, in unbound cases such as palimpsest manuscript XL, each loose bifolio contains two non-sequential pages of the original codex, for instance, pages 4 and 13 in the second bifolio of a quaternion (Figure 2, Case 2C). Here, segmentation is necessary to ensure proper sequencing.
Sequence reconstruction. The final requirement is the digital reconstruction of the *codex antiquior* according to its hypothesized codicological structure. This entails organizing reconstructed and layered pages into their hypothesized positions while accounting for *lacunae* and missing images. Additionally, the implementation must support the reunification of fragments originating from multiple host codices or separate institutional collections.

4. PROPOSED IMPLEMENTATION WITH IIIF PRESENTATION API 3.0

Beyond image interoperability, which enables the reuse of images from institutions adopting the framework, the IIIF Presentation API offers a flexible data model for describing complex digital objects.

The primary resource for modeling is the Manifest, a JSON-LD document that encapsulates the object and its constituent parts. Images are rendered on Canvases listed inside the "items" property. Structural ordering is defined by the "structures" property, which contains Ranges that logically organize Canvases by their content. Spatial manipulation is achieved through targeting mechanisms such as SpecificResource and FragmentSelector.
A simplified Manifest structure for a palimpsest with MSI image stacks and alternative orderings may be outlined as follows:

- Manifest
    - Items
        - Canvas A (a folio or bifolio)
            - AnnotationPage
                - Annotation
                    - Choice (alternative image bodies for the same Canvas)
                        - Items
                            - SpecificResource (used when rotation is needed)
                                - Image A_1 (e.g., **natural light**)
                                - Image A_2 (e.g., **infrared**)
                                - Image A_3 (e.g., **ultraviolet**)
        - Canvas B
    - Structures
        - Range A (behavior "sequence", can represent a specific reading order, e.g., that of the **overtext**)
            - Items
                - Reference to a Canvas URI
                    - SpecificResource (used when spatial targeting is needed)
                - Reference to another Canvas URI
        - Range B (behavior "sequence", can represent an alternative reading order, e.g., that of an **undertext**)

MSI and multiple images in a single view. The IIIF Cookbook[10] includes a recipe for *Multiple Choice of Images in a Single View*,[11] designed for cases where several registered image stacks (such as MSI or PCA derivatives) align to a single Canvas. It employs the Choice body type defined by the W3C Web Annotation Data Model as an ordered list of resources from which the client selects a single item for processing or display (World Wide Web Consortium, 2017). Each item in the Choice element is positioned so it extends across the entire Canvas; therefore, each image in the stack must have at least the same aspect ratio.
Alternative page sequences. A core contribution of this work is the proposed *Alternative Page Sequences*[12] recipe, which can address the sequencing-related structural complexities of palimpsests. The recipe, not limited exclusively to palimpsests, can represent the pages of a misbound or structurally altered book in their current order, while allowing users to toggle to alternative (reconstructed) reading sequences.
We define alternative orderings through Ranges in the "structures" property to facilitate navigation logic beyond the default linear array of Canvases. These Ranges function as ordered lists that may reference entire Canvases, specific Canvas segments, or other Ranges. A Range specifies a reading path only when its behavior value is set to "sequence".
Segmentation and spatial manipulation. Certain structural conditions, as illustrated in Figure 2,[13] require image cropping, rotation, or composition.

- **Bifolio segmentation (Cases 2C and 2D).** The image of a loose bifolio often contains two non-sequential pages (Case 2C). To restore proper sequencing, we may isolate the relevant page region by referencing a SpecificResource on a Range that points to the full bifolio Canvas. Alternatively, a FragmentSelector may specify the page region to render rather than the entire bifolio.

[10] The IIIF Cookbook (https://iiif.io/api/cookbook/, cons. 20/04/2026) collects IIIF "recipes", reusable code snippets for the creation of IIIF Manifests for common use cases.
[11] https://iiif.io/api/cookbook/recipe/0033-choice/ (cons. 05/02/2026)
[12] https://iiif.io/api/cookbook/recipe/0027-alternative-page-order/ (cons. 05/02/2026)
[13] Figure 2, Case 2 is modeled on another palimpsest of the Biblioteca Capitolare di Verona, manuscript LXII, an 8th-century copy of Cresconius' *Concordia Canonum* written on partially palimpsest parchment sourced from a heavily annotated 6th-century copy of the *Codex Iustinianus*. Each palimpsest leaf was rotated 90° and folded in two to accommodate a bifolio of the *codex recentior*. The palimpsest part is preserved unbound.

In Case 2D, reconstructing the undertext reading sequence spanning the entire bifolio requires rotating its view. While one might expect behavior analogous to cropping when rotating a Canvas, Canvas rotation is not yet standardized in the Presentation API 3.0. A possible workaround is to create a new Canvas and rotate each image in the Choice stack using ImageApiSelector. However, this introduces duplicated Canvases of each rotated page containing Choices, and possibly Annotations that refer to the same page but in different orientations.

- **Joining disparate images (Case 1B).** In bound palimpsests, when a single undertext page spans an entire bifolio of the *codex recentior*, the original page is captured in two non-sequential images. To reconstruct the original, larger page, we must unify the two images within a single view. As implemented in the Cookbook's *Composition from Multiple Images*,[14] each half of the Canvas can be targeted by an Annotation that assigns the appropriate image to a defined region with a Selector. The Canvas rotation complication applies here as well.

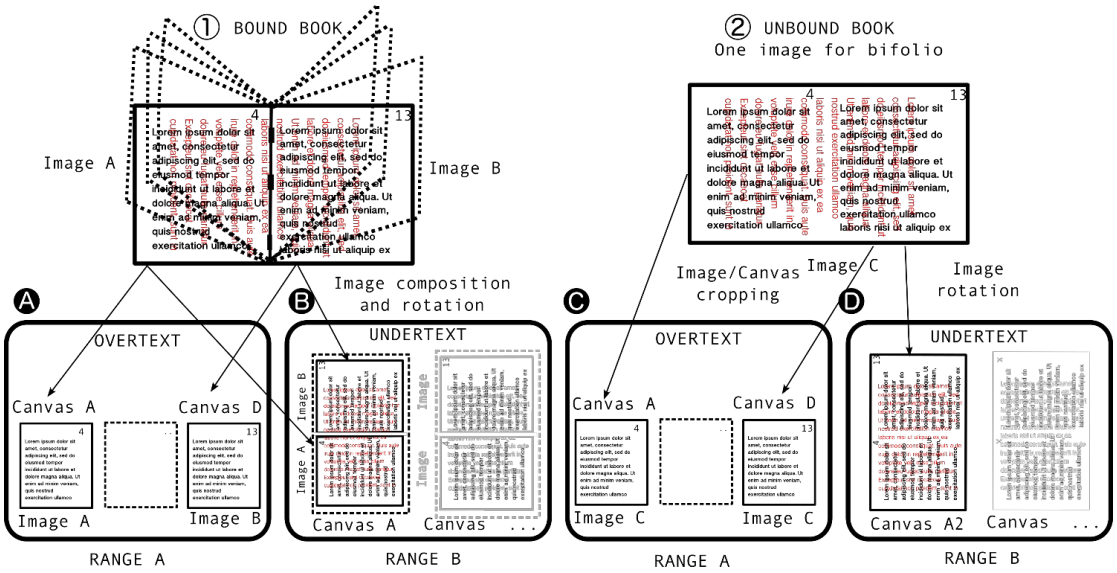


Figure 2. A palimpsest where the undertext is rotated 90° to accommodate a bifolio in the *codex recentior*. Case 1, bound palimpsest codex: (A) the structure of the overtext is maintained after the imaging of the individual pages; (B) the undertext page must be reconstructed, for it exists in two different, non-sequential images of the individual pages. Case 2, loose bifolio: (C) each image contains two non-sequential pages of the overtext that must be isolated and placed in the correct reading order. (D) The undertext is not disturbed by binding; the folio can be rotated to its original orientation to facilitate reading.

5. DISCUSSION

In palimpsest visualization, the image stack (comprising the various spectral or processed layers associated with a page) should be treated as a single conceptual unit. Any subsequent transformation, such as rotation or cropping, should therefore be applied to the entire stack. Nevertheless, under the current Presentation API 3.0, accommodating rotation requires the creation of separate Canvases. Creating a Canvas for each rotation implies that Annotations targeting a specific region on the primary Canvas do not translate to the corresponding region in the secondary rotated Canvas. While this complication may prove useful when creating annotations specific to different textual layers, it ultimately raises the question whether the Canvas is a temporary container for resources or a "common reference system" where all resources and annotations coexist.

6. CONCLUSION

Structural and material complexity are central challenges in the study of palimpsests, and the community currently lacks a standardized reference implementation for navigating their digitizations. Therefore, we propose a specification-level approach to representing palimpsests and other complex books using the

[14] https://iiif.io/api/cookbook/recipe/0036-composition-from-multiple-images/ (cons. 05/02/2026)

IIIF Presentation API 3.0 that addresses identified requirements, such as layering, rotation, cropping, composition, and sequencing. In doing so, we highlight both the specification's flexibility and identify a potential "limiting case" related to rotation.
The work is developed incrementally, with individual requirements addressed through discrete proposals (such as the *Alternative Page Sequences* recipe), as initial steps towards a more integrated modeling framework. While still a work in progress, the proposal offers a structured approach for disseminating complex manuscript research and reconstructions to specialists and broader digital audiences, materials traditionally confined to catalog descriptions and collation schemata.

ACKNOWLEDGEMENTS

The authors thank Trip Kirkpatrick (Yale Library) and Robert Casties (MPIWG) for the valuable feedback on the IIIF implementation of palimpsest visualization provided during the IIIF Cookbook community calls.
Research on the palimpsests of the Biblioteca Capitolare of Verona is part of the project *From Antiquity to Community: Rethinking Classical Heritage through Citizen Humanities* (AntCom). The AntCom project has received funding from the Horizon Europe research and innovation program under Horizon MSCA Doctoral Networks Grant Agreement No. 101073543.
The authors acknowledge the use of AI-assisted tools for grammar and stylistic revision.

REFERENCES

Baschera, C. (Ed.). (1999). *Gli scolii veronesi a Virgilio: Introduzione, edizione critica e indici* (1. ed.). Casa Ed. Mazziana.

De Paolis, P. (2025). I Codices Vergiliani Antiquiores e la Scuola. In M. Bassetti & P. Degni (Eds), *Città Cattedrali Biblioteche al Passaggio tra Tardoantico e Alto Medioevo: A partire dal caso di Verona* (pp. 117–145). Fondazione Centro Italiano di Studi sull'Alto Medioevo.

Duba, W. O. (2019). ‚Fragmentarium'. *Das Mittelalter*, 24(1), 221–223. https://doi.org/10.1515/mial-2019-0015

Koroleva, E. (2015). *Les Manuscrits de Florus de Lyon: Opérations de reconstitution virtuelle et mapping de données dans le cadre du projet Biblissima*. École nationale des Chartes.

Németh, A. (2023). Interactive Learning in Palimpsest Research. In C. Rapp, G. Rossetto, J. Grusková, & G. Kessel (Eds), *New light on old manuscripts: The Sinai palimpsests and other advances in palimpsest studies*. Austrian Academy of Sciences Press. https://doi.org/10.1553/978OEAW91575s499

Phelps, M. B. (2023). The Sinai Palimpsests Project (2011–2016): Goals, Methods, and Contributions. In C. Rapp, G. Rossetto, J. Grusková, & G. Kessel (Eds), *New light on old manuscripts: The Sinai palimpsests and other advances in palimpsest studies*. Austrian Academy of Sciences Press. https://doi.org/10.1553/978OEAW91575s31

Porter, D., Campagnolo, A., & Connelly, E. (2017). VisColl: A New Collation Tool for Manuscript Studies. In H. Busch, F. Fischer, & P. Sahle (Eds.), *Kodikologie und Paläographie im digitalen Zeitalter*. 4. BoD.

Rossetto, G. (2023). *Greek palimpsests at Saint Catherine's monastery (Sinai): Three euchologia as case studies*. Austrian Academy of Sciences Press.

Toth, M. B., Christens-Barry, W. A., Calabro, D., Heintzelman, M. Z., Moreton, M., & Stewart, C. (2020). Integrating Advanced Imaging of Ancient Manuscripts. *Archiving Conference*, 17(1), 33–38. https://doi.org/10.2352/issn.2168-3204.2020.1.0.33

World Wide Web Consortium. (2017, February 23). *Web Annotation Data Model*. https://www.w3.org/TR/annotation-model/

[5/C] Crowdsourcing e public engagement

Tag ANIMALx: Crowdsourcing the Non-Human in Cultural Heritage Collections

DANIELA TEIXEIRA GOMES[1], CARLA VIEIRA[1], JOANA VIEIRA PAULINO[2]

[1]CHAM
[2]IHC/IN2PAST

ABSTRACT (ENGLISH)

Faced with evolving forms of collaboration between humanities researchers and society, as well as with the expanding landscape of digital methods, this communication presents the development workflow, challenges, and preliminary results of *Tag ANIMALx*, a digital tool developed within the project "Animals of Lisbon" (ANIMALx), funded by the Portuguese Foundation for Science and Technology. Situated at the intersection of Digital Humanities (DH), Animal Studies and Citizen Science, this initiative addresses a significant gap in historiography: the invisibility of non-human animals in standard museum inventories. While animals are ubiquitous in historical records, they are rarely indexed as primary subjects, rendering Animal History difficult to research systematically.

To address this, *Tag ANIMALx* employs a gamified digital interface designed to interact with the Museum of Lisbon's collections. By engaging the public with simple, structured inquiries regarding inventory images, the tool crowdsources the identification and classification of animal representations. This process serves a dual purpose: it generates essential data for the ANIMALx project's research on historical human-animal relations and significantly enriches the Museum's metadata, making the collection "searchable" regarding non-human agents, augmenting its descriptions and accessibility. Furthermore, it engages communities in knowledge co-construction, promoting an active participation from a wider audience.

Beyond detailing the tool's development status, this communication aims to highlight a workflow built on replicable, open-source software widely adopted in the DH community. This ensures the model can serve as a reference for future heritage projects. More than a data-collection utility, *Tag ANIMALx* tests the viability of public collaboration in complex archival tasks and the potential for automating metadata enrichment. Ultimately, by fostering the co-construction of knowledge, the project promotes a multispecies historical narrative, demonstrating that scientific authority is shared between institutions, citizens, and the non-human subjects of history.

Keywords: Citizen Science; Animal History; Gamification; Metadata Enrichment; Cultural Heritage

ABSTRACT (ITALIANO)

Tag ANIMALx: Crowdsourcing dei soggetti non umani nelle collezioni del patrimonio culturale

Alla luce delle forme di collaborazione in continua evoluzione tra i ricercatori umanistici e la società, nonché del panorama in espansione dei metodi digitali, questa comunicazione presenta il flusso di lavoro di sviluppo, le sfide e i risultati preliminari di *Tag ANIMALx*, uno strumento digitale sviluppato nell'ambito del progetto esplorativo "Animals of Lisbon" (ANIMALx), finanziato dalla Fondazione portoghese per la scienza e la tecnologia. Situata all'incrocio tra Digital Humanities (DH), Animal Studies e Citizen Science, questa iniziativa affronta una lacuna significativa nella storiografia: l'invisibilità degli animali non umani negli inventari standard dei musei. Sebbene gli animali siano onnipresenti nei documenti storici, raramente vengono indicizzati come soggetti primari, rendendo difficile la ricerca sistematica della storia degli animali.

Per ovviare a questo problema, *Tag ANIMALx* utilizza un'interfaccia digitale gamificata progettata per interagire con le collezioni del Museo di Lisbona. Coinvolgendo il pubblico attraverso semplici richieste strutturate relative alle immagini dell'inventario, lo strumento raccoglie in crowdsourcing l'identificazione e la classificazione delle rappresentazioni animali. Questo processo ha un duplice scopo: genera dati essenziali per la ricerca del progetto ANIMALx sulle relazioni storiche tra uomo e animale e arricchisce in modo significativo i metadati del Museo, rendendo la collezione "ricercabile" per quanto riguarda gli agenti non umani, aumentando le descrizioni e l'accessibilità. Inoltre, coinvolge le comunità nella co-costruzione della conoscenza, promuovendo una partecipazione attiva da parte di un pubblico più ampio. Oltre a descrivere in dettaglio lo stato di sviluppo dello strumento, questa comunicazione mira a mettere in evidenza un flusso di lavoro basato su un software open source replicabile e ampiamente adottato nella

comunità DH. Ciò garantisce che il modello possa servire da riferimento per futuri progetti sul patrimonio culturale. Più che un'utilità per la raccolta di dati, *Tag ANIMALx* verifica la fattibilità della collaborazione pubblica in compiti di archiviazione complessi e il potenziale di automazione dell'arricchimento dei metadati. In definitiva, promuovendo la co-costruzione della conoscenza, il progetto promuove una narrazione storica multispecie, dimostrando che l'autorità scientifica è condivisa tra istituzioni, cittadini e soggetti non umani della storia.
Parole chiave: Scienza dei cittadini; Storia degli animali; Gamification; Arricchimento dei metadati; Patrimonio culturale

1. INTRODUCTION

In the contemporary landscape of Digital Humanities (DH), the role of Galleries, Libraries, Archives, and Museums (GLAM) has undergone a profound transformation. No longer static repositories dedicated solely to the custody and preservation of physical objects, these institutions are increasingly positioning themselves as dynamic platforms for open access, social interaction, and knowledge co-creation. This paradigm shift aligns with the broader Open Science movement, which advocates not only transparency in scientific results but also the opening of the research process to active civil society participation. Within this context, collaboration has become a structural component of research, fostering the rise of "Citizen Humanities"—a specific and evolving branch of Citizen Science. Unlike traditional citizen science projects often focused on observational data collection in the natural sciences, Citizen Humanities centers on the interpretation, classification, and enrichment of cultural artefacts, recognizing that the data in humanities is intrinsically interpretative, qualitative, and subjective (Heinisch, 2021).

However, the democratization of historiographical research through digital means faces critical epistemological and operational challenges. Firstly, there is a lingering scepticism within the academic community regarding the reliability and the quality of data produced by non-specialists. Secondly, GLAM institutions and research projects face the practical difficulty of sustaining volunteer engagement over long periods, preventing initiatives from becoming ephemeral "communication campaigns" rather than robust research infrastructures.

This communication addresses these challenges through the case study of *Tag ANIMALx*, a digital tool developed within the exploratory project "Animals of Lisbon" (ANIMALx), undertaken at NOVA University Lisbon. The project identifies and seeks to resolve a significant gap in current historiography: the archival invisibility of non-human animals in standard museum inventories. While animals are ubiquitous in historical records—present in iconography, descriptions of daily life, economic logs, and urban planning documents—they are rarely indexed as primary subjects. Consequently, Animal History remains difficult to research systematically due to anthropocentric cataloging practices that prioritize human authorship, artistic technique, or general themes over non-human presence (Bonnell & Kheraj, 2022).

To bridge this gap, we present the comprehensive, reproducible workflow for *Tag ANIMALx*, a gamified crowdsourcing platform designed to interact with the Museum of Lisbon's digital collection. Unlike *ad-hoc* initiatives that often lack documentation and long-term sustainability, this project relies on FAIR principles (Findable, Accessible, Interoperable, Reusable), on community knowledge co-collaboration and on enriching the descriptions and accessibility of museum collections. By detailing the methodological architecture, from the initial diagnosis of metadata gaps to the implementation of hybrid validation protocols, this contribution demonstrates how gamified co-construction can generate robust metadata, effectively transforming the public from passive consumers into active producers of historical knowledge.

2. THEORETICAL FRAMEWORK AND STATE OF THE ART

The theoretical foundation of the ANIMALx project is grounded in the animal turn, an intellectual movement that challenges the traditional anthropocentric view positioning humans as the exclusive agents of history. As argued by current scholarship in the field, particularly by authors such as Mahesh Rangarajan and Erica Fudge, Animal History seeks to reconstruct narratives by considering the "animal point of view" and the agency of animals in shaping social, economic, and ecological contexts (Fudge, 2017; Rangarajan, 2013). This perspective argues that history is not constructed solely by one species, but unfolds within shared circumstantial constraints where humans and animals adapt and react to one another.

However, operationalizing this "zoocentric" perspective is hindered by the structure of traditional archives, which are organized according to human-centered logics. As noted by Bonnell and Kheraj

(2022), animals in archives often appear merely as statistics, commodities, or background scenery, lacking biographical or specific indexing. A researcher seeking the history of a specific species often encounters fragmented references hidden under disparate categories like "public hygiene," "property," or "crime," without a unified access point. The solution, therefore, lies not in refunding archives but in enriching them with new metadata layers that describe species, behaviors, and interactions, effectively making the "animal trace" findable and computable.

To process the vast visual data required for this historiographical shift—specifically, identifying animals in thousands of unindexed images—the project employs crowdsourcing. While pioneering projects such as *Old Weather* or *Transcribe Bentham* proved the efficacy of volunteer contributions for transcription tasks (Causer & Terras, 2014), sustaining volunteer motivation remains a significant hurdle. The "duty to science" is often insufficient to maintain long-term engagement from a heterogeneous public.

In order to address this hurdle, *Tag ANIMALx* integrates principles from "serious games". As defined by Laamarti et al. (2014), serious games are applications whose primary purpose is not entertainment, but which use entertainment mechanics to achieve a practical goal, in this case, data curation. By applying gamification mechanics, such as feedback loops, progression bars, community goals, and leaderboards, tasks that require repetitive cognitive effort, such as image classification, become engaging and rewarding. This approach transforms the crowdsourcing experience from a micro-tasking obligation into a meaningful, playful interaction. Projects like *Steve.museum* or *Tag! You're It* (Brooklyn Museum) have previously demonstrated that gamified interfaces can successfully engage younger demographics and produce valid folksonomies that enhance collection accessibility (Ridge, 2013).

3. METHODOLOGY

The development of Tag ANIMALx as a replicable digital tool follows a structured four-phase workflow designed to ensure interoperability, data quality, and replicability. This section details the technical and curatorial decisions taken from initial data diagnosis to public operationalization.

3.1 Phase 1: Data Acquisition and Standardization

The project operates on the digital collection of the Museum of Lisbon, which comprises approximately 55,000 records hosted on the *InPatrimonium* repository. During the diagnostic phase, it was determined that direct web scraping was not viable due to security constraints. Consequently, a strategy of CSV data export was adopted. To ensure the tool's reliability during development, a controlled sample of 127 records was established. These records were manually validated by the research team to serve as a "ground truth" for calibrating the crowdsourcing tool and testing user accuracy in the early stages. Later, a representative corpus of around 8,000 records was established, which would, progressively, feed the database on which Tag ANIMALx operates.

Following extraction, extensive data cleaning was conducted using *OpenRefine*. This step was crucial to resolving semantic ambiguities inherent in natural-language metadata. For instance, the system had to distinguish the Portuguese surname "Lobo" (Wolf) from the animal "lobo" (wolf), or differentiate "rato" (mouse) from "prato" (plate), or toponyms like "Largo do Rato" (Rato Square). Crucially, the metadata structure was harmonized by mapping the Museum's original schema to a hybrid model combining Dublin Core (for standard cultural description) and Darwin Core (for biological description). This integration ensures that new data fields, for example, *Scientific Name* and *Taxonomic Category* are scientifically valid and interoperable with global biodiversity databases, adhering to the project's goal of interdisciplinary connectivity.

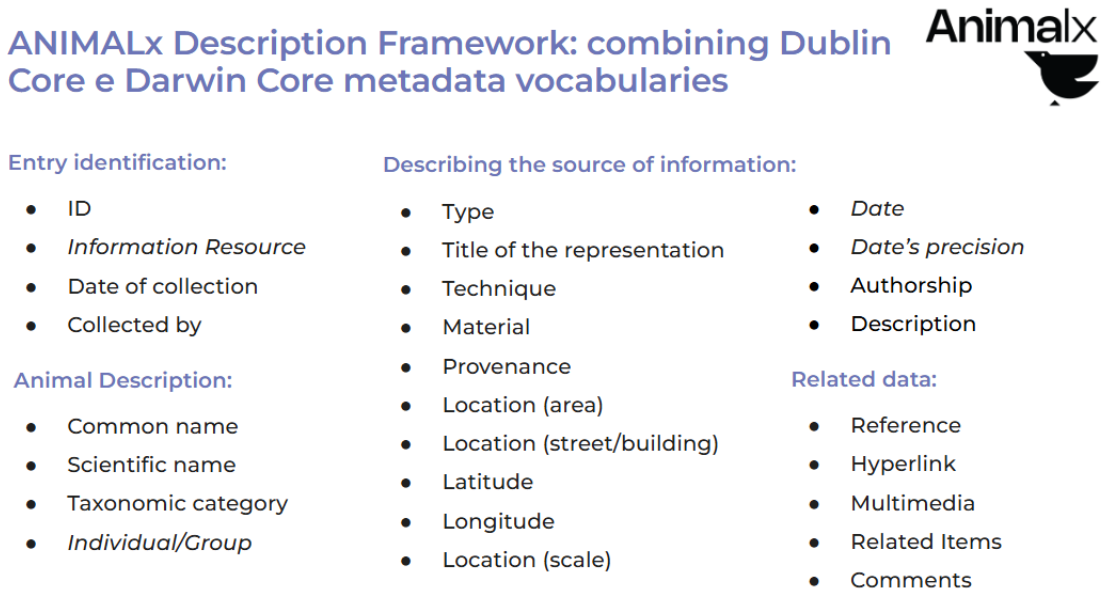


Figure 1. Metadata Model. The schema demonstrates the integration between Dublin Core (for cultural data) and Darwin Core (for biological data).

To ensure future interoperability with macro-level cultural heritage ontologies like CIDOC CRM, our hybrid Dublin Core (DC) and Darwin Core (DwC) approach functions as a foundational, structured dataset ready for semantic crosswalking. While CIDOC CRM provides a complex, event-centric ontology, the descriptive metadata generated through our Dublin and Darwin Core model can be systematically mapped to CRM entities. For instance, a Darwin Core taxonomic classification, such as "*Scientific Name*" can be mapped to a CIDOC CRM E55 Type, while a Dublin Core descriptive field maps to an E28 Conceptual Object or an E39 Actor linked via an E65 Creation event. This strategy ensures that the Tag ANIMALx dataset remains agile: it provides immediate practical utility for current GLAM and biodiversity aggregators, while maintaining the structural integrity required to be ingested into broader, Linked Open Data (LOD) ecosystems in the future (Doerr et al., 2003; Page, 2019).

3.2 Phase 2: Technical Architecture and Software Selection

A critical phase of the methodology involved selecting the infrastructure. Four solutions were evaluated: Flickr, Zooniverse, MetaData Games, and a custom Omeka S implementation. While *Zooniverse* offers a vast user base, its closed API architecture limits deep integration with the project's main website (developed in Omeka S) and custom metadata fields. *Flickr*, while accessible, lacks the structured data capabilities required for scientific research. *MetaData Games*, though theoretically aligned, is no longer actively supported. Therefore, the project opted for a customized implementation within Omeka S. This strategic choice guarantees data centralization, keeping the crowdsourced metadata within the same ecosystem as the main ANIMALx database. The interface was developed using custom HTML/CSS and JavaScript plugins to transform the standard data entry process into a "serious game". To sustain engagement, the tool plans to implement immediate feedback mechanisms, such as progress bars and counters, alongside extrinsic rewards, such as community rankings. The design prioritizes accessibility, replacing complex archival jargon with accessible prompts (e.g., asking "Where is the animal in this image?" instead of requesting spatial coordinates), thereby lowering the barrier to entry for non-specialist participants and aiming for community active engagement through the sense of achievement and accomplishment.

3.3 Phase 3: The Hybrid Validation Workflow

To address the reliability gap often associated with citizen science, Tag ANIMALx employs a multi-layered validation protocol designed to ensure data integrity before it enters the institutional catalogue.

1. Consensus Algorithm: The first layer is automated. Each record is served to multiple independent users. A "valid" classification is only generated when a statistical consensus is reached, for

example, of 5 identical tags from unique users. This approach, which relies on a crowd consensus, filters out individual errors and malicious tagging.

2. Expert Review: The second layer is curatorial. Aggregated data that passes the consensus threshold enters a staging area ("Treated Records"). Here, the research team performs a final manual verification of the results obtained through the first validation process. Only after this expert approval are the new metadata fields published to the main ANIMALx catalog. This hybrid model ensures that gamification serves precision, filtering out errors while maintaining high throughput, and validates the premise that non-specialists can contribute meaningfully to specialized datasets.

3.4 Institutional Return and Open Science Strategy

The operational cycle concludes with the production of a comprehensive report documenting the new metadata generated for each record, as well as identifying gaps or errors in the original institutional descriptions detected during the crowdsourcing process. This structured dataset is submitted to the Museum of Lisbon as a form of institutional return, offering value-added information that—pending institutional deliberation—may be directly integrated into the museum's digital repository.

Furthermore, to ensure the sustainability and replicability of the proposed workflow, the project commits to the full availability of its technical documentation and source code. These resources will be deposited in open-access repositories such as GitHub and Zenodo under a Creative Commons license. This strategy ensures that the architectural model of Tag ANIMALx can be audited, reused, or adapted by other researchers and GLAM institutions, contributing to the collective advancement of digital strategies in heritage preservation.

4. CURRENT STATUS AND EXPECTED RESULTS

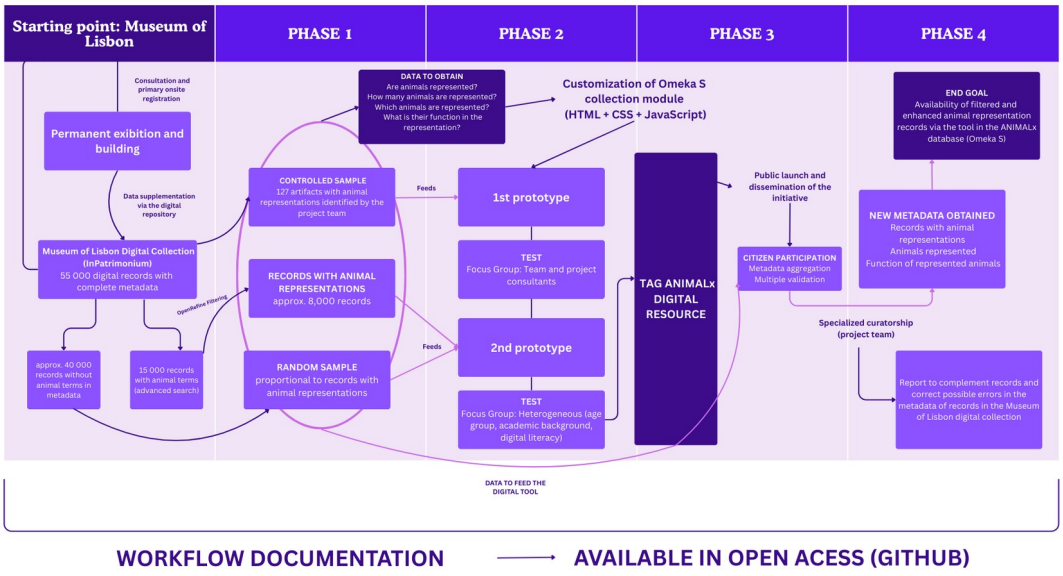


Figure 2. Tag ANIMALx Workflow Diagram. The process outlines the four project phases: from data diagnosis at the Museum of Lisbon, through prototype development and testing, to public launch and validation.

The project is currently in the prototype development phase. The preliminary usability tests will be conducted with the control sample (127 records), indicating that the gamified interface significantly lowers the entry barrier for non-specialists compared to traditional forms. In this year's first semester, we anticipate the processing of a broader dataset of approximately 2,000 records, including a randomized set of items to test for "negative results" (images where users confirm no animal is present). This will generate a structured corpus of "animal traces" that were previously unsearchable. The expected results are twofold: scientifically, the creation of a dataset that enables quantitative analysis of animal presence in Lisbon's history; and museologically, the demonstration that standard inventory descriptions can be enriched at scale without overburdening institutional staff.

While the current contribution focuses on the system architecture and the preliminary results are limited to the control sample, a robust evaluation framework is being designed. Future empirical evaluation will measure two main axes: Firstly, the data reliability, by quantitatively comparing crowdsourced tags against the expert 'ground truth'; Secondly, user engagement, by tracking retention rates and the efficacy of specific gamification mechanics over time.

5. CONCLUSION AND LEGACY

In conclusion, Tag ANIMALx proposes a shift in how we envisage digital heritage: from static display to active co-construction. By integrating Citizen Science methods with rigorous historiographical inquiry, the tool validates the public's role in producing knowledge about the "Animal Past," effectively socializing the archive. Looking forward, making the historical 'animal trace' computable and interoperable positions projects like Tag ANIMALx as foundational steps to reach broader European infrastructures, such as the ECHOES-ECCCH initiative (European Commission, 2024), contributing to the creation of rich, multispecies digital twins within the cultural heritage domain.
Crucially, the project is committed to sustainability and Open Science principles. A recurring issue in DH is the obsolescence of tools once a project's funding ceases. To combat this, the entire workflow, source code, and documentation for Tag ANIMALx will be released under a Creative Commons license in Open Access, ensuring that the model is replicable by other GLAM institutions, offering a sustainable roadmap for the "humanization"—and "animalization"—of digital archives.
Ultimately, Tag ANIMALx aims to prove that technology, when guided by ethical commitment and methodological rigor, can bridge the gap between society and its history, giving visibility to the non-human agents that helped shape it.

ACKNOWLEDGEMENTS

AUTHOR CONTRIBUTIONS:
The Tag ANIMALx development is being carried out as Daniela Teixeira Gomes Master's Project in Digital Curation and Digital Humanities at NOVA FCSH, therefore she served as the lead author, responsible for the conceptualization, methodology, software adaptation, and writing of the original draft of this communication. Joana Vieira Paulino supervised the entire research and development workflow and contributed to the review and editing of this proposal. Carla Vieira, as one of the ANIMALx project researchers and responsible for the development and management of ANIMALx's database, also contributed to the supervision and critical review of this proposal.
FUNDING:
This study was funded by FCT through the project ANIMALx - Animals of Lisbon: Tracking the presence and role of (non-human) animals in the history and landscape of the city (2023.12750.PEX — https://doi.org/10.54499/2023.12750.PEX). This research also received support from CHAM (NOVA FCSH / UAc), through the strategic project sponsored by FCT (UID/04666/2025 — https://doi.org/10.54499/UID/04666/2025).
The IHC is funded by national funds through FCT — Fundação para a Ciência e a Tecnologia, I.P., under the projects UID/04209/2025 (DOI: https://doi.org/10.54499/UID/04209/2025) and LA/P/0132/2020 (DOI: https://doi.org/10.54499/LA/P/0132/2020).

DECLARATION OF GENERATIVE AI USE:
The authors acknowledge the use of generative artificial intelligence (DeepL) to assist in the translation and linguistic refinement of this proposal from Portuguese to English and Italian, as well as the translation of figure contents from Portuguese to English. The authors have reviewed and edited the output as needed and take full responsibility for the content of the publication.

REFERENCES

Balázs, B., Mooney, P., Nováková, E., Bastin, L., & Arsanjani, J. J. (2021). Data Quality in Citizen Science. In K. Vohland et al. (Eds.), *The Science of Citizen Science* (pp. 139–157). Springer International Publishing. https://doi.org/10.1007/978-3-030-58278-4_8

Bonnell, J., & Kheraj, S. (2022). Introduction: Traces of the Animal Past. In J. Bonnell & S. Kheraj (Eds.), *Traces of the Animal Past: Methodological Challenges in Animal History*. University of Calgary Press.

Causer, T., & Terras, M. (2014). Crowdsourcing Bentham: Beyond the Traditional Boundaries of Academic History. *International Journal of Humanities and Arts Computing*, 8(1), 46–64. https://doi.org/10.3366/ijhac.2014.0119

Doerr, M., Hunter, J., & Lagoze, C. (2003). Towards a core ontology for information integration. *Journal of Digital Information*, *4*(1).

European Commission. (2024). *European Cloud for Heritage OpEn Science (ECHOES)*. CORDIS EU research results. https://cordis.europa.eu/project/id/101157364

Fudge, E. (2017). What Was It Like to Be a Cow? History and Animal Studies. In L. Kalof (Ed.), *The Oxford Handbook of Animal Studies* (pp. 258–278). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199927142.013.28

Heinisch, B. (2021). Citizen Humanities. In K. Vohland et al. (Eds.), *The Science of Citizen Science* (pp. 97–114). Springer International Publishing. https://doi.org/10.1007/978-3-030-58278-4_6

Laamarti, F., Eid, M., & Saddik, A. E. (2014). An Overview of Serious Games. *International Journal of Computer Games Technology*, 2014(1), 358152. https://doi.org/10.1155/2014/358152

Page, R. D. M. (2019). Biodiversity informatics: The challenge of linking data and the role of shared identities. *Briefings in Bioinformatics*, *20*(2), 653–668. https://doi.org/10.1093/bib/bbx153

Rangarajan, M. (2013). Animals with rich histories: The case of the lions of Gir Forest, Gujarat, India. *History and Theory*, *52*(4), 109–127. https://doi.org/10.1111/hith.10690

Ridge, M. (2013). From Tagging to Theorizing: Deepening Engagement with Cultural Heritage through Crowdsourcing. *Curator: The Museum Journal*, 56(4), 435–450.

From perception to memory: Public Engagement and Neurohumanistic Approaches to the Collective Reconstruction of Archaeological Heritage

GRAZIA SOLENNE[1], VINCENZA FERRARA[2], ELISA CORRÒ[3]

[1] CA' FOSCARI UNIVERSITY OF VENICE, GRAZIA.SOLENNE.RESEARCH@GMAIL.COM
[2] SAPIENZA UNIVERSITY OF ROME, VINCENZA.FERRARA@UNIROMA1.IT
[3] CA' FOSCARI UNIVERSITY OF VENICE, ELISA.CORRO@UNIVE.IT

ABSTRACT (ENGLISH)

This paper aims to document a process that, starting from experimental research, led to the definition and application of *Digital Humanities and Public Engagement* practices through the integration of an interdisciplinary methodology that includes neurohumanistic approaches and public archaeology tools. The study focuses on research activities conducted and directed by Ca' Foscari University of Venice at the archaeological site of the monastery of Santi Ilario e Benedetto in Mira (Venice, Italy), conducted between 2023 and 2024. The archaeological context, characterised by a strong connection with the local community, makes the reconstruction of collective memory a determining factor in enhancing cultural heritage. The project focused on the cognitive and perceptual aspects of heritage experience and applied eye-tracking and *Visual Thinking Strategies (VTS)* methodologies to develop and analyse perception, attention, and comprehension within an archaeological excavation context.
Based on the collected data, an operational protocol for participatory enhancement was developed and applied to dissemination activities and to a series of public engagement events held during the 2024 excavation campaign. For the latter, the initiative focused on the movement of archaeological artefacts associated with the monastery, which are considered artefacts of collective memory. Being currently conserved in the National Archaeological Museum of Venice (MAV), it was possible to investigate the change in meaning produced by this shift. In conclusion, the study demonstrates how the development of more constructive, shared, and inclusive archaeological and historical knowledge can be generated through the systemic integration of digital technology, experimental methodologies, and participatory research frameworks.
Keywords: cultural memory; digital humanities; public archaeology; visual thinking strategies; neurohumanities.

ABSTRACT (ITALIANO)

Dalla percezione alla memoria: public engagement e approcci neuroumanistici per la ricostruzione collettiva del patrimonio archeologico.
L'articolo si propone di documentare un processo che, a partire da una ricerca sperimentale, ha permesso di definire e applicare pratiche di *Digital Humanities e di Public Engagement*, attraverso l'integrazione di una metodologia interdisciplinare che include un approccio neuroumanistico e strumenti di archeologia pubblica. Il progetto si concentra sulle attività di ricerca condotte e dirette dall'Università Ca' Foscari di Venezia presso il sito archeologico del monastero di S. Ilario e Benedetto a Mira (Venezia, Italia), condotte tra il 2023 e il 2024. Il contesto archeologico, caratterizzato da una forte connessione con la comunità locale, pone la ricostruzione della memoria collettiva quale fattore determinante per la valorizzazione del patrimonio culturale. Il progetto si è concentrato sugli aspetti cognitivi e percettivi dell'esperienza del patrimonio e ha adottato l'*eye-tracking* e le *Visual Thinking Strategies (VTS)* come metodologie per sviluppare e analizzare la percezione, l'attenzione e la comprensione nel contesto dello scavo archeologico. Sulla base dei dati raccolti è stato elaborato un protocollo operativo di valorizzazione partecipata, applicato alle attività di divulgazione e a una serie di eventi di coinvolgimento del pubblico svoltisi durante la campagna di scavo del 2024. Per quest'ultimo, l'iniziativa si è concentrata sullo spostamento dei reperti archeologici associati al monastero, considerati artefatti della memoria collettiva. Essendo attualmente conservati presso il Museo Archeologico Nazionale di Venezia (MAV), è stato possibile indagare il cambiamento di significato prodotto dal trasferimento. In conclusione, lo studio dimostra come lo sviluppo di conoscenze archeologiche e storiche più costruttive, condivise e inclusive

possa essere generato attraverso l'integrazione sistemica tra tecnologia digitale, metodologie sperimentali e framework di ricerca partecipativa.
Parole chiave: memoria culturale; public engagement; archeologia pubblica; visual thinking strategies; neuroumanistica.

1. INTRODUCTION

In recent decades, the literature on digital and cultural heritage studies has demonstrated the central role of public engagement and community involvement in heritage enhancement. Within this broader framework, the importance of community engagement in archaeological practice is further reinforced by Manacorda's statement (Manacorda, 2008, 258-259, author's translation from Italian), "Archaeology speaks for everyone, [...]". This vision, which sees archaeology as a means to develop a sense of equality in diversity, further legitimises the development of public engagement practices in the archaeological context. An open, dynamic, and layered view of cultural heritage can be proposed through archaeological research instruments, placing citizens at the centre of heritage knowledge and restoring sense (Volpe, 2020). The integration of digital technology has created new opportunities for public engagement in public and community archaeology (Bonacchi, 2017). In this context, Corrò (2021) highlights the positive impacts of adopting a citizen-science approach to foster cooperation among institutions, researchers, and citizens, supported by digital tools. Digital storytelling, narrative construction, and design thinking facilitate awareness of the territory's cultural value and a sense of belonging within the community (Corrò, 2021). While a participatory and digital approach emphasises the social aspect of heritage public engagement, an interdisciplinary approach, shaped by neurohumanities, enables the examination of cognitive aspects of these experiences (Corrò *et al.*, 2025), investigating concepts such as cultural memory and perceptions as mechanisms through which users and communities engage with cultural heritage (Corrò *et al.*, 2024). Cultural memory is a key notion at the intersection of neuroscience approaches and heritage studies, as it represents a social process through which a community constructs its cultural identity by selecting, transmitting, and reshaping the meaning of the past within the present. At the same time, heritage can be understood as meaning constructed and transmitted across generations through interpretive and participatory practices (Corrò, 2021). This process helps strengthen cultural identity, operating through external supports such as objects, places, monuments, institutions, and texts. In this context, the concept of the "places of memory" can represent an entire landscape capable of activating cultural memory (Assmann, 2011). In the archaeological environment, Malafouris & Renfrew (2008) propose adopting a framework of material engagement, moving beyond a neurocentric approach and conceiving cognition as an element of interaction among the brain-body-world system, modelled by material culture. This research, grounded in this theoretical framework, aims to document the interdisciplinary research process undertaken to illustrate how public engagement initiatives, adhering to neuroarchaeological principles and prioritising cognitive and perceptual experiences, can significantly enhance the local community's interaction with their culture. The contribution stems from prior research on local community engagement and perceptions of cultural heritage, applying a neurohumanistic methodological approach. This framework was inspired by the results of Neurocities and Ruinscapes: Reconstructing Ancient Cities and Ruins Using Virtual Reality (2023–2024) project, which demonstrates the effectiveness of experimental tools for analysing attention, perception, and engagement (Forte *et al.*, 2025). On this basis, the methodology was adapted to the context of the 2023 archaeological excavation of the monastery of Santi Ilario e Benedetto, in the Municipality of Mira (VE), to analyse how the local community perceives and interprets archaeological ruins and how that experience relates to sensitivity to heritage and the construction of cultural identity. To conclude, the study's contribution lies in integrating digital tools for data collection, interpretative and quantitative analysis, and public dissemination. These technologies act as core methodological components that connect experimental observation with the quantitative analysis of eye-tracking data, the digital data collection and interpretive analysis of Visual Thinking Strategies responses supported by topic modeling, and participatory heritage communication.

2. ARCHAEOLOGICAL CONTEXT: THE MONASTERY OF SANTI ILARIO AND BENEDETTO

The monastery of Santi Ilario e Benedetto (9th century AD) represents a context of strong cultural identity and historical importance, as it symbolises a founding institution of the nascent Venice. In 2023 and 2024, two excavation campaigns were conducted at the site under the scientific direction of Prof.

Sauro Gelichi (Ca’ Foscari University of Venice), with technical direction by Dr. Alessandro A. Rucco (Ca' Foscari University of Venice) and under the executive supervision of Dr. Elisa Corrò (Ca’ Foscari University of Venice). The excavation campaigns revealed significant historical foundations, bringing to light noteworthy features, including the foundations of some of the pillars of the famous 13th-century triapsidal church, which were excavated at the end of the 19th century under the direction of Marquis Lorenzo Saibante. Until now, only intangible traces of this church had survived, including a remarkable historical photograph from the period. The investigations also highlighted an earlier, smaller three-aisled church, the substantial foundations of which, built from large stone blocks, have been preserved. To conclude, during the most recent campaign, the discovery of a large cemetery area highlighted the extensive occupation of the site and confirmed the presence of a wide and long-lasting settlement.
The archaeological materials from the monastery of Santi Ilario e Benedetto shed light on the site's history and cultural significance, as well as on the long journey of excavation, restoration, relocation, and preservation. The remains, including mosaics, sarcophagi, and architectural elements, come from excavations conducted in the late 19th century, when buildings survived only as ruins. The archaeological findings were transferred to Venice, where they were first displayed at the Civic Museum and later permanently housed in the National Archaeological Museum (MAV) (Moine *et al.*, 2017).

3. FROM RESEARCH TO PROTOCOL

The preliminary research pursued multidisciplinary objectives by connecting two complementary methodologies: Visual Thinking Strategies (VTS) and eye-tracking. Rooted in constructivism and pedagogical principles, Visual Thinking Strategies is a methodology that uses analysis and discussion activities regarding the perception of an image, in this case an archaeological object, to improve observation, communication, and interpretation skills as active cognitive tools to develop a shared meaning. VTS not only encourages participants to discuss what is being observed but also facilitates constructive dialogue among them, develops visual literacy and critical thinking, and fosters engagement at the emotional and perceptual levels, improving the relationship between cultural artefacts, participants’ perceptions, and collective understanding (Hailey *et al*., 2015; Ferrara, 2023). Eye-tracking, in turn, enables the investigation of visual attention and perceptual behaviour by collecting quantitative process-level data, such as saccades, fixations, and gaze patterns, during interactions with artefacts, exhibitions, and archaeological sites (Banzi, 2023; Forte *et al*., 2023).
Building on this basis, the methodological approach combined VTS to foster collective meaning-making and eye-tracking to provide data on visual attention processes. Questionnaires facilitated the gathering of socio-demographic, emotional, and evaluative information, allowing for a multi-layered understanding of participants’ experiences. The experimental protocol was composed of two data collection sessions. Group A (n = 8), combined questionnaires and VTS with eye-tracking: before expert explanation, participants completed a registration questionnaire (Q1), carried out a first VTS session (VTS-A), and completed an emotions/aesthetic judgment questionnaire (Q2). After attending the guided geologists’ and archaeologists’ explanation of the site, they performed a second VTS session (VTS-B) and finally undertook eye-tracking recording with the Pupil Invisible eye tracker. The second data collection, composed of group B (n=15), was conducted during a public excavation visit and, due to organisational constraints, excluded VTS: eye-tracking was recorded before any explanation, followed by Q1, Q2, and the final feedback form. An additional eye-tracking session was conducted with archaeologists (n=5) to compare experts and non-experts. Each data collection session concluded with a final survey to gather feedback and assess local community's attitudes toward heritage enhancement. Informed consent sections were included in each questionnaire to ensure confidentiality and anonymous use of the data for research purposes (*see infra*). The registration questionnaire provided information on participants’ familiarity with the archaeological context and their sociodemographic and territorial characteristics, indicating a predominantly adult and mature audience, largely from the localities surrounding the archaeological excavation in the province of Venice. Regarding eye-tracking, each recording fixed participants to a single observation point with a predefined field of view, asking them to observe the archaeological excavation for one minute. *For group A,* two participant recordings were produced at two different points in the excavation*. For pos*t-production lighting issues, it was possible to rework the data into a heatmap for only one point in the excavation. For group B, which was divided into three shifts, separate heatmaps were produced for each time interval, as lighting varied between time slots.

Visual Thinking Strategies, included in the first investigation, supported the analysis of cognitive engagement and observational skills. The VTS prompts were adapted to the specific context of the excavation tests. The protocol[1] was structured around the following questions, to which each visitor had to answer by filling out a form within the museum space or excavation:

(q.1) Observe: What is going on in front of you?
(q.2) What especially strikes you?
(q.3) What do you wonder about? Is there anything you would like to know about what you are observing?

The VTS method is usually carried out orally with a group discussion led by a facilitator; however, previous experiences have shown that this method can also help collect visitor comments useful for better designing the visit route. During VTS sessions, the involvement of neutral and impartial facilitators played an essential role in supporting participants. The decision to implement VTS before and after the guided explanation of the excavation context provided data to assess the evolution of cognitive and perceptual attention levels, as well as the impact of this methodology on participants' emotional involvement. An evaluation grid was implemented to investigate the number of identified elements, the number of words, and the attention paid to the excavation. To identify recurring lexical patterns and thematic clusters in VTS responses, an exploratory topic modeling analysis in Python was applied on a subset of VTS-A and VTS-B responses. After basic text preprocessing, the responses were analysed in Italian, their original language, to preserve interpretive reliability and the results were visualised through word clouds and an intertopic distance map (Solenne, 2024). Questionnaire 2, focused on the emotional and aesthetic judgement of the experience, assessed valence, arousal, beauty, and liking, referring to the scale proposed by Van Paasschen *et al.* (2015) and to the questions developed within the *Neuroartifact* project (Forte *et al.*, 2025), in order to make the instrument more aligned with the archaeological context. The analysis of mean values revealed that both groups rated the experience well in terms of aesthetic and preference criteria, while the emotional impact (arousal) was significantly greater for participants in the first trial (group A with VTS). Overall, the preliminary results on VTS data demonstrate a significant perceptual and cognitive change before and after expert explanation. The initial uncertain and probabilistic description (VTS-A) becomes rich in technical and taxonomic terms (VTS-B). At the lexical level, if participants' greater curiosity emerges in the first VTS session, the session following the explanation demonstrates a thematic change and more structured questions and sentences. Furthermore, words such as "valorisation", "reconstruction", "public", "documentation", and "community" emerge more clearly, highlighting the impact of the experience on the critical thinking of the participants (*Figure 1*). Moreover, the VTS proved to be effective in fostering debate among local community members, who, after the initial VTS session, raised numerous questions to the experts (Solenne, 2024).



Figure 1. (left) word frequency cloud from VTS-B, q.3 (Solenne, 2024)

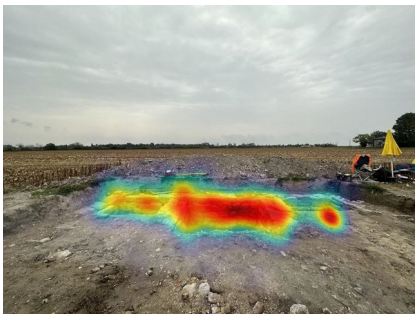


Figure 2. (right) eye-tracking heatmap, group A (Solenne, 2024)

The analysis of eye-tracking data, based on the comparison of heatmaps generated in Pupil Cloud software, shows a strong concentration on the central core of the observed area. In light of VTS responses, this focal area corresponds to participants' attention to stratigraphy, materials and structures

[1] The questions and protocol were defined by Vincenza Ferrara.

emerging from the excavation (*Figure 2*). Analysis of group B eye-tracking data indicates a strong impact of external conditions on perception. Furthermore, the comparison between non-experts and archaeologists indicates different perceptual tendencies, with non-experts focusing on the centre of the scene, and compared to specialists, demonstrating a more selective and detail-oriented observation. Furthermore, the evidence emerging from Q1 and the final feedback questionnaire provided insights into understanding the perceived value of heritage and the identity of the local community, which declares a strong identification with its territory and responds positively to the possibility that the history of the excavations will be narrated, highlighting the need to reconstruct cultural and territorial memory. The arguments outlined show interest in the site, local history, and landscape, as well as in aspects related to the territory. As already demonstrated by Q2 data, group A, which conducted the VTS sessions, highlights the emotional impact and connection with the territory, while group B emphasises what they learned during the visit and their own knowledge (Solenne, 2024). Given the limited size of the sample, the first 2023 experiment was conceived as an exploratory and experimental study. In this sense, its main strengths lie not in the scale of the dataset but in the integration of different methodologies and in their application in a real field context, despite the numerical limitations and operational constraints encountered. These findings supported the development and implementation of an operational protocol to investigate cognitive responses and foster public engagement through the use of Visual Thinking Strategies (VTS) and guided storytelling. As part of the 2024 public engagement programme, visitors recruited through an online form were divided into two groups: one group participated in VTS-based activities following the same questions adopted in the 2023 study (*see above*), while a control group followed a standard visit without VTS. This structure enabled comparison of outcomes and evaluation of the methodology's impact on narrative-based visiting paths. For both groups, informed consent was collected, together with questionnaires assessing prior knowledge of art and archaeology (Q1), as well as aesthetic and emotional responses (Q2). The protocol was carried out in two primary settings: the National Archaeological Museum of Venice (MAV), where 8 out of 15 participants took part in the VTS activities (VTS-A), and the excavation site, where 15 out of 30 visitors completed the VTS-based activity (VTS-B)[2]. The narrative framework centres on the archaeological materials currently preserved in the internal courtyard of the National Archaeological Museum of Venice and on their movements after excavations conducted in the 19th century. Following the VTS session, the remaining mosaics, sarcophagi, and structural elements of the ancient monastery are read as "places of memory," capable not only of recounting the artefacts' history but also of reshaping their symbolic and cultural significance. Afterwards, VTS activities were conducted during the open days at the new excavation site, followed by the archaeologists' narration and discussion with the public. Participants were also asked, through an online form, to produce narratives about the artifacts' history and movement, rephrasing and imagining what might have happened to the materials.

4. RESULTS AND CONCLUSION

The 2023 preliminary results analysis provided key insights into the operational valorisation activities conducted in 2024. Specifically, the community's strong connection to local heritage, the need to enhance accessibility and understanding of the excavation site and the possibility of greater active participation provided an operational framework for designing public engagement activities. In July 2024, the exhibition "*Memories of the Lagoon: from its medieval origins to the Archaeological Discoveries at the site of Sant'Ilario*" presented a narrative journey through the history of the site, retracing the excavations conducted between the 19th and 21st centuries and highlighting the most recent archaeological investigations through a digital storytelling approach[3]. More specifically, the exhibition reconstructed the history of the monastery and retraced the archaeological investigations through explanatory panels, QR codes, and videos. One video presented the history of the excavation and highlighted the main results of the 2023 campaign, while another narrated the broader history of the monastery. Together, these materials made the history of the site and its excavations more accessible to the public.

[2] Only 6 non-VTS participants completed both Q1 and Q2 (5 for the MAV activity, 1 for excavation activity); therefore, the comparative reading of the non-VTS group should be considered preliminary.
[3] Developed under the scientific guidance of Prof. Sauro Gelichi and Dr. Alessandro A. Rucco, and coordinated by Dr. Elisa Corrò with the support of Dr. Grazia Solenne, in collaboration with the Associazione Gruppo La Malcontenta and the Parrocchia S. Ilario.

In December 2024, this narrative approach continued with the program of events "*From Excavation to Storytelling – Dallo scavo alla narrazione*", dedicated to the rediscovery of the monastery through a dialogue between archaeology, history, and imagination[4]. The initiative culminated in the exhibition "*Storie dal monastero dei Santi Ilario e Benedetto*". Here, archaeological knowledge was interwoven with stories created by students, offering a renewed and participatory perspective on the site and its cultural heritage, combining scientific research and creative interpretation to reactivate the memory of the place for a wider community. Moreover, the preliminary analysis of the 2024 implementation of the operational protocol produced several interesting results. At the National Archaeological Museum of Venice (MAV), the VTS activity focused on the objects on display, with curiosity about their provenance, decoration, and history. The observation of artefacts called for a key to understanding the monastery's history, with an implicit demand for greater valorisation and the development of learning (*Figure 3*). During the open excavation experience, the focus shifts to the actions of archaeologists, rather than the recognition and interpretation of structures. Greater interpretive difficulty and a need for mediation tools emerge to better understand and interpret the observed data (*Figure 4*). The narratives also highlight concrete actions related to valorisation opportunities, namely, active involvement during visits, the possibility of enhancing the use of VTS as a tool to enhance citizen collaboration and facilitate a shared vision of the site's history, and the possibility of more explicitly connecting the finds exhibited at the MAV to the excavation. Finally, it was possible to further assess the impact of VTS on participants' perception of comfort, an element that positively affects the visit and facilitates more active participation.



Figure 3. (left) word frequency cloud from VTS-A, q.3



Figure 4. (right) word frequency cloud from VTS-B, q.3

To conclude, this study demonstrates how the intersection of neurohumanistic approaches with experimental and interdisciplinary methodologies that connect public engagement and public archaeology can provide concrete tools for the collective reconstruction of local cultural heritage. The combination of VTS and eye-tracking as methodologies highlights how cognitive and perceptual aspects can positively impact the collective and individual interpretation of heritage, leading to the development of critical thinking and cultural memory. The development of the operational protocol confirms the importance of adopting data-oriented public engagement practices to foster greater community involvement and collective awareness. Beyond this particular case study, the protocol offers a repeatable framework that can be adapted to other cultural heritage settings, shaped by community participation and complex, layered memory processes. In conclusion, the findings suggest future directions for developing heritage valorisation practices that prioritise cognitive and perceptual aspects of the visitor experience, fostering active, conscious participation of the local community, which becomes the protagonist and promoter of its own "places of memory".

ETHICS STATEMENT

An informed consent form was developed for participants taking part in the study program. Participation was voluntary, and participants were informed of their right to withdraw from the study at any time without penalty. Confidentiality and anonymity were guaranteed, and all data were securely stored and processed in accordance with applicable ethical and data protection regulations. All information collected during this research was used exclusively for scientific and educational purposes.

[4] Curated by Dr. Elisa Corrò, Prof. Marco Giraldi, Dr. Alessandro A. Rucco, and Dr. Grazia Solenne, and developed in collaboration with the lower secondary school San Domenico Savio of Oriago.

ACKNOWLEDGEMENTS

The project was carried out in collaboration with several institutional and community partners, whose support was fundamental to its development and realisation. We gratefully acknowledge the *Comune di Mira; Università Ca' Foscari Venezia, Dipartimento di Studi Umanistici; Venice Centre for Digital and Public Humanities; Fondazione Università Ca' Foscari Venezia; Soprintendenza Archeologia, Belle Arti e Paesaggio per il Comune di Venezia e Laguna; Centro Studi Archeologia Venezia; Associazione Culturale Gruppo "La Malcontenta"; Parrocchia di S. Ilario Vescovo, Malcontenta (VE); Città di Venezia – Le Città in Festa; and Venezia Unica*. We would like to express our sincere gratitude to Prof. Sauro Gelichi for his scientific guidance and to Dr. Alessandro A. Rucco for their valuable contribution to the project. We also wish to thank Prof. Federico Bernardini (supervisor) and Nevio Danelon (co-supervisor) of the thesis in Digital and Public Humanities for its support throughout the research. Special thanks go to the lower secondary school San Domenico Savio of Oriago, to the teachers who supported the initiative, and to all the students whose creativity and participation gave a new voice to the history of the monastery. This project was supported by funds from *Università Ca' Foscari Venezia* and the *Comune di Mira*.

AUTHOR CONTRIBUTION

Grazia Solenne contributed to data collection, quantitative analysis, and visualisations.
Vincenza Ferrara contributed to methodological design and the VTS framework.
Elisa Corrò contributed to conceptualisation, supervision, and project administration.
All authors contributed to manuscript preparation and approved the final version.

REFERENCES

Assmann, J. (2011). *Cultural memory and early civilization: Writing, remembrance, and political imagination.* Cambridge University Press. https://doi.org/10.1017/CBO9780511996306.

Solenne, G. (2024*). Public archaeology through the lens of neuroarchaeology: Exploring emotional, cognitive responses and heritage significance in Rome and Mira (Venice) via eye tracking and visual thinking strategies* (master's thesis). Università Ca' Foscari Venezia. Supervisor: Bernardini, Federico; co-supervisors: Corrò, Elisa; Danelon, Nevio. https://hdl.handle.net/20.500.14247/1949.

Banzi, A. (Ed.). (2023). *The Brain-Friendly Museum: Using Psychology and Neuroscience to Improve Visitor Experience.* Routledge, Taylor & Francis. https://doi.org/10.4324/9781003304531.

Bonacchi, C. (2017). Digital media in public archaeology. In G. Moshenska (Ed.) *Key concepts in public archaeology* (pp. 60-72). https://doi.org/10.14324/111.9781911576419.

Moine, C., Corrò, E., & Primon, S. (2017). *Paesaggi artificiali a Venezia. Archeologia e geologia nelle terre del monastero di Sant'Ilario tra alto Medioevo ed Età Moderna*. All'Insegna del Giglio.

Corrò, E. (2021). Heritage Thinking. Approaching Communication in the Venice Countryside. Knowledgescapes, *Insights on Public Humanities. Filologie medievali e moderne*, 25. https://hdl.handle.net/10278/5002074.

Corrò, E., Danelon, N., Forte, M., & Ferrara, V. (2024). La decifrazione dei paesaggi con rovine nelle vedute degli artisti del Settecento: un approccio neuro-umanistico attraverso l'intelligenza artificiale. In *Italian* Association for Urban History (Eds.), *Beyond the gaze: Interpreting and understanding the city / Oltre lo sguardo: Interpretare e comprendere la città* (XI AISU Congress, Ferrara, September 13–16, 2023) (pp. 1008–1010). ISBN 978-88-31277-08-2.

Corrò, E., Danelon, N., Forte, M., Ferrara, V., Solenne, G. (2025). From Observation to Engagement: Neuro-Humanistic and Visual Thinking Strategies Applied to Cultural Landscapes. In *Beyond the Gaze. Interpreting and understanding the city. Oltre lo sguardo. Interpretare e comprendere la città. XI Congresso AISU / 11th AISU Congress*, (pp. 369-380). Congress PROCEEDINGS, AISU International, Convegno: Beyond the gaze. Interpreting and understanding the city. Oltre lo sguardo. Interpretare e comprendere la città XI Congresso AISU / 11th. ISBN: 978-88-31277-11-2.

Ferrara, V. (2023). The Museum and Quality of Life. In A. Banzi (Ed.) *The Brain-Friendly Museum* (pp. 118-129). Routledge, Taylor & Francis. https://doi.org/10.4324/9781003304531.

Forte, M., White, L. E., Straneva, K., & Woytowitz, S. (2023). Virtual Reality and Neuroarchaeology: Visual Perception and Cognition of an Archaeological Excavation. In A. Banzi (Ed.), *The Brain-*

Friendly Museum (pp. 147-158). Routledge, Taylor & Francis. https://doi.org/10.4324/9781003304531.

Forte, M., Ferrara, V., Mingione, M., Alaimo di Loro, P., Giorgi, A., Menicocci, S., Babiloni, F., Iosa, M. (2025). The NeuroARTifact Project: Seeing with the mind. In M. Forte (Ed.), *A new Etruscan archaeology: Twenty-first century techniques and methods* (online ed.). Oxford Academic. https://doi.org/10.1093/9780197582053.003.0007.

Hailey, D., Miller, A., & Yenawine, P. (2015). Understanding visual literacy: The visual thinking strategies approach. In Baylen, D., D'Alba, A. (Eds) *Essentials of Teaching and Integrating Visual and Media Literacy* (pp. 49-73). Springer, Cham. https://doi.org/10.1007/978-3-319-05837-5_3.

Malafouris, L., & Renfrew, C. (2008). Introduction. *Cambridge Archaeological Journal*, 18(3), 381–385. https://doi.org/10.1017/S0959774308000425.

Manacorda, D. (2008). *Lezioni di archeologia* (pp. 258-259). Editori Laterza.

Van Paasschen, J., Bacci, F., & Melcher, D. P. (2015). The influence of art expertise and training on emotion and preference ratings for representational and abstract artworks. *PloS one*, 10(8), e0134241. https://doi.org/10.1371/journal.pone.0134241.

Volpe, G. (2020). *Archeologia pubblica. Metodi, tecniche, esperienze.* Carocci editore. https://hdl.handle.net/11586/258068.

Comunicare la ricerca archeologica nell'ecosistema digitale: pratiche, metodi e sperimentazioni su Instagram

MARCO DEMURU
UNIVERSITÀ DI BARI, M.DEMURU@PHD.UNIBA.IT

ABSTRACT (ITALIANO)

Il contributo presenta una ricerca, tuttora in corso, sulla condivisione delle indagini archeologiche sui social media e, nello specifico, sulla piattaforma Instagram, che mira a indagare il rapporto tra archeologia e comunicazione digitale, con particolare attenzione al ruolo dei social network nella valorizzazione delle attività di scavo e nella diffusione dei risultati scientifici. A partire da un caso studio specifico, il progetto prende forma dalla creazione e gestione del profilo Instagram @archeosantelia con l'idea di superare la narrazione del diario di scavo, facendo dell'account non una semplice pagina di aggiornamento sull'avanzamento delle indagini archeologiche sul Capo Sant'Elia, ma uno spazio per condividere la ricerca in modo aperto e partecipato, fondato sul dialogo e sul confronto tra i diversi soggetti che hanno un legame con il sito. Per dare concretezza a questo approccio, gli stakeholder sono coinvolti attivamente attraverso la compilazione di un breve form di domande, appositamente strutturato, che raccoglie temi e punti di vista. Queste informazioni costituiscono la base per i contenuti da condividere, trasformando la pagina Instagram in una bibliografia condivisa del sito, capace di restituire la pluralità delle voci e dei significati che animano il patrimonio archeologico di Capo Sant'Elia. Lo scopo del progetto è quello di sperimentare e valutare la definizione di un prototipo di comunicazione social della ricerca archeologica concepito come modello metodologico ripetibile e adattabile a contesti differenti, capace di coniugare rigore scientifico, accessibilità dei contenuti e coinvolgimento attivo dei pubblici al fine di conseguire una più efficace disseminazione delle ricerche archeologiche, intesa non come semplice diffusione dei risultati, ma come processo continuo di condivisione e dialogo e in grado di ottimizzare l'uso consapevole dei principali strumenti di comunicazione digitale.
Parole chiave: Archeologia pubblica; Social network; Comunicazione digitale della ricerca; Coinvolgimento dei pubblici; Modelli di condivisione della ricerca

ABSTRACT (ENGLISH)

Archaeological Research Communication in the Digital Ecosystem: Practices, Methods, and Instagram-Based Experimentation.
This contribution presents an ongoing research project focused on the use of social media, specifically Instagram, for sharing archaeological investigations. The study explores the relationship between archaeology and digital communication, with particular attention to the role of social networks in enhancing excavation activities and disseminating scientific results. Based on a specific case study, the project is developed through the creation and management of Instagram profile @archeosantelia. The aim is to move beyond the traditional excavation diary format, rethinking the account not as a simple update tool on the progress of fieldwork on Capo Sant'Elia, but as an open and participatory space for sharing research. This approach is grounded in dialogue and exchange among the various actors connected to the site. To give concrete form to this approach, stakeholders are actively involved through a short and structured questionnaire designed to collect themes, perspectives, and viewpoints. These contributions form the basis of the published content, transforming the Instagram page into a shared bibliography of the site that reflects the plurality of voices and meanings shaping archaeological heritage of Capo Sant'Elia.
The project aims to test and evaluate a prototype model for the social media communication of archaeological research. Conceived as a repeatable and adaptable methodological framework, the model seeks to combine scientific rigor, content accessibility, and active public engagement. Dissemination is therefore understood not as the simple transmission of results, but as an ongoing process of sharing and dialogue, supporting a more effective and conscious use of digital communication tools.
Keywords: Public archaeology; Social networks; Digital research communication; Public engagement; Models of research sharing

1. TECNOLOGIE DIGITALI E RIDEFINIZIONE DEI PROCESSI CONOSCITIVI

Viviamo oggi all'interno di una cultura pienamente digitale, frutto di una lunga traiettoria di trasformazioni tecnologiche che incidono in modo decisivo sulla società odierna.
Spesso descritta come erede delle grandi rivoluzioni del passato, la rivoluzione digitale si presenta al tempo stesso innesco di nuovi processi di cambiamento in un momento storico che si va caratterizzando sempre più per mutamenti profondi e repentini (Balbi 2021, p. 28).
L'innovazione, però, non riguarda esclusivamente la dimensione tecnologica, ma investe i comportamenti, le aspettative e il modo di vivere la *digital life*. A fronte della natura eminentemente tecnico-scientifica dei suoi dispositivi e delle sue procedure, l'informatica contemporanea appare attraversata da una diffusa fascinazione di carattere quasi magico-misterico, in cui i sistemi informatici tendono a essere percepiti come entità autorevoli a cui l'utente si affida, spesso in modo passivo, con atteggiamenti che sfiorano una nuova forma di devozione rivolta al digitale. L'onnipresenza di dati nelle nostre vite costituisce uno degli aspetti più evidenti di questa narrazione e, al tempo stesso, uno dei più rilevanti, poiché contribuisce a rafforzare l'idea di un sapere automatizzato e incontestabile (Previtali 2022, p. 43).
La tradizione italiana dell'informatica umanistica ha storicamente rifiutato una visione meramente strumentale del digitale, privilegiando invece un approccio epistemologico fondato sull'idea dell'informatica come scienza teorica della rappresentazione dell'informazione (Ciotti 2018). La promozione della libera circolazione di idee, metodi e pratiche che eccedono la tradizionale chiusura degli ambiti disciplinari, rimarcano la capacità delle Digital Humanities di favorire forme di accesso più ampie e democratiche alla conoscenza in una dimensione transdisciplinare che non può essere pensata al di fuori del paradigma digitale. Si è discusso a lungo se la rivoluzione digitale sia stata capace di sovvertire l'ontologia dei linguaggi mediali o piuttosto se ne abbia ridefinito le potenzialità; oggi, tali dibattiti sembrano aver perso parte della loro centralità in un mondo in cui si è affermata la categoria "postmediale". I gesti quotidiani, liquidati come semplici automatismi, quali mettere un *like* su un social network o effettuare un acquisto online, si rivelano in realtà atti complessi all'interno di una contemporaneità digitale dove abitudini e comportamenti vengono continuamente processati in un sistema ricorsivo (Previtali 2022, p. 34).
Le trasformazioni innescate dal processo di digitalizzazione hanno investito anche gli strumenti di comunicazione. Basti pensare alla rapidità con cui vengono immessi sul mercato nuovi *devices* che integrano un numero crescente di funzionalità e che sono entrati stabilmente nella quotidianità di una larga parte della popolazione. La diffusione capillare delle tecnologie digitali ha reso possibile una connessione costante e una comunicazione istantanea su scala globale, incidendo sulla vita quotidiana di miliardi di persone e modificando in modo sostanziale il modo in cui gli individui si informano, producono contenuti e costruiscono relazioni. Le innovazioni proposte hanno inciso in modo profondo sui processi comunicativi, ridefinendo l'intero sistema dei media, le pratiche di diffusione e fruizione dei contenuti e, forse soprattutto, l'*agency* degli utenti. Lo spazio, inteso come distanza da percorrere in un certo tempo, viene annullato a favore di immediatezza e simultaneità, generando una fruizione dei contenuti su base emozionale e non sulla riflessività critica, cui si aggiunge discontinuità e incoerenza nel flusso di contenuti di cui viene costantemente messa in discussione la gerarchia (Biscaldi, Matera 2019, p. 18).
La collaborazione tra soggetti diversi, in linea con l'idea di intelligenza collettiva, che emerge dalla convergenza mediale, attiva forme di costruzione della conoscenza favorendo nuovi legami culturali, proponendo una concezione più reticolare e condivisa del sapere in cui ogni individuo può offrire il proprio contributo in un processo dinamico, partecipativo e aperto. La sfida che deve essere raccolta oggi è quella di interrogarsi criticamente sulle modalità attraverso cui le tecnologie digitali vengono integrate nei processi culturali e scientifici, evitando letture meramente celebrative o deterministiche e riconoscendo, invece, la complessità dei rapporti tra strumenti, pratiche e soggetti che li utilizzano.

2. MEDIA DIGITALI E ARCHEOLOGIA

Se il digitale ridefinisce i processi di produzione, circolazione e condivisione della conoscenza, tali trasformazioni chiamano direttamente in causa l'archeologia pubblica e il rapporto con i pubblici nei processi di indagine delle modalità di relazione tra ricerca e società. Esiste un legame profondo tra la comunicazione a distanza e la materialità intrinseca dell'archeologia, come dimostra la lunga tradizione di prodotti audiovisivi che, sin dagli albori della televisione, hanno portato i temi archeologici prima sul piccolo schermo, poi al cinema e, oggi, all'interno dei social network. L'archeologia pubblica non è

comunicazione, ma sicuramente la comunicazione costituisce un elemento fondamentale del rapporto tra archeologia e società, tanto che si ritiene irrinunciabile il coinvolgimento del pubblico attraverso nuovi paradigmi comunicativi (Volpe 2020, p. 47).
L'universo dei social network è popolato da un'ampia varietà di piattaforme, ognuna con vocazioni specifiche, linguaggi propri e differenti modalità di interazione tra utenti che incidono in modo significativo sulle forme di produzione e fruizione dei contenuti. Questa eterogeneità rende i social media ambienti complessi, all'interno dei quali la comunicazione della ricerca non può essere pensata in modo uniforme, ma richiede scelte consapevoli, calibrate su obiettivi specifici, su precisi pubblici di riferimento e sulle caratteristiche dei singoli contesti digitali.
Sin dalle prime fasi di diffusione dei social network, l'archeologia ha riconosciuto queste piattaforme come uno spazio privilegiato dove sperimentare nuove forme di comunicazione, coinvolgimento e dialogo con le persone. Alle prime esperienze pionieristiche (Uomini e cose a Vignale, Miranduolo in Alta Val di Merse – Il progetto) sono seguite numerose sperimentazioni che hanno mostrato come le potenzialità offerte dai social possano essere interpretate in modi differenti, in relazione agli obiettivi, ai contesti e ai pubblici di riferimento. In tutte le sue declinazioni, l'attivazione di profili social ha contribuito ad avvicinare l'archeologia alle comunità, favorendo il superamento di un'immagine distante, burocratica e prescrittiva che a lungo ha caratterizzato il rapporto tra la disciplina e i cittadini. In questo senso, la comunicazione social può agire come un catalizzatore di partecipazione culturale, rendendo le persone parte attiva dei processi di conoscenza e contribuendo a una gestione del patrimonio più attenta ai bisogni delle diverse comunità, pratica che funziona se inserita all'interno di modelli progettuali flessibili e capaci di adattarsi ai diversi contesti territoriali e sociali (Pinna, Sanna Montanelli 2024, p. 239).
Instagram è una piattaforma molto interessante nella quale osservare le modalità attraverso cui l'archeologia trova spazio e visibilità all'interno dell'ecosistema digitale. L'esplorazione di profili e *hashtag* tematici evidenzia la presenza di una *community* ampia, attiva e partecipe, interessata ai temi archeologici e alla loro divulgazione. Un'indagine esplorativa condotta nel 2020, finalizzata a rilevare la diffusione e l'utilizzo degli hashtag legati all'archeologia su Instagram, ha fornito elementi utili a comprendere le dinamiche di visibilità, aggregazione dei contenuti e partecipazione sulla piattaforma (Pinna, Pruneddu 2020, p. 445). Il confronto con i dati aggiornati all'inizio del 2025 mostra una crescita significativa del fenomeno, testimoniata dall'aumento degli account che includono il termine 'archeologia' nel nome utente e dalla presenza di numerosi profili che fanno riferimento a scavi o siti archeologici attraverso le informazioni in *bio*, confermando il progressivo consolidamento della disciplina su Instagram (Demuru, Pinna 2025, p. 245).

3. IL PROFILO INSTAGRAM COME STRUMENTO OPERATIVO DELLA SPERIMENTAZIONE

L'osservazione delle modalità con cui tali trasformazioni si riflettono in pratiche operative concrete consente di cogliere come i principi della comunicazione partecipativa e della costruzione condivisa del sapere possano essere messi alla prova in contesti reali di ricerca, caratterizzati da condizioni territoriali, sociali e culturali specifiche.
Il sito archeologico di Capo Sant'Elia è situato in un'area periferica della città Cagliari, sulla cima di un promontorio alle cui pendici si dipana un borgo, che porta il suo nome, le cui vicende storiche e sociali hanno risentito in modo significativo della collocazione marginale rispetto al centro urbano. L'isolamento si è ulteriormente accentuato in seguito alle scelte di sviluppo urbanistico che, a partire dagli anni Cinquanta, hanno portato Sant'Elia a essere definito un vero e proprio ghetto sociale (Santucciu 2020, p. 65). Ciononostante, per quanto permanga una certa marginalità urbanistica e sopravviva ancora una rappresentazione stigmatizzante del borgo, ancora associato, nel discorso pubblico, a una raffigurazione problematica e poco attrattiva, l'area si configura come un contesto particolarmente fertile per indagini orientate a una lettura più ampia e complessa del sito archeologico, capace di integrare la dimensione materiale dello scavo con quella sociale e relazionale rappresentata dal borgo e dalla comunità che lo abita e perfettamente inserito nel contesto cittadino[1].

[1] Le attività di Vestigia UniCa – Laboratorio di didattica e comunicazione dei Beni Culturali, condotte in occasione della manifestazione locale Monumenti Aperti 2024, hanno rappresentato un'occasione di sperimentazione di forme di apertura verso il territorio, attraverso il coinvolgimento diretto di stakeholder, cittadini e soggetti legati al contesto locale. Le indagini sulla percezione del patrimonio svolte tra i residenti del quartiere hanno restituito una sostanziale distanza tra la comunità e il sito archeologico. Questo scollamento ha portato a riflettere su modalità ulteriori di

La Sella del Diavolo, la propaggine che ospita il sito archeologico, che da un lato offre un affaccio sul litorale e dall'altro regala uno sguardo sulla città, costituisce un punto di riferimento cittadino e connota lo skyline paesaggistico di Cagliari.
Il profilo Instagram @archeosantelia[2] rappresenta la componente operativa della sperimentazione in corso, uno spazio di raccolta e restituzione di temi, istanze e punti di vista espressi dagli stakeholder che, a diverso titolo, si interfacciano con il sito archeologico, contribuendo alla costruzione di una narrazione plurale e condivisa. Tale approccio mira a superare i limiti della narrazione lineare del diario di scavo, spesso concentrata sull'avanzamento delle indagini e sulla presentazione puntuale dei dati, ma meno efficace nella restituzione di una contestualizzazione più ampia del sito e del suo rapporto con il territorio e le comunità di riferimento. Il sito archeologico viene così calato in una dimensione che eccede i confini fisici dell'area di scavo, assumendo una valenza culturale e relazionale più ampia, costruita attraverso il contributo di soggetti diversi, portatori di prospettive eterogenee.
All'interno di questo quadro, il feed di Instagram diventa lo spazio in cui tali contributi si stratificano e si mettono in relazione, un ambiente capace di connettere gli stakeholder tra loro e di offrire una visione complessiva del sito, non riducibile a una sequenza di aggiornamenti, ma restituita come sistema complesso di significati, pratiche e relazioni.
L'intervista è il mezzo individuato come strumento privilegiato di raccolta dati, in grado di rispondere alle esigenze operative del profilo. La costruzione dei contenuti pubblicati non si fonda esclusivamente sui risultati delle indagini archeologiche, ma prende forma soprattutto a partire dalle istanze emerse dal coinvolgimento degli stakeholder raccolte attraverso le interviste. Questa tecnica, basata su un rapporto bilaterale tra intervistatore e intervistato, consente la produzione di dati discorsivi utili a far emergere prospettive, temi e interpretazioni differenti (Corrao 2015). I soggetti coinvolti comprendono una pluralità di professionisti, istituzioni e associazioni, selezionati in funzione del loro rapporto con il sito e del ruolo che rivestono nei processi di conoscenza e valorizzazione, con l'obiettivo di restituire narrazioni plurali e di rendere visibili relazioni e significati spesso marginali o impliciti.
La raccolta delle informazioni si struttura attorno a un form predisposto per orientare l'intervista verso una coerenza tematica che garantisca, al contempo, l'omogeneità dei dati e la comparabilità dei contenuti emersi, senza tuttavia limitare la libertà espressiva degli interlocutori. Il modello, destinato a enti, associazioni, cittadini e professionisti coinvolti, a vario titolo, nella vita del sito, esplicita sin da subito l'idea di intendere lo scavo archeologico come area più ampia di valore storico, paesaggistico e comunitario, con l'obiettivo di definire temi, argomenti e contenuti che troveranno successivamente spazio nel profilo Instagram dedicato. Inizialmente pensato per essere compilato autonomamente, il form, costituisce una traccia per un'intervista che si svolge in presenza (vd. Fig.1). Questa modalità consente di un'interazione diretta tra intervistatore e intervistato in cui è possibile utilizzare formule e strategie atte a orientare il dialogo in funzione dei temi emersi durante la conversazione, adattando le domande all'andamento dell'intervista per cogliere ulteriori elementi che difficilmente sarebbero potuti emergere durante la compilazione autonoma del form (Nistor 2024, p. 38).



Figura 1. Un momento dell'intervista all'Assessora Giulia Andreozzi (foto S. Pambira)

L'approccio dialogico, inoltre, porta con sé un carattere generativo, attivando un meccanismo di rimando grazie al quale vengono segnalati nuovi interlocutori, ampliando così la rete di stakeholder e di confronti ulteriori per una produzione continua di contenuti.

comunicazione della ricerca archeologica, orientata alla definizione di pratiche capaci di intercettare le istanze della comunità e di integrarle all'interno di processi di costruzione condivisa della conoscenza.
[2] La pagina è raggiungibile a questo link: https://www.instagram.com/archeosantelia/ (cons. 05/02/2026)

La suddivisione del form in sezioni consente, in primo luogo, di delineare il profilo dello stakeholder, raccogliendo informazioni relative alla natura del legame con il sito, alle modalità e alla durata di tale relazione, nonché alle attività svolte in rapporto ad esso. Questo primo blocco è seguito da una sezione dedicata alla raccolta di informazioni di carattere interpretativo sul valore che lo stakeholder attribuisce al sito, con un focus sugli aspetti dell'area più importanti da preservare e valorizzare e quelli più vulnerabili o a rischio, sia dal punto di vista materiale che culturale, nonché quelli che hanno un potenziale ancora inespresso e potrebbero essere sviluppati o valorizzati. Le domande, di natura prevalentemente aperta, sono pensate per stimolare una riflessione articolata, favorendo l'emersione di narrazioni personali e di punti di vista differenziati, che contribuiscono ad ampliare il quadro conoscitivo oltre i dati strettamente archeologici. Si apre, quindi, una terza sezione dedicata all'individuazione di temi e argomenti ritenuti rilevanti per la narrazione del sito. Questa parte del form è finalizzata a raccogliere indicazioni sugli ambiti che, secondo gli stakeholder, meritano di essere approfonditi e condivisi, nonché sulle possibili fonti o sui materiali utili a documentarli. Gli intervistati sono invitati a segnalare questioni di interesse legate al sito, che possono riguardare aspetti storici, archeologici, ambientali e sociali, così come il rapporto con le comunità locali o la presenza di progetti, iniziative e attività, in corso o recentemente concluse. L'obiettivo è far emergere tematiche e prospettive differenti, utili a restituire una visione articolata e contestualizzata del sito.
Il quarto blocco di domande verte sulle prospettive future del sito archeologico, mettendo l'intervistato nelle condizioni di immaginare il futuro così da poter costruire una visione condivisa e orientare azioni concrete, sostenibili e partecipate. Se da un lato gli si chiede di riflettere su un quadro ipotetico nel medio periodo senza alcun cambiamento sostanziale rispetto alle dinamiche attuali, dall'altro si chiede di offrire un'introspezione sul suo possibile scenario ideale e su cosa vorrebbe vedere realizzato davvero. Il confronto tra una proiezione di continuità in linea con i processi attualmente in atto e una visione ideale libera da ogni vincolo permette di individuare criticità, potenzialità e priorità di intervento, offrendo elementi utili a orientare ulteriori strategie di ricerca, comunicazione e valorizzazione, ma fondate su una visione condivisa, consapevole e partecipata.
Le informazioni raccolte durante le interviste vengono successivamente rielaborate e tradotte nei linguaggi propri di Instagram, tenendo conto dei diversi formati messi a disposizione dalla piattaforma. Il processo di trasformazione dei dati discorsivi conduce alla realizzazione di contenuti strutturati, caroselli che combinano immagini e testi, in cui le frasi chiave emerse dalle interviste vengono evidenziate nelle didascalie e associate a fotografie capaci di restituire visivamente i temi e le istanze emerse (vd. Fig. 2). Di fatto, il processo di rielaborazione delle informazioni dà luogo a una co-creazione dei contenuti secondo un metodo collaborativo che prevede il trasferimento dei dati discorsi dall'intervista al post Instagram secondo una riformulazione e un adattamento ai linguaggi della piattaforma. La dimensione collaborativa non è introdotta a posteriori, ma risulta intrinseca nell'intervista stessa, poiché costituisce lo spazio primario di produzione e condivisione dei contenuti.



Figura 2. *Overlay* testuale delle ultime due immagini del post pubblicato il 22/12/2025 (foto ed elaborazione M. Demuru)

4. DALLA SPERIMENTAZIONE ALLA DEFINIZIONE DI UN MODELLO

Il profilo Instagram viene qui assunto non come semplice canale di diffusione, ma come strumento partecipativo attraverso cui sperimentare, in un contesto reale di ricerca archeologica, la possibilità di costruire una comunicazione fondata sul coinvolgimento diretto degli stakeholder, capace di mettere in connessione i diversi contenuti emersi nel corso delle interviste, che vengono, quindi, riorganizzati e trasformati da informazioni accessorie a elemento strutturale del processo di condivisione. L'impostazione adottata si propone di superare il modello comunicativo di tipo top-down, unidirezionale, riconducibile al

cosiddetto "modello del deficit", nel quale la conoscenza viene trasmessa da un'autorità scientifica a un pubblico passivo, e che è stato progressivamente messo in discussione in favore di approcci partecipativi che valorizzano il coinvolgimento attivo dei soggetti nei processi di costruzione e condivisione del sapere (Bucchi 2008, Volpe 2020).
L'osservazione delle modalità con cui i contenuti attivano relazioni tra i soggetti coinvolti costituisce il punto di partenza per la valutazione del grado di raggiungimento degli obiettivi prefissati, letti in relazione al contesto specifico del progetto. L'analisi si fonda sull'osservazione sistematica dei contenuti pubblicati, articolata attraverso il tracciamento delle ricorrenze tematiche e una mappatura delle connessioni tra gli stakeholder, su un corpus costituito da 15 interviste e da contenuti pubblicati con una frequenza media di circa un post settimanale, spesso accompagnato dal relativo reel. L'esame trasversale dei post ha permesso di individuare alcune ricorrenze tematiche significative. In particolare, emerge con continuità il tema della fragilità del sito, legata sia a fattori ambientali sia all'impatto delle attività umane, cui si affianca la necessità di una tutela condivisa e continuativa. Parallelamente, ricorre in modo sistematico il tema del rapporto con la comunità, intesa non come semplice destinataria della comunicazione, ma come parte attiva nei processi di valorizzazione e conservazione del patrimonio. A questi aspetti si affianca una riflessione diffusa sulle potenzialità inespresse del sito e sulle prospettive future, spesso articolate in termini di apertura, accessibilità e integrazione con il contesto urbano e culturale. Infine, si evidenzia una costante attenzione al rapporto tra ricerca, didattica e territorio, che configura il sito come spazio di formazione, sperimentazione e costruzione condivisa della conoscenza.
La mappatura delle connessioni tra gli stakeholder ha evidenziato come i contributi non si configurino come unità isolate, ma come elementi interconnessi tra loro. In alcuni casi, tali connessioni si manifestano in forma esplicita, attraverso richiami diretti tra studiosi[3], mentre più frequentemente emergono sotto forma di convergenze tematiche, in cui soggetti diversi - ricercatori, studenti e rappresentanti istituzionali – affrontano questioni analoghe da prospettive differenti. Questo processo evidenzia la capacità del dispositivo di costruire un sistema relazionale in cui le istanze individuali confluiscono in un quadro interpretativo condiviso. La ricorrenza di tali nuclei tematici, insieme alla presenza di richiami incrociati tra gli stakeholder e all'apertura di nuovi filoni narrativi, consente di valutare in che misura il feed funzioni come spazio di connessione e non soltanto di esposizione. La lettura dei dati quantitativi offerti dalla piattaforma, invece, quali interazioni, livelli di partecipazione e continuità dell'engagement, viene integrata all'interno di questo quadro e interpretata non come indicatore assoluto, ma in relazione alla dimensione locale del progetto e al grado di coinvolgimento effettivo dei soggetti che gravitano attorno al sito[4]. L'analisi congiunta di questi aspetti permette di cogliere non solo la visibilità dei contenuti, ma soprattutto la loro incidenza nell'attivazione di dinamiche partecipative radicate nel territorio, restituendo un sistema di osservazione continuo che accompagna l'evoluzione della sperimentazione.

5. PRIME RIFLESSIONI

L'obiettivo, prefissato in fase di progettazione, di restituire il sito archeologico di Capo Sant'Elia come un luogo inserito in uno spazio che supera i confini dell'area di scavo, sta trovando una prima concreta attuazione attraverso il progressivo coinvolgimento degli stakeholder e la rielaborazione dei contributi raccolti. Le risposte emerse dalle interviste iniziano, infatti, a delineare connessioni tra i diversi soggetti coinvolti, mettendo in relazione punti di vista, competenze e istanze eterogenee, facendo del feed del profilo uno spazio di sintesi, in cui la somma dei contributi restituisce un quadro articolato del sito, costruito a partire dalla pluralità delle voci e dalla molteplicità delle relazioni che lo attraversano.
La comunicazione, impostata consapevolmente in questa direzione, non si limita a descrivere l'area archeologica o a documentare l'avanzamento delle attività di scavo, ma attribuisce al profilo Instagram il ruolo di strumento operativo orientato alla sperimentazione di un modello innovativo di comunicazione applicato a un contesto reale di ricerca archeologica, consentendo di proseguire la comunicazione della ricerca anche al di là delle singole campagne di scavo e dei momenti di attività sul campo. All'interno del

[3] È il caso del post generato dall'intervista a Nicoletta Usai, Prof.ssa di Storia dell'Arte Medievale all'Università di Cagliari, che chiama in causa direttamente il collega Andrea Pala, Prof. di Storia dell'Arte Medievale anch'egli. Il post è consultabile a questo link: https://www.instagram.com/p/DUA3cDiiItA/?utm_source=ig_web_copy_link&igsh=MzRlODBiNWFlZA==
[4] Una prima analisi dei dati di insight del profilo evidenzia come le interazioni siano prevalentemente generate dai follower rispetto ai non follower e come questi siano geolocalizzati principalmente a Cagliari.

form, due domande svolgono una funzione chiave: l'individuazione degli argomenti ritenuti significativi da raccontare attraverso il profilo e le modalità di reperimento delle informazioni utili alla loro costruzione. Tali quesiti non si limitano a raccogliere suggerimenti tematici, ma attivano un processo dinamico e generativo, in cui le risposte diventano il punto di partenza per l'apertura di nuovi filoni narrativi e per l'ampliamento progressivo del quadro interpretativo del sito. Questo meccanismo consente di costruire una narrazione incrementale e aperta, capace di riorganizzare e ampliare nel tempo i contenuti del profilo attraverso rubriche dedicate, nuovi format e differenti linee narrative, sfruttando in modo sistematico le funzionalità offerte dalla piattaforma. La strutturazione del processo comunicativo secondo questa logica potrebbe rendere possibile la definizione di un modello di comunicazione replicabile in contesti analoghi, adattabile a siti archeologici inseriti in ambiti territoriali complessi e caratterizzati da una pluralità di stakeholder. Al tempo stesso, la natura aperta del modello consente una continua implementazione della strategia comunicativa, favorendo l'introduzione di nuovi strumenti, linguaggi e modalità di condivisione, nonché ulteriori sviluppi attraverso l'integrazione con altri social network, in una prospettiva di ecosistema comunicativo più ampio e interconnesso.

BIBLIOGRAFIA

Balbi, G. (2021). L'ultima ideologia. Breve storia della rivoluzione digitale. Roma: Edizioni Laterza.

Biscaldi A., Matera V. (2019). Antropologia dei social media. Comunicare nel mondo globale Roma: Carocci editore.

Bucchi, M (2008). Dal deficit al dialogo, dal dialogo alla partecipazione – e poi? Modelli di interazione tra scienza e pubblico. Rassegna italiana di sociologia. Vol. 3. Bologna: Il Mulino, 377-402.

Ciotti F. (2018). Le Digital humanities in Italia. La tradizione del nuovo. Aib studi, 58, 2, 177-179.

Corrao S. (2015). L'intervista nella ricerca sociale. Quaderni di Sociologia, 38. http://journals.openedition.org/qds/1058.

Demuru M., Pinna F. (2025). L'archeologia medievale italiana su Instagram. Situazione, difficoltà e prospettive nella comunicazione della ricerca archeologica attraverso i social network. X Congresso nazionale di archeologia medievale, Udine – Cividale del Friuli 9-13 settembre 2025. Minguzzi, S. (Ed.). 3, 242-247.

Eugeni R. (2022). La condizione postmediale. Brescia: Scholé-Editrice Morcelliana.

Nistor P. P. (2024). The Interview as a Research Technique in Social Work –Methodological and Practical Challenges. Logos Universality Mentality Education Novelty: SocialSciences, 13, 2, 29-42.

Pinna F., Sanna Montanelli M. (2024). Citizen Archaeology. Un contributo alla costruzione di una archeologia pubblica come scienza normale. Post Classical Archaeology, 14, 237-256.

Pinna F., Pruneddu F. (2020). L'archeologia medievale italiana su Instagram: osservazioni sulla situazione presente e prospettive. VI Ciclo di Studi Medievali, Firenze 8-9 giugno 2020, 442-448.

Previtali G. (2022). Che cosa sono le digital humanities. Roma: Carocci Editore.

Volpe G. (2020). Archeologia Pubblica. Metodi, tecniche, esperienze. Carocci Editore: Roma.

Public engagement nell'era dei synthetic media: per una co-interpretazione critica tra Digital Humanities e Media Education

GABRIELE PROSPERI[1], MARIO A. BOCHICCHIO[2]
[1]UNIVERSITÀ DEL SALENTO, [2]UNIVERSITÀ DEGLI STUDI DI BARI

ABSTRACT (ITALIANO)

L'articolo propone un contributo metodologico relativo al progetto *LUM.IA – Learning and Understanding Media through Artificial Intelligence*, finanziato dal Fondo Italiano per la Scienza (FIS3) in corso di avviamento, che mira ad analizzare processi informali di consumo audiovisivo su piattaforme come TikTok, Instagram e YouTube, con attenzione alle pratiche di rielaborazione e circolazione dei contenuti, a sviluppare metodologie interdisciplinari per la *media education* e a integrare strumenti IA, in particolare sistemi per la *Retrieval-Augmented Generation* (RAG) e chatbot educativi, in percorsi trasparenti, inclusivi e controllabili. L'obiettivo è delineare una cornice operativa di co-interpretazione assistita che connetta *public engagement*, curatela delle evidenze e interazione conversazionale, e descrivere un protocollo adattabile di laboratori con scuole e comunità. Il contributo intende fornire alla comunità DH un modello trasferibile per progettare approcci di *public engagement* capaci di affrontare i *synthetic media* senza puntare all'infallibilità degli strumenti ma rafforzando pratiche argomentative e trasparenza epistemica.
Parole chiave: Media education; Synthetic media; Co-interpretazione assistita; Retrieval-Augmented Generation.

ABSTRACT (ENGLISH)

Public Engagement in the Age of Synthetic Media: For a Critical Co-Interpretation Between Digital Humanities and Media Education.
The paper proposes a methodological contribution related to the LUM.IA – Learning and Understanding Media through Artificial Intelligence project, funded by the Italian Science Fund (FIS3), currently in its start-up phase, which aims to analyze informal processes of audiovisual consumption and production on platforms such as TikTok, Instagram, and YouTube, to develop interdisciplinary methodologies for media education, and to integrate AI tools, in particular Retrieval-Augmented Generation (RAG) systems and educational chatbots, into transparent, inclusive, and controllable pathways. The goal is to outline an operational framework of assisted co-interpretation that connects public engagement, evidence curation, and conversational interaction, and to describe an adaptable protocol for workshops with schools and communities. The contribution aims to provide the DH community with a transferable model for designing public engagement approaches capable of addressing synthetic media, without aiming for the infallibility of tools, but rather by strengthening argumentative practices and epistemic transparency.
Keywords: Media education; Synthetic media; Assisted co-interpretation; Retrieval-Augmented Generation

1. INTRODUZIONE

La rapida diffusione di contenuti audiovisivi sintetici o manipolati sta cambiando il modo in cui scuole e comunità attribuiscono le fonti e praticano la verifica. Il problema non è solo la distinzione tra vero e falso tipica della post-verità (Keyes, 2004; d'Ancona, 2017), ma un "disordine informativo" che colpisce particolarmente pubblici vulnerabili come quello degli adolescenti. Al contempo, la dipendenza da tecnologie e media digitali è in crescita (Mauceri & Di Censi, 2020), ridefinendo coordinate spazio-temporali e forme di esperienza. Le pratiche adolescenziali sui social si sviluppano, infatti, in continuità tra online e offline (Tirocchi & Serpieri, 2020): una condizione *onlife* (Floridi, 2015) in cui vita digitale e vita "reale" risultano intrecciate in un medesimo quadro di significato (Boccia Artieri et al., 2017) e nella quale i contenuti audiovisivi funzionano come ambienti di relazione, identità e apprendimento (Boyd, 2007).
Ciò si inserisce all'interno di un preesistente contesto di produzione informale – *short video*, *meme-culture*, *remix* (Lessig, 2008) – determinato da una cultura partecipativa (Jenkins, 2006) in cui gli utenti possono archiviare, annotare, appropriarsi e far circolare contenuti. È qui che la mediazione algoritmica

rende le pratiche interpretative inseparabili dalle infrastrutture di visibilità, che determinano cosa viene reso rilevante nello spazio pubblico (Gillespie, 2014). L'emergere dei *synthetic media* – contenuti generati o significativamente modificati da processi automatizzati, spesso basati su IA – democratizza la manipolazione e incide sulla reputazione e sulla fiducia (Chesney & Citron, 2019, 1785): oltre al danno puntuale, cresce uno scetticismo generalizzato che porta a dubitare anche di prove audiovisive autentiche.

2. OBIETTIVI E POSIZIONAMENTO

Il progetto LUM.IA, in corso di avviamento, nasce dall'osservazione che la centralità dell'audiovisivo nelle generazioni più giovani, unita alla pervasività delle piattaforme social e alla disponibilità di strumenti generativi, produce vulnerabilità specifiche (esposizione a manipolazione, disinformazione, dipendenze), ma anche opportunità di partecipazione e creatività che non possono essere gestite con logiche solo prescrittive. Da qui tre obiettivi: (i) l'analisi dei processi informali di consumo e produzione audiovisiva, con particolare focus sulle abitudini degli adolescenti (10-17 anni) e sulle interazioni con i media digitali; (ii) lo sviluppo di metodologie interdisciplinari per migliorare i percorsi di *media education* (Buckingham, 2019) includendo un modulo di alfabetizzazione IA di base per i docenti; (iii) l'integrazione dell'IA nei percorsi educativi garantendo trasparenza e inclusività.
Il progetto studia come gli adolescenti si muovono tra creazione e fruizione, come *trend* e algoritmi influenzano fiducia e comportamenti mediali, e come contenuti sintetici (*deepfake*, *script* automatizzati, *video AI-generated*) incidano su identità e percezione. L'analisi include anche le pratiche produttive: la creazione di contenuti rielaborati o sintetici da parte degli stessi adolescenti è parte del campo osservato, poiché consente di comprendere le competenze già presenti, le soglie di riconoscimento della manipolazione e le dinamiche di circolazione dei contenuti nelle reti di pari. Sul piano applicativo il progetto prevede sistemi di *Retrieval-Augmented Generation* (RAG) e *chatbot* educativi per *feedback* personalizzati per età e competenze, supporto alla distinzione tra materiali autentici e manipolati, e funzioni di accessibilità (*text-to-speech*, supporti grafici) per studenti con *Specific Learning Disabilities* (SLD). Il *public engagement* è integrato nella sua traiettoria d'impatto: la sostenibilità e la diffusione saranno testate anche tramite iniziative nel quadro del Piano Nazionale Cinema e Immagini per la Scuola (CIpS) – promosso congiuntamente dal MiC e dal MIM, finalizzato a introdurre in modo strutturato il linguaggio audiovisivo nelle scuole come strumento educativo trasversale – e con partner territoriali, così da rendere riusabili laboratori e materiali (*knowledge base*, protocolli, linee guida) in contesti diversi.
Per rendere operativo il legame tra consumo, produzione audiovisiva e dispositivo educativo, l'analisi delle pratiche di piattaforma produce un repertorio di casi (autentici, borderline, manipolati o sintetici) tradotti in schede strutturate costituite da metadati, descrizione, *claim*, contesto di circolazione, piste di verifica. Queste schede costituiscono l'*input* dei laboratori: vengono discusse in gruppo per ricostruire il *claim*, il contesto e possibili verifiche, e per produrre *output* restituiti (*checklist* e schede argomentate). L'interazione con il sistema conversazionale le supporta recuperando dalla *knowledge base* i criteri e le fonti pertinenti, esplicitandone la provenienza e suggerendo piste di verifica, senza sostituire la discussione collettiva. I casi sono inoltre referenziati tramite collegamenti e descrizioni contestuali, così da garantire la tracciabilità.

3. IL RUOLO DELL'INTELLIGENZA ARTIFICIALE

Nel contesto della crescente attenzione della comunità scientifica per l'impiego dell'intelligenza artificiale generativa nei contesti educativi (Zawacki-Richter et al., 2019; Kasneci et al., 2023) i *Large Language Models* (LLM) si caratterizzano per le loro potenzialità di personalizzazione dell'apprendimento, generazione di contenuti e interazione conversazionale. Tuttavia, la letteratura ne evidenzia limiti strutturali che risultano particolarmente critici quando l'obiettivo didattico non è la semplice fruizione di informazioni, ma lo sviluppo di competenze critiche e argomentative: la tendenza a produrre allucinazioni, ossia contenuti plausibili ma non fondati su evidenze verificabili (Groza & Marginean, 2023); la scarsa copertura di domini specialistici e di conoscenze locali non presenti nei dati di addestramento (Theophilou et al., 2023); l'opacità delle fonti sottostanti alla generazione, che impedisce all'utente di risalire alle evidenze e ne favorisce una delega epistemica acritica. Tali limiti risultano particolarmente problematici nel contesto del progetto qui delineato, che si propone di accompagnare scuole e comunità nell'analisi dei *synthetic media* e nella costruzione di pratiche di verifica informata. Un sistema che generasse risposte non ancorate a fonti controllate e curate riprodurrebbe, paradossalmente, le stesse

dinamiche di opacità informativa che il progetto intende contrastare. Per queste ragioni, la scelta architetturale fondamentale è quella di adottare un sistema di *Retrieval-Augmented Generation* (RAG), affiancato da *chatbot* educativi progettati per operare come mediatori conversazionali all'interno di percorsi strutturati.

3.1 PRINCIPI DI FUNZIONAMENTO DELLA RETRIEVAL-AUGMENTED GENERATION

La RAG, introdotta da Lewis et al. (2020), combina la generazione linguistica dei modelli di grandi dimensioni con un meccanismo di recupero di informazioni da basi di conoscenza esterne. A differenza di un LLM usato in modalità puramente generativa, un sistema RAG integra nella risposta frammenti documentali recuperati da un *repository* controllato e indicizzato, così da vincolare la produzione testuale a evidenze selezionate (Lewis et al., 2020). In termini architetturali, un sistema RAG comprende tipicamente tre moduli (Gao et al., 2024): (i) retrieval, che individua i frammenti più pertinenti rispetto alla domanda; (ii) *augmentation*, che inserisce tali frammenti nel contesto della richiesta; (iii) *generation*, in cui il LLM genera l'*output* tenendo conto sia delle proprie capacità linguistiche sia delle evidenze recuperate.
Operativamente, la collezione documentale viene segmentata in unità testuali di dimensione controllata (*chunking*) e trasformata in rappresentazioni vettoriali (*embedding*), memorizzate in un database vettoriale (*vector store*). La domanda dell'utente viene a sua volta convertita in *embedding* e confrontata, tramite ricerca per prossimità semantica, con quelli del *repository*; i frammenti più rilevanti vengono selezionati e passati al modello come contesto per la generazione della risposta (Lewis et al., 2020; Gao et al., 2024). Rispetto all'uso diretto di un LLM, questo impianto consente di: (a) limitare l'*output* a un *corpus* definito e verificato, riducendo incoerenze e informazioni spurie; (b) rendere le fonti richiamate rintracciabili, facilitando controllo e confronto; (c) aggiornare la base di conoscenza senza riaddestrare il modello, mantenendo flessibilità e adattabilità dei contenuti (Swacha & Gracel, 2025).

3.2 IL CHATBOT COME MEDIATORE CONVERSAZIONALE

La scelta RAG è coerente con gli obiettivi del progetto perché permette un'interazione in cui l'IA non sostituisce il giudizio critico, ma lo supporta rendendo più trasparenti le condizioni di produzione della risposta. La letteratura segnala una crescente adozione dei *chatbot* RAG in ambito educativo per ridurre l'impatto delle allucinazioni e migliorare accuratezza e aggiornabilità della conoscenza (Swacha & Gracel, 2025; Li et al., 2025), in linea con la necessità di un'integrazione critica dei LLM nei contesti didattici (Kasneci et al., 2023). In questa direzione, la sperimentazione mostra come un RAG basato su fonti curate possa rispondere alle esigenze didattiche (richiamate anche in §4) e come la qualità dell'*output* dipenda soprattutto da due leve: la curatela del *repository* e la progettazione dei *prompt* per tipologie di utente e fasce d'età, così da calibrare registro e profondità dell'interazione. All'interno del progetto, il *chatbot* non è un erogatore di risposte definitive ma un mediatore conversazionale che supporta fasi specifiche del laboratorio, definito tanto da ciò che fa quanto da ciò che evita di fare. Nella preparazione e nell'orientamento recupera dalla *knowledge base* curata i frammenti pertinenti, li riformula in un registro adeguato al profilo d'uso ed esplicita la provenienza. Come supporto argomentativo, propone domande e piste di verifica, stimolando la continuità dell'interazione e valorizzando errore e incertezza come occasioni di apprendimento critico. Quando la domanda eccede il perimetro della *knowledge base* o le evidenze sono insufficienti, segnala esplicitamente il limite, evitando *output* non verificabili e rendendo visibile l'incertezza epistemica. Al contempo, il *chatbot* non certifica l'autenticità dei contenuti, non emette giudizi definitivi e non sostituisce il confronto in presenza: in una cornice di *co-interpretazione assistita*, esso prepara e rende disponibili evidenze, mentre la costruzione del giudizio resta affidata alla discussione collettiva e alla mediazione pedagogica.

3.3 COMPONENTI DEL SISTEMA E CURATELA DELLA *KNOWLEDGE BASE*

La progettazione del sistema prevede quattro componenti interdipendenti che concorrono alla qualità dell'interazione: la curatela dei contenuti, il *retrieval*, la generazione e la tracciabilità.
La curatela della *knowledge base* rappresenta il fondamento dell'intero impianto. Il *repository* documentale sarà composto da fonti selezionate dal gruppo di ricerca e dai partner sulla base di criteri espliciti: autorevolezza, pertinenza tematica, aggiornamento e diversità delle prospettive. Per il dominio specifico dei *synthetic media*, la base di conoscenza includerà materiali relativi alle tecniche di manipolazione audiovisiva, ai criteri di verifica delle fonti, a casi studio documentati, a linee guida istituzionali e a letteratura scientifica di riferimento. La selezione dei contenuti non è un'operazione

puramente tecnica, ma un atto epistemico che definisce il perimetro entro cui il sistema può operare: ciò che non è presente nella *knowledge base* non potrà essere recuperato, rendendo trasparente il confine tra ciò che il sistema è in grado di supportare e ciò che eccede le sue possibilità.
Il modulo di *retrieval* si occupa della segmentazione, indicizzazione e ricerca semantica dei frammenti. La calibrazione dei parametri di *chunking* (dimensione dei frammenti, sovrapposizione, granularità) e la scelta del modello di *embedding* incidono direttamente sulla qualità delle risposte, poiché determinano quali porzioni di testo vengono effettivamente recuperate e fornite al modello generativo.
Il modulo di generazione opera sulla base dei frammenti recuperati e dei *prompt* progettati per il contesto specifico. La progettazione dei *prompt* (*prompt engineering*) costituisce una leva cruciale per adeguare il registro linguistico, il livello di approfondimento e la struttura delle risposte alle esigenze delle diverse tipologie di utenti: studenti di diverse fasce d'età, docenti, operatori culturali (Prosperi et al., 2025). I prompt saranno inoltre progettati per includere l'indicazione esplicita delle fonti richiamate, la segnalazione dei limiti della risposta e la proposta di ulteriori piste di approfondimento.
La tracciabilità chiude il circuito della trasparenza epistemica; a livello di interfaccia, l'utente potrà visualizzare i frammenti documentali su cui si fonda la risposta, verificandone provenienza e adeguatezza: l'interazione con il *chatbot* passa da consumo passivo a pratica attiva di valutazione critica delle fonti.

3.4 CORPUS DI RIFERIMENTO, CASI AUDIOVISIVI E CONDIZIONI D'USO

La *knowledge base* sarà costruita come un insieme di risorse testuali verificabili e di "casi audiovisivi" descritti, collegati da metadati e criteri di verifica. Da un lato, il corpus includerà fonti autorevoli (letteratura scientifica selezionata, linee guida istituzionali, *checklist* e materiali educativi), che definiscono concetti, tecniche e procedure di analisi. Dall'altro, includerà una collezione di casi provenienti dal contesto piattaforma, selezionati per valore didattico e rappresentatività. La selezione dei casi coprirà tipologie ricorrenti di *synthetic media* (ad esempio *face-swap*, *lip-sync*, *voice cloning*, contenuti generati/ibridi e manipolazioni contestuali), così da rendere esplicito il perimetro del dominio affrontato nei laboratori. Ciascun caso sarà rappresentato nel *repository* tramite schede descrittive e campi testuali – comprensivi di descrizione del contenuto, *claim*, contesto di pubblicazione e ricezione, indicatori discutibili, piste di verifica – con uso delle *caption* già disponibili e, ove utile, con una trascrizione del parlato ottenuta come supporto all'analisi (ad esempio mediante procedure automatizzate e/o revisione), così da consentire al *retrieval* di operare su contenuti testuali tracciabili e confrontabili.
La natura audiovisiva di questi contenuti pone una sfida specifica per il *retrieval*: a differenza di un *corpus* testuale, i casi non sono direttamente indicizzabili come testo continuo. La soluzione adottata – rappresentare ciascun caso tramite schede descrittive testuali, metadati, trascrizioni e *caption* – consente al sistema RAG di operare su contenuti tracciabili e confrontabili senza richiedere l'elaborazione diretta dei file audiovisivi, che restano accessibili tramite link con metadati contestuali. Questo approccio delimita consapevolmente il perimetro del sistema: la *knowledge base* supporta l'analisi argomentativa e la verifica delle fonti, non la rilevazione automatica di artefatti tecnici nei media, coerentemente con l'obiettivo educativo del progetto.
L'impianto è coerente con l'obiettivo di evitare una *detection* automatica dell'autenticità: il focus resta sulla co-interpretazione, sulla verifica informata e sulla costruzione argomentativa, sostenute da evidenze curate e dalla trasparenza delle fonti richiamate. Contestualmente, il *corpus* è progettato tenendo esplicitamente conto di *copyright* e condizioni d'uso dei dati: si privilegiano risorse open e istituzionali e, quando disponibili, dataset con licenze compatibili; per contenuti soggetti a protezione o a *policy* di piattaforma, la *knowledge base* adotta un approccio *link/metadata-first*, evitando la redistribuzione del contenuto integrale.

3.5 METRICHE E VALUTAZIONE DEL CHATBOT

La valutazione del sistema conversazionale si articolerà su due piani distinti ma complementari: la qualità tecnica del sistema e la qualità della mediazione didattica.
Sul piano tecnico, saranno monitorati indicatori relativi al comportamento del sistema RAG: la percentuale di risposte effettivamente ancorate ai frammenti del *repository* (*groundedness*), la rilevanza e la copertura dei frammenti recuperati rispetto alla domanda posta (*retrieval precision* e *recall*), la coerenza e la fedeltà della risposta generata rispetto alle evidenze recuperate (*faithfulness*), e la qualità dell'esplicitazione dell'incertezza, ovvero la capacità del sistema di segnalare quando la domanda eccede

il perimetro della *knowledge base* o quando le evidenze sono insufficienti. La riduzione delle allucinazioni, intesa come generazione di contenuti non fondati sulle fonti del *repository*, sarà un indicatore-chiave. Sul piano della mediazione didattica, la valutazione si concentrerà sull'efficacia del sistema nel facilitare pratiche critiche anziché sostituirle. Verranno considerati: la capacità dell'interazione di stimolare la formulazione autonoma di domande e ipotesi verificabili da parte degli utenti; la riduzione della delega epistemica allo strumento, osservabile attraverso l'analisi delle conversazioni e dei percorsi di verifica intrapresi dopo l'interazione; la coerenza dei rifiuti motivati, ovvero la qualità con cui il sistema declina di rispondere a domande fuori perimetro senza produrre output inattendibili; l'adeguatezza del registro linguistico e degli strumenti compensativi (sintesi vocale, supporti grafici, mappe concettuali) per le diverse tipologie di utenti, inclusi gli studenti con *Specific Learning Disabilities*.
La valutazione sarà condotta attraverso cicli iterativi di test, coinvolgendo docenti e studenti nei contesti laboratoriali previsti dal progetto. Le metriche tecniche saranno integrate con osservazioni qualitative e con i dati raccolti attraverso i protocolli di valutazione multi-asse descritti nella sezione dedicata.
Dal punto di vista implementativo, il sistema prevede una pipeline iterativa: selezione e curatela delle fonti; produzione di schede caso e segmentazione in *chunk*; generazione di *embedding* e indicizzazione in un *vector store*; progettazione di *prompt* differenziati per profilo d'uso; interfaccia conversazionale con tracciabilità delle fonti e gestione esplicita dell'incertezza. L'aggiornamento della *knowledge base* avviene per cicli successivi, alimentato dai laboratori (schede e materiali restituiti), così da migliorare progressivamente la copertura del dominio e la qualità del *retrieval* senza ricorrere al riaddestramento del modello. L'implementazione può basarsi su componenti standard e riusabili (modelli di *embedding* e un LLM orchestrato via *prompt*), privilegiando soluzioni open-source e infrastrutture istituzionali dove possibile. Lo sviluppo è previsto per fasi iterative: prototipo funzionale su corpus limitato, validazione nei laboratori, quindi estensione progressiva del *repository* e affinamento del *retrieval* e dei *prompt* in base ai *feedback*. Questa impostazione consente di rendere l'*effort* scalabile e di aggiornare contenuti e criteri senza dover riaddestrare il modello.

4. DISCUSSIONE: PER UNA CO-INTERPRETAZIONE ASSISTITA

La prima sperimentazione di queste idee è maturata all'interno di un precedente progetto di ricerca applicata dedicato a tecniche e tecnologie per la gestione di contenuti culturali. Il sistema didattico basato su RAG, costruito attorno a un *repository* interrogabile tramite un'interfaccia conversazionale, fu concepito per rispondere a tre esigenze didattiche specifiche: (1) superare i limiti dei LLM in domini specialistici, dove modelli generalisti tendono a produrre risposte plausibili ma non coerenti con fonti circoscritte, lessico disciplinare e obiettivi di apprendimento; (2) ridurre il rischio di "allucinazioni" vincolando la generazione a evidenze recuperate dal *repository* e rendendo possibile (anche a livello di interfaccia) la tracciabilità di ciò che viene richiamato; (3) offrire un registro linguistico e un livello di complessità modulati sul profilo d'uso (docenti e studenti) e sulle fasce d'età, così da trasformare l'*output* in uno *scaffolding* (definizioni, esempi, domande guida, riformulazioni) anziché in una risposta risolutiva uguale per tutti.
Farinacci (2024) sottolinea il potenziale dell'IA nella *media education*, evidenziando la possibilità di offrire apprendimento personalizzato e migliorare l'accessibilità dei contenuti, abilitando analisi più avanzate delle interazioni digitali. In questa stessa direzione, la prima sperimentazione presentava due aspetti che diventano fondativi per l'attuale progetto: da un lato, l'idea che la qualità dell'output dipenda dalla curatela delle fonti e dalla progettazione del processo di recupero (segmentazione/*chunking*, indicizzazione e *retrieval*) più che da una generazione "libera"; dall'altro, che l'integrazione didattica richieda un impianto pedagogico interattivo (nella sperimentazione la chiave di volta fu la *flipped classroom*), in cui l'interazione col bot avviene come preparazione e orientamento (una sorta di pre-lettura guidata, o un chiarimento di concetti tramite la costruzione di domande), mentre il tempo in presenza viene dedicato ad attività di discussione, verifica delle fonti, esercizi di argomentazione e produzione di elaborati, con il docente nel ruolo di facilitatore e garante metodologico. Gli strumenti compensativi (sintesi vocale e visiva, mappe concettuali) diventano, così, dei dispositivi che rendono effettiva la partecipazione e permettono di mantenere l'apprendimento centrato su pratiche critiche e inclusive.
L'attuale progetto assume, quindi, questa esperienza come punto di partenza e ne compie il salto di scala: dal dominio storico-locale e dalla classe sperimentale a un problema pubblico più ampio e contemporaneo – la *media literacy* nell'ecosistema audiovisivo delle piattaforme e dei *synthetic media* – e

da un uso "strumentale" dell'IA a una progettazione esplicita della conversazione come pratica di co-interpretazione assistita, pensata per essere trasferibile e scalabile attraverso reti e programmi (come, ad esempio, CIpS).
Con co-interpretazione assistita intendiamo il dispositivo metodologico che, nel progetto, metterà in continuità (i) l'esperienza precedente di RAG vincolato a fonti curate sperimentata in ambito didattico, (ii) la nuova sfida dei *synthetic media* come problema pubblico nell'ecosistema delle piattaforme, e (iii) un impianto di public engagement in cui scuole e comunità non sono destinatari ma co-produttori di criteri. Il *chatbot* (così come gli eventuali altri strumenti *AI-based* che saranno sviluppati nel corso della ricerca) diventa un mediatore che sostiene la discussione: rende visibili evidenze e contesti disponibili nella *knowledge base*, esplicita i limiti e guida l'argomentazione.
L'idea di fondo è che, in merito ai *synthetic media*, l'obiettivo educativo e civico non sia quello di certificare automaticamente l'autenticità di un prodotto, ma costruire condizioni affinché pubblici eterogenei possano argomentare in modo informato, riconoscere i limiti della conoscenza disponibile, formulare ipotesi verificabili e distinguere tra evidenza e interpretazione in un ecosistema in accelerazione.

5. ACCESSIBILITÀ: ETICA E VALUTAZIONE

La fiducia dei partecipanti e delle istituzioni dipende dalla trasparenza delle regole e dalla comprensibilità dei limiti del sistema. Per questo, la conformità al GDPR, l'uso di tecniche di anonimizzazione e protocolli di sicurezza per garantire confidenzialità e integrità dei dati, insieme a una forte attenzione ai rischi di *bias* e disinformazione, sono al centro del progetto. L'inclusione assume un ruolo cardine, perché costituisce una condizione di efficacia dell'intero impianto. Di fatto, l'attenzione alla *user experience* e alle funzioni inclusive (*text-to-speech*, supporti grafici) passa da obiettivo progettuale a strumento di mediazione e, al tempo stesso, di lettura della partecipazione. Strumenti fondamentali per il sostegno di studenti con SLD, in questo caso essi consentono anche di osservare come differenti profili di accessibilità incidano sulla qualità della partecipazione e sulla costruzione del giudizio critico, e se (e in che misura) la riduzione delle barriere cognitive e comunicative favorisca pratiche argomentative più solide e una distinzione più consapevole tra evidenza e interpretazione. In un dispositivo di co-interpretazione, dunque, l'accessibilità non è un requisito periferico, ma un fattore che modella direttamente le possibilità di interazione.
Per queste ragioni l'accessibilità entra anche nelle metriche: la valutazione deve considerare non solo se il dispositivo funziona, ma per chi funziona e a quali condizioni. Poiché l'obiettivo non è massimizzare la fluidità della conversazione, ma la qualità della mediazione, la valutazione si concentrerà su quanto l'interazione faciliti pratiche critiche invece di sostituirle. Un primo asse riguarderà l'apprendimento critico (ovvero distinguere evidenze e interpretazioni, formulare ipotesi verificabili, riconoscere l'incertezza, motivare un giudizio); un secondo l'engagement (qualità e continuità della partecipazione, riuso di criteri e griglie di lettura, produzione di materiali restituiti); un terzo il comportamento dello strumento conversazionale (percentuale di risposte ancorate al *repository*, qualità dell'esplicitazione dell'incertezza, coerenza dei rifiuti motivati e, soprattutto, riduzione della delega epistemica al *device*).

6. CONCLUSIONI E SVILUPPI FUTURI

Gli sviluppi futuri seguono una traiettoria di scalabilità già inscritta nell'impianto delineato. Sul piano didattico, l'obiettivo è partire dall'implementazione degli strumenti sviluppati all'interno di percorsi di educazione al linguaggio audiovisivo, e costruire un percorso che intrecci *film/media literacy* e competenze d'uso dell'IA, assumendo come punto educativo centrale non la caccia all'errore, ma la sua valorizzazione. Anziché proporre un rilevamento sempre più performante dell'allucinazione o del *fake*, si intende considerare l'errore come occasione di apprendimento critico.
L'esperienza in classe ha mostrato che, di fronte a risposte plausibili ma non sempre affidabili, gli studenti sono stati spinti a interrogare fonti, contesti e presupposti, sviluppando una postura riflessiva verso lo strumento e i contenuti: l'*output* dell'IA diventa un oggetto mediale da analizzare e discutere, non un'autorità. In questo senso l'IA è funzionale a percorsi di educazione ai media e con i media (Buckingham, 2019), perché rende visibili condizioni di produzione e circolazione dei messaggi e allena alla verificabilità, all'argomentazione e alla consapevolezza dei limiti degli strumenti adottati.
Obiettivo complessivo è stabilizzare un *framework* riusabile: un kit di protocolli laboratoriali, griglie di analisi, *knowledge base* tematiche e linee guida di progettazione conversazionale, adattabile ad altre

didattiche (dalla storia alla comunicazione scientifica), mantenendo i principi di trasparenza epistemica, inclusione e controllabilità e la possibilità di aggiornamento iterativo. Sul piano territoriale, la sperimentazione mira a consolidarsi attraverso reti e partner inizialmente locali, per poi estendersi in modo interregionale valorizzando le collaborazioni inter-accademiche attive (Salento, Bologna, Bari, Cattolica Milano), su cui fonda il *network*, e la circolazione di competenze e materiali tra contesti regionali differenti, anche in vista di nuove interfacce sociotecniche (es. robotica di consumo).
La sostenibilità a medio periodo dipende dalla capacità di collocare il modello dentro infrastrutture nazionali già orientate all'adozione scolastica. In questa direzione, l'innesto in CIpS e la collaborazione con attori istituzionali consentono di testare sia la replicabilità sia l'impatto in scuole e contesti diversi, e di promuovere il riuso di pratiche e materiali oltre il singolo pilota, con un orizzonte di trasferibilità nazionale. La validazione della scalabilità passerà attraverso un piano di valutazione multi-asse (apprendimento critico, qualità dell'engagement, fiducia informata; e, lato sistema, ancoraggio alle fonti e comportamento del sistema conversazionale), rendendo confrontabili risultati e *trade-off* in diversi contesti applicativi.

RINGRAZIAMENTI

Il lavoro è svolto nell'ambito del progetto *LUM.IA – Learning and Understanding Media through Artificial Intelligence*, finanziato dal Ministero dell'Università e della Ricerca nell'ambito del Bando FIS 3 (Fondo Italiano per la Scienza, 2024).

BIBLIOGRAFIA

Boccia Artieri G., Gemini L., Pasquali F., Carlo S., Farci M., Pedroni M. (2017). Fenomenologia dei social network. Milan: Guerini.

Boyd, D. (2007). Why youth (heart) social network sites: The role of networked publics in teenage social life. In D. Buckingham (Ed.), Youth, identity, and digital media (pp. 119–142). The MIT Press.

Buckingham, D. (2019). The Media Education Manifesto. Cambridge: Polity Press.

Chesney, R. M., & Citron, D. K. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. California Law Review, 107: 1753–1819.

d'Ancona, M. (2017). Post-truth: The new war on truth and how to fight back. Ebury Press.

Farinacci, E. (2024). Towards a renewed understanding of screen and audiovisual education. Scholé, 62(1).

Floridi, L. (Ed.). (2015). The Onlife Manifesto. Cham: Springer.

Gao, Y., Xiong, Y., Diber, A., Wang, T., Hao, J., Pan, L., Wang, D., Liang, Z., Wang, R., Shu, L., et al. (2024). Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv preprint arXiv:2312.10997.

Gillespie, T. (2014). The Relevance of Algorithms. In Gillespie, T., Boczkowski, P. J., & Foot, K. A. (Eds.). (2014). Media technologies: Essays on communication, materiality, and society (Inside Technology). The MIT Press.

Groza, A., & Marginean, A. (2023). Brave new world: ai in teaching and learning. In ICERI2023 Proceedings. IATED: 8706–8713.

Jenkins, H. (2006, October 19). Confronting the challenges of participatory culture: Media education for the 21st century (Part one). Pop Junctions: Reflections on Entertainment, Pop Culture, Activism, Media Literacy, Fandom and More. https://henryjenkins.org/blog/2006/10/confronting_the_challenges_of.html

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Stadler, M., Weller, J., Kuhn, J., & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103, 102274.

Keyes, R. (2004). The post-truth era: Dishonesty and deception in contemporary life. St. Martin's Press.

Lessig, L. (2008). Remix. New York: Penguin Press.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In Advances in Neural Information Processing Systems 33 (NeurIPS 2020).

Li, Y., Wang, S., Yang, R., Liu, Y., & Zeng, D. (2025). Retrieval-augmented generation for educational application: A systematic survey. Computers and Education: Artificial Intelligence, 8, 100360.

Mauceri, S., & Di Censi, L. (2020). Adolescenti iperconnessi. Rome: Armando.

Prosperi, G., Miglietta, G., Miccoli, E., & Bochicchio, M. (2025). Tecnologie AI per la didattica. In S. Rebora, M. Rospocher, & S. Bazzaco (Eds.), Diversità, equità e inclusione: Sfide e opportunità per l'informatica umanistica nell'era dell'intelligenza artificiale. Proceedings del XIV Convegno Annuale AIUCD2025 (Verona, 11–13 giugno 2025). AIUCD.

Swacha, J., & Gracel, M. (2025). Retrieval-Augmented Generation (RAG) Chatbots for Education: A Survey of Applications. Applied Sciences, 15(8), 4234.

Theophilou, E., Koyutürk, C., Yavari, M., Bursic, S., Donabauer, G., Telari, A., et al. (2023). Learning to prompt in the classroom to understand AI limits: a pilot study. In International Conference of the Italian Association for Artificial Intelligence. Cham: Springer Nature Switzerland: 481–496.

Tirocchi, S., & Serpieri, R. (2020). Piattaforme, adolescenti e forme di soggettivazione. Mediascapes Journal, 16 (2020).

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? International Journal of Educational Technology in Higher Education, 16, 39.

[6/A] Modellazione testuale e linguistica computazionale

Verso un dizionario narrativo computazionale degli usi linguistici in aree ad alta vulnerabilità sociale

MICHELA BANDINI, SILVIA PICCINI, ANDREA BELLANDI, EMILIANO GIOVANNETTI
CNR-ISTITUTO DI LINGUISTICA COMPUTAZIONALE "A. ZAMPOLLI", ITALY – NAME.SURNAME@ILC.CNR.IT

ABSTRACT (ITALIANO)

Questo contributo presenta la costruzione di un dizionario narrativo computazionale del linguaggio giovanile in contesti di vulnerabilità sociale, sviluppato attraverso un corpus orale di interviste nell'area napoletana nell'ambito di un progetto di Ricerca-Azione. L'analisi integra approcci qualitativi e quantitativi per indagare le strategie definitorie dei parlanti, alla luce della distinzione bruneriana tra pensiero paradigmatico e narrativo. I risultati confluiscono in un dizionario non-standard pensato come una rete concettuale formalizzata secondo il modello OntoLex-Lemon, in linea con i principi FAIR dei Linguistic Linked Open Data.
Parole chiave: dizionario narrativo; linguaggio giovanile; corpus orale; lessico computazionale; OntoLex-Lemon

ABSTRACT (ENGLISH)

Toward a Computational Narrative Dictionary of Language Use in Areas of High Social Vulnerability. This paper presents the construction of a computational narrative dictionary of youth language in contexts of social vulnerability, developed from an oral corpus of interviews run in the Naples area as part of an Action-Research project. The analysis combines qualitative and quantitative approaches to investigate speakers' definitional strategies, in light of Bruner's distinction between paradigmatic and narrative thinking. The results are combined in a non-standard dictionary structured as a conceptual map formalised according to the OntoLex-Lemon model, in line with the FAIR principles of Linguistic Linked Open Data.
Keywords: narrative dictionary; youth language; oral corpus; computational lexicon; OntoLex-Lemon

1. INTRODUZIONE

Il presente contributo illustra le fasi di ideazione e una prima rappresentazione formale di un dizionario narrativo computazionale del linguaggio giovanile in contesti caratterizzati da elevata vulnerabilità sociale. Il lavoro si colloca nell'ambito del progetto sociale *Futuri (im)possibili: diagnosticare il presente e immaginare il futuro con i giovani di Caivano*, promosso dalla Fondazione Rut in collaborazione con la Fondazione Don Calabria per il Sociale, l'Istituto di Linguistica Computazionale "A. Zampolli" (CNR-ILC), l'Istituto di Ricerca su Innovazione e Servizi per lo Sviluppo (CNR-IRISS) e diverse realtà locali. Il dizionario, concepito come uno strumento di supporto alla comprensione dell'ipercomplessità (Morin, 1990) dei fenomeni antropo-sociali, è pensato per coloro che operano quotidianamente a fianco dei giovani – insegnanti, educatori, operatori sociali, psicologi – ai quali vuole fornire una base di conoscenza utile per progettare futuri interventi educativi e sociali.
L'approccio adottato è quello della Ricerca-Azione, metodo sviluppato dallo psicologo tedesco Kurt Lewin (1946) e ampiamente utilizzato negli studi sociologici in contesti fragili, dove emergono problemi che ostacolano il benessere collettivo. La ricerca non si limita, pertanto, all'osservazione dei fenomeni sociali, ma si configura come un processo dialogico e trasformativo, fondato sull'interazione continua tra teoria e prassi. I dati raccolti orientano l'azione e, al contempo, le azioni producono nuova conoscenza, in un circolo interpretativo dinamico, che si basa su un confronto tra gli attori del contesto di riferimento guidati da una specifica "idea generale" (Elliott, 1993). Proprio in virtù dei destinatari e delle finalità del progetto, il dizionario non è concepito come un elenco statico di significati denotativi, ma include associazioni connotative, interpretazioni ironiche e reti concettuali implicite. Si parte, infatti, dall'assunto che "definire" costituisca un processo cognitivo complesso, che non si riduce a imbrigliare il reale in una formula astratta dai confini rigidi. Nel tentativo di formulare una definizione spesso la mente mobilita reti associative, richiama esperienze vissute, esempi concreti e narrazioni personali. Detto in altri termini, come già sottolineato dallo psicologo statunitense Jerome Bruner (Bruner, 1986; Bruner, 1990), nel

definire entrano in gioco due modalità differenti e complementari che contribuiscono, entrambe, alla costruzione della realtà: da un lato il *pensiero paradigmatico* che tende alla formalizzazione e all'oggettività, e dall'altro il *pensiero narrativo* che fornisce interpretazioni situate, soggettive e culturalmente mediate, organizzate attorno all'esperienza personale e al sé. È proprio il significato narrativo a costituire il tratto distintivo del dizionario che qui presentiamo.
Prima di entrare nel dettaglio della sua struttura e delle scelte di modellazione adottate per la rappresentazione computazionale delle informazioni lessicali, presenteremo brevemente i contesti di raccolta dei dati linguistici, le modalità di acquisizione del materiale e il processo di costruzione del corpus che ha costituito la base empirica di analisi.

2. LA METODOLOGIA DELLA RICERCA: UN CONTRIBUTO LINGUISTICO

In linea con un approccio etnografico-sociolinguistico (Hymes, 1974), la prima fase della ricerca si è concentrata sulla raccolta di dati linguistici. Attraverso tavoli comunitari e consultazioni con educatori, assistenti sociali e operatori del territorio, sono stati individuati tre ambiti di intervento ritenuti rilevanti per osservare pratiche linguistiche spontanee, differenziati in base al grado di strutturazione dei contesti e alle modalità di coinvolgimento dei partecipanti. Nel *contesto strutturato*, rappresentato dall'ambiente scolastico, è stato adottato un approccio controllato, basato sulla somministrazione di questionari cartacei. Gli studenti sono stati coinvolti in attività di riflessione sul significato e sull'uso delle parole attraverso esercizi di comprensione e completamento, domande a scelta multipla e quesiti su scenari ispirati alla vita quotidiana. I *contesti semi-strutturati* hanno incluso ambienti come i doposcuola e i centri diurni, nei quali sono state privilegiate l'osservazione diretta, le interazioni informali e le attività di gruppo pianificate. Infine, i *contesti non strutturati* hanno compreso incontri in presenza con giovani, principalmente in spazi pubblici, nei quali i dati sono stati raccolti mediante questionari digitali e interviste orali registrate.
I contesti strutturato e semi-strutturato, esplorati nell'area di Afragola, hanno svolto una funzione di pre-test metodologico, finalizzata alla messa a punto degli strumenti di raccolta dati da adottare nella fase di test principale che si è svolta a Caivano, in un contesto non strutturato. In questo scenario, i volontari del *Centro "Le Salicelle" di Afragola - Istituto Don Calabria*, operanti come unità di prossimità (o *unità di strada*), hanno interagito con i giovani del territorio attraverso due principali modalità di raccolta dei dati: *i*) la somministrazione di un questionario digitale tramite *Google Form*; *ii*) la conduzione di interviste semi-strutturate (Adams, 2015), registrate in formato audio o video.
Il questionario digitale[1], diffuso tramite QR code, era composto da domande a risposta chiusa (a scelta multipla) e aperta, articolato in diverse sezioni, volte a raccogliere informazioni specifiche, quali: *i*) dati demografici generali (ad esempio età, genere, luogo di residenza); *ii*) significati attribuiti a una selezione di termini chiave (ad esempio 'capo', 'stupro', 'abuso' e 'ghetto'); iii) esempi di uso dei termini in contesti di vita quotidiana e associazioni semantiche. A questo si aggiungevano domande aperte finalizzate a stimolare riflessioni sui significati personali, sulle rappresentazioni culturali e sui potenziali stereotipi associati ai termini chiave.
Le interviste semi-strutturate sono state condotte dai volontari secondo un protocollo composto da sei domande principali, incentrate su tematiche chiave quali il successo, il genere e la legalità. Pur fornendo una base comune di riferimento — a tutti gli intervistati venivano richiesti età, luogo di origine e significato di termini come 'femmina', 'criminale' e 'infame' — il protocollo di intervista prevedeva una certa flessibilità, lasciando agli intervistatori la libertà di adattare o approfondire le domande in base alle esigenze situazionali e alle risposte dei partecipanti. In particolare, gli intervistatori erano incoraggiati a introdurre sollecitazioni esplorative quali *"Cosa intendi con questa parola?"*, *"Hai un esempio concreto?"*, *"Conosci qualcuno che per te rappresenta questa idea?"*, *"Dipende dal contesto?"* o *"Usi questa parola anche con i tuoi amici?"*, al fine di elicitare chiarimenti, sfumature contestuali ed esempi d'uso situati.
Tutte le interviste[2] sono state raccolte mediante registrazioni audio o, ove possibile, video, previa firma

[1] Ad oggi, sono state raccolte 38 risposte valide da partecipanti di età compresa tra i 12 e i 23 anni, con una netta prevalenza di rispondenti di genere femminile (oltre il 54%), a fronte di una percentuale inferiore al 40% di partecipanti maschi. Dal punto di vista demografico, oltre il 94% dei rispondenti risiede a Caivano, confermando l'efficacia della strategia di R-A territorialmente radicata adottata in questa fase del progetto.
[2] Tra il 24 aprile e l'8 giugno 2025 sono state complessivamente raccolte 58 interviste, un corpus orale di dimensioni significative. Dal punto di vista demografico, il campione comprende 32 partecipanti di genere femminile e 26 di genere maschile; la maggioranza degli intervistati (58 individui) risiede a Caivano, confermando ulteriormente il forte radicamento territoriale della raccolta dati. Per quanto riguarda la distribuzione anagrafica, i partecipanti hanno un'età

del consenso informato da parte dei partecipanti; nel caso di minori, sono state applicate rigorose misure di tutela della privacy, garantendo che non fossero registrati tratti facciali identificabili[3].

3. ANALISI PRELIMINARE DEI DATI: CREAZIONE DI UN DATASET

L'intero processo si è articolato in cinque fasi sequenziali, finalizzate a garantire una rigorosa preparazione del corpus e un'analisi accurata dei dati linguistici.

Fase 1. Le registrazioni audio e video (61 file audio, per un totale di 58 interviste, con una durata complessiva di 2 ore, 18 minuti e 12 secondi) sono state trascritte digitalmente mediante l'ausilio di *NotebookLM* di Google. Una successiva revisione manuale ha consentito di correggere gli errori, standardizzare la formattazione, identificare gli interlocutori e garantire l'anonimizzazione, a tutela della privacy dei partecipanti (vedi nota 3).

Fase 2. Le trascrizioni sono confluite in un corpus orale organizzato in modo sistematico, con pieno allineamento tra file audio/video, metadati e testi trascritti. A ciascun elemento del corpus sono stati assegnati identificatori univoci: le interviste sono state contrassegnate da ID progressivi (ad es. *int01–int58*), i partecipanti da ID corrispondenti (*P01–P58*) e gli intervistatori da identificativi dedicati (*I01–I03*). I materiali sono stati archiviati secondo una rigorosa convenzione di denominazione, che mette in relazione in modo biunivoco registrazioni, trascrizioni e file di metadati in formato JSON, preservando l'integrità di ciascuna intervista[4].

Fase 3. Una volta strutturato, il corpus ha consentito lo svolgimento di analisi descrittive relative al profilo demografico dei partecipanti e di analisi quantitative sui dati linguistici, utili ad acquisire una panoramica complessiva del dataset, già illustrata nella sezione precedente (note 1 e 2).

Fase 4. È stata quindi condotta un'analisi combinata quantitativa e qualitativa, con particolare attenzione al comportamento terminologico e alle strategie definitorie adottate dai partecipanti. Le analisi sui singoli termini sono state realizzate mediante matrici a doppia entrata, con l'obiettivo di indagare non solo la distribuzione dei significati, ma anche le modalità cognitive attraverso cui i partecipanti hanno espresso il contenuto semantico delle parole (se, ad esempio, attraverso sinonimi o strategie più complesse come la parafrasi esplicativa che enuclea i tratti distintivi del concetto). Le risposte sono state classificate sulla base del quadro binomiale di Bruner illustrato nell'introduzione: le definizioni di tipo scientifico sono state associate ai significati denotativi e interpretate come manifestazioni di pensiero paradigmatico, mentre tutte le altre risposte — incluse le associazioni connotative positive e negative — sono state analizzate come espressioni di pensiero narrativo. Nel loro insieme, le analisi tengono conto in modo integrato dei dati provenienti dai questionari e dalle interviste orali raccolte a Caivano.

Il termine 'femmina' offre un esempio emblematico di tale rappresentazione. Da un lato, i partecipanti hanno attivato una modalità di ragionamento paradigmatica, producendo significati denotativi coerenti con le definizioni repertoriate nei dizionari (ad esempio, 'femmina' intesa come "sesso biologico opposto al maschio, atto alla riproduzione"). Tuttavia, nella maggioranza dei casi, attraverso il pensiero narrativo la parola è stata associata a significati connotativi positivi o negativi. I partecipanti hanno fatto spesso ricorso a esperienze autobiografiche, anche mediante esempi prototipici (ad esempio, 'femmina' come "mia madre che fa tutto per me" o associazioni con pratiche quali il vestirsi in modo curato o l'uso del trucco). Il significato è stato organizzato soggettivamente attorno al sé, facendo emergere l'importanza della dimensione culturale: il termine 'femmina' ha dato vita a rappresentazioni socialmente condivise, stereotipizzazioni, giudizi negativi (inferiorità, violenza o sessualizzazione), nonché a narrazioni legate all'identità ("sentirsi donna dentro"). Le associazioni semantiche sono state ricavate direttamente dai testi prodotti dai partecipanti. Gli esempi testuali originali, con le relative frequenze, sono stati conservati e associati alle tipologie definitorie, consentendo analisi qualitative e quantitative. In una fase successiva si

compresa tra i 12 e i 30 anni, con una chiara concentrazione nella fascia 14–17 anni, in linea con il focus primario del progetto sull'uso linguistico adolescenziale.

[3] La raccolta e gestione dei dati linguistici è stata condotta nel rispetto del General Data Protection Regulation (GDPR), con particolare attenzione alle specificità del contesto di ricerca, caratterizzato dalla presenza di partecipanti minorenni e situazioni di vulnerabilità sociale. I dati sono stati sottoposti a procedure di anonimizzazione e conservati in ambienti protetti, accessibili esclusivamente ai membri del progetto.

[4] Il corpus è attualmente conservato in un repository di progetto privato, organizzato per tipologia di dati (audio, trascrizioni, metadati) e documentato tramite un file README. I metadati adottano uno schema personalizzato basato su JSON, sviluppato in funzione delle esigenze specifiche del progetto. Le fasi future prevedono sia il rilascio pubblico del corpus su piattaforme dedicate, come il CLARIN Virtual Language Observatory (VLO), sia la mappatura dei metadati verso standard comunitari per i corpora orali, come Dublin Core, al fine di garantire interoperabilità e riutilizzo a lungo termine.

procederà all'annotazione sistematica del corpus, che permetterà interrogazioni combinate sui dati (cfr. *infra*).

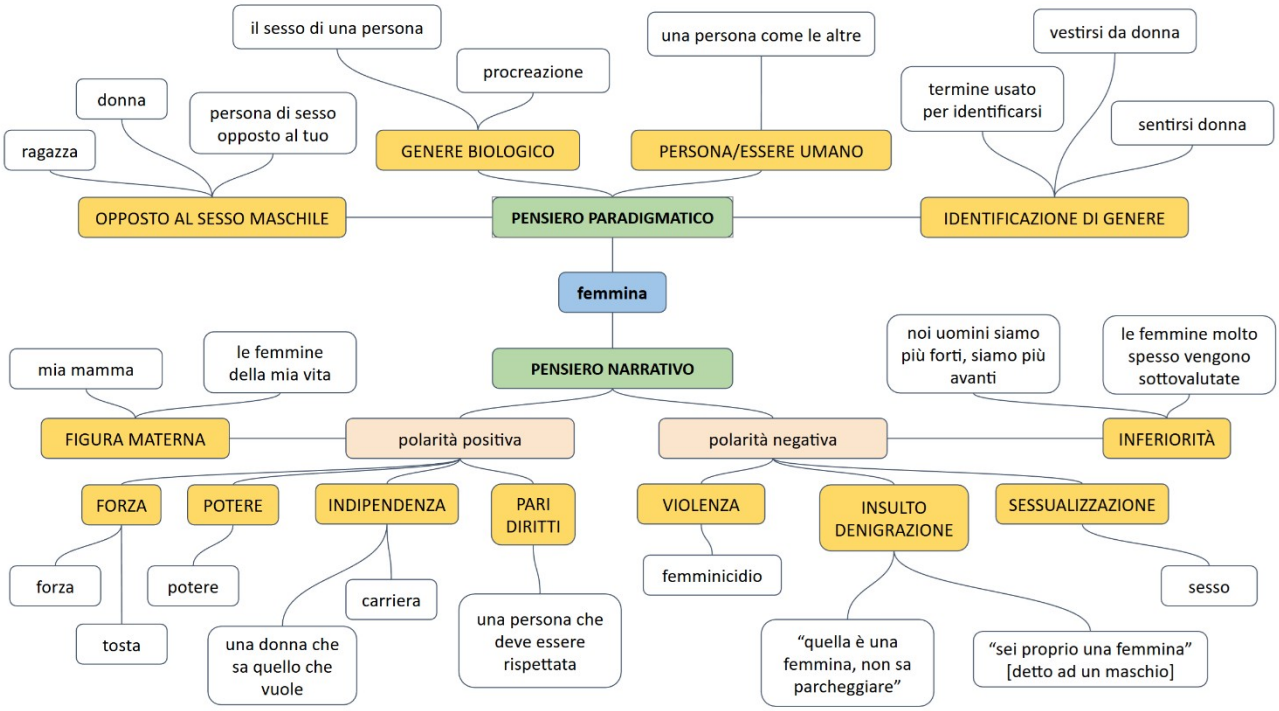


Figura 1. La voce lessicale 'femmina'

Fase 5. Nella fase finale dell'analisi è stato definito il modello di voce lessicografica più adeguato a rappresentare i dati linguistici emersi dalle interviste e dai questionari. Se i dizionari standard tendono a registrare principalmente il significato denotativo delle parole, presentandolo attraverso definizioni paradigmatiche fondate su rapporti di sinonimia e inclusione semantica, un dizionario concepito come strumento di intervento in contesti ad alta vulnerabilità sociale deve invece porre al centro il significato narrativo. Ciò implica dare rilievo alle narrazioni, alle associazioni positive e negative e agli esempi socialmente situati evocati dai giovani intervistati, attraverso i quali il significato si costruisce e si negozia all'interno dell'esperienza vissuta.

Ciascuna voce si articola pertanto in due macro-dimensioni: una paradigmatica e una narrativa, quest'ultima ulteriormente organizzata in base alla polarità (positiva o negativa) e ai concetti evocati dai parlanti. La struttura reticolare che ne deriva, esemplificata in Figura 1, configura le voci come reti di relazioni semantiche.
Al fine di rappresentare e sfruttare appieno questa articolata struttura a "grafo", sia per la consultazione da parte di utenti sia per l'elaborazione automatica dei dati, il dizionario è stato concepito nativamente come risorsa computazionale. La proposta di rappresentazione formale adottata verrà discussa nel paragrafo successivo.
Allo stato attuale, il dizionario comprende 16 voci lessicali, costituite sia da termini monolemmatici sia da espressioni colloquiali o polirematiche ('aborto', 'bastardo', 'capo', 'criminale', 'essere malato', 'essere uno schiaffo', 'fare abuso', 'femmina', 'ghetto', 'guardia', 'infame', 'rispetto', 'si 'na petra', 'sistema', 'soldi', 'stupro'). A ciascuna voce sono associate informazioni morfologiche di base.

4. VERSO UNA RAPPRESENTAZIONE FORMALE DEL LESSICO

La formalizzazione presentata in questa sezione costituisce una prima proposta di modellazione dei dati lessicali emersi dall'analisi del corpus, con particolare attenzione alla rappresentazione del significato narrativo, che rappresenta l'elemento di maggiore originalità del contributo. Al fine di garantire trovabilità, accessibilità, interoperabilità e riuso, in linea con i principi FAIR (Wilkinson et al., 2016), nonché l'integrazione con l'ecosistema dei *Linguistic Linked Open Data* (Cimiano et al., 2020), si è scelto di adottare il modello *OntoLex-Lemon*[5] (McCrae et al., 2017), ampiamente utilizzato per la

[5] https://www.w3.org/2016/05/ontolex/ (cons. 22/04/2026).

rappresentazione di risorse lessicali. La sua flessibilità consente di modellare forme di significato non standard, in particolare quelle legate alla dimensione narrativa, difficilmente rappresentabili attraverso strutture lessicografiche tradizionali. Per la formalizzazione delle parole considerate e dei significati loro attribuiti è stato utilizzato il modulo "core", denominato *ontolex*. La Figura 2 riporta un esempio di formalizzazione, serializzato in formato *Turtle*.

```
01  :fem_entry a ontolex:LexicalEntry ;
02      lexinfo:partOfSpeech lexinfo:noun ;
03      lexinfo:gender lexinfo:feminine ;
04      ontolex:canonicalForm :lex_femmina ;
05      ontolex:sense :fem_sense1, :fem_sense2 .

06  :lex_fe a ontolex:Form ;
07      ontolex:writtenRep "femmina"@it ;
08      lexinfo:number lexinfo:singular .

09  :fem_sense1 a ontolex:LexicalSense ;
10      ontolex:reference :WOMAN ;
11      skos:definition "Essere di sesso opposto a
           quello maschile, atto a procerare."@it ;
12      dc:type :paradigmatic .

13  :fem_sense2 a ontolex:LexicalSense ;
14      skos:definition "simbolo di forza, potere."@it ;
15      dc:type :narrative .

16  :meaningModalityScheme a skos:ConceptScheme ;
17     skos:prefLabel "Schema delle modalità di costruzione del
                        significato"@it .

18  :paradigmatic a skos:Concept ;
19     skos:prefLabel "paradigmatic"@en , "paradigmatico"@it ;
20     skos:definition "Senso riconducibile a una costruzione
          paradigmatica del significato: contenuto denotativo e
          definitorio, allineato a una concettualizzazione
          lessicografica o scientifica."@it ;
21     skos:inScheme :meaningModalityScheme .

22  :narrative a skos:Concept ;
23     skos:prefLabel "narrative"@en , "narrativo"@it ;
24     skos:definition "Senso riconducibile a una costruzione
          narrativa del significato: contenuto situato, valutativo
          ed esperienziale, modellato dal contesto sociale e
          culturale."@it ;
25     skos:inScheme :meaningModalityScheme .
```

Figura 2. Formalizzazione in OntoLex-Lemon dell'entrata 'femmina': serializzazione in Turtle.

L'entrata 'femmina' è stata rappresentata come istanza della classe ontolex:LexicalEntry (riga 01), ossia un'unità astratta che raggruppa le informazioni linguistiche e semantiche associate al termine. L'entrata è descritta tramite proprietà morfosintattiche di base, come la parte del discorso e il genere grammaticale, ed è collegata alla sua forma canonica, modellata come ontolex:Form con la corrispondente rappresentazione scritta.
L'aspetto centrale dell'esempio riguarda la rappresentazione dei due tipi di significato introdotti in precedenza, e in particolare la distinzione tra un senso "paradigmatico" e uno "narrativo". Per rappresentarli si è scelto di non estendere il modello *OntoLex-Lemon* (ad esempio introducendo sottoclassi di ontolex:LexicalSense), ma di ricorrere a una formalizzazione di questi due elementi come concetti SKOS[6] (Simple Knowledge Organization System) esterni. Questo approccio riflette la filosofia portante dei *Linked Data*, volta a limitare la proliferazione di entità *ad hoc* e favorire il riuso di ontologie standardizzate.
In dettaglio, i due tipi di senso vengono rappresentati mediante l'utilizzo di due concetti SKOS (skos:Concept): "paradigmatic" e "narrative", a loro volta descritti come appartenenti a un skos:ConceptScheme che rappresenta l'insieme delle possibili modalità di costruzione del significato (righe 16-17). Attraverso l'utilizzo del vocabolario di metadatazione Dublin Core Terms[7], e più specificatamente della proprietà dct:type, i tipi vengono assegnati opportunamente ai rispettivi sensi lessicali (righe 12 e 15). La scelta di utilizzare il predicato dct:type risiede nella sua versatilità semantica: l'assenza di vincoli di dominio (*domain*) e un intervallo (*range*) esteso a rdfs:Class rendono dct:type formalmente compatibile con il riferimento a concetti SKOS. Negli sviluppi della ricerca, in ottica di raffinamento semantico, valuteremo la possibilità di introdurre una proprietà dedicata definita come rdfs:*subProperty* di dct:type.
Per ragioni di chiarezza espositiva, nell'esempio relativo all'entrata 'femmina' vengono riportati soltanto due sensi, rappresentativi delle due tipologie individuate — uno paradigmatico e uno narrativo — modellati come due istanze di ontolex:LexicalSense. Il primo corrisponde al significato standard, denotativo, di 'femmina', come potrebbe essere riportato in un dizionario. Questo senso è stato collegato a un concetto (<WOMAN>[8]), inserito a titolo di esempio, appartenente a un'ontologia di conoscenza condivisa di riferimento. Il secondo senso, invece, formalizza uno degli usi colloquiali individuati

[6] https://www.w3.org/TR/skos-reference/ (cons. 22/04/2026).
[7] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/ (cons. 22/04/2026).
[8] I concetti sono indicati tra parentesi uncinate (< >) al fine di distinguerli dai termini lessicali, riportati tra apici singoli (' ').

attraverso le interviste. La descrizione formale di questo tipo di sensi "narrativi", qui presentata in forma preliminare, è attualmente oggetto di ulteriore elaborazione. Sulla base delle analisi condotte, tali sensi non dovranno essere concepiti come semplici accezioni aggiuntive, ma come nodi semantici collegati a concetti lessicali e valutativi specifici, suscettibili di essere qualificati secondo differenti gradi e orientamenti di polarità. Nel caso dell'entrata 'femmina' illustrata nell'esempio, il senso narrativo potrà dunque essere messo in relazione a concetti quali <FORZA> o <POTERE>, attraverso collegamenti caratterizzati da polarità positiva. La formalizzazione di questa componente del modello seguirà le *best practice* dei *Linked Data* e sarà orientata al riuso di vocabolari esistenti, selezionati in funzione della loro adeguatezza a rappresentare dimensioni valutative, concettuali e relazionali del significato.

5. CONCLUSIONI

Il dizionario narrativo degli usi linguistici in aree ad alta vulnerabilità sociale si configura come una risorsa computazionale integrabile nell'ecosistema dei *Linguistic Linked Open Data* e conforme ai principi FAIR, in grado di rendere esplicite le relazioni semantiche tra termini e i molteplici concetti che essi evocano. Le fasi successive della ricerca prevedono un ulteriore affinamento della formalizzazione, attraverso un'annotazione sistematica del corpus e una modellazione più dettagliata delle relazioni semantiche e del loro valore. Parallelamente, i dati verranno resi progressivamente fruibili mediante strumenti di *data visualization* ed esplorazione interattiva. In questa prospettiva, il dizionario si configura come una risorsa operativa e accessibile, a supporto di interventi educativi, sociali e di ricerca interdisciplinare, in linea con i principi della Ricerca-Azione che ne costituiscono il quadro metodologico di riferimento.

BIBLIOGRAFIA

Adams, W. C. (2015). Conducting semi-structured interviews. In J. S. Wholey, H. P. Hatry & K. E. Newcomer (Ed. 4), Handbook of Practical Program Evaluation (pp. 492–505). John Wiley & Sons. https://doi.org/10.1002/9781119171386.ch19

Bruner, J. (1986). Actual Minds, Possible Worlds. Harvard University Press. http://www.jstor.org/stable/j.ctv26070s7

Bruner, J. (1990). Acts of meaning: Four lectures on mind and culture (Vol. 3). Harvard University Press

Cimiano, P., Chiarcos, C., McCrae, J. P., & Gracia, J. (2020). Linguistic linked data. Springer International Publishing

Elliott, J. (1993). La ricerca azione: un quadro di riferimento per l'autovalutazione nelle scuole. Pozzo G. & Zappi L.(a cura di) Elliott, Giordan, Scurati. La ricerca azione. Metodiche, strumenti, casi. Torino, Bollati Boringhieri, 95-120.

Hymes, Dell. (1974). Foundations in Sociolinguistics: An Ethnographic Approach. Philadelphia: University of Pennsylvania Press.

Lewin, K. (1946). Action research and minority problems. Journal of social issues, 2(4), 34-46.

McCrae, J. P., Bosque-Gil, J., Gracia, J., Buitelaar, P., & Cimiano, P. (2017). The OntoLex-Lemon model: Development and applications. In Proceedings of eLex 2017.

Morin, E. (1990). Introduction à la pensée complexe. Paris : Ed. Du Seuil

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg N. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific data 3.

Verso la creazione di una Treebank per il sardo

NICOLETTA PUDDU[1], MANUELA SANGUINETTI[1], LUIGI TALAMO[2]

[1]UNIVERSITÀ DEGLI STUDI DI CAGLIARI
[2]UNIVERSITY OF SAARBRÜCKEN

ABSTRACT (ITALIANO)

In questo contributo intendiamo presentare una prima proposta per la creazione di una treebank per il sardo secondo il framework delle Universal Dependencies (UD). Verranno illustrate le considerazioni teoriche introduttive, i principi alla base della selezione dei testi, le fasi di tokenizzazione e annotazione e i primi risultati di training di un annotatore sintattico. Delineeremo quindi i passi futuri per l'estensione dell'annotazione.
Parole chiave: Universal Dependencies, sardo, lingue di minoranza, annotazione morfosintattica

ABSTRACT (ENGLISH)

This paper presents an initial proposal for the development of a Sardinian treebank within the Universal Dependencies (UD) framework. We outline the theoretical motivations informing the design of the resource, the criteria adopted for corpus construction and text selection, and the methodological workflow implemented for tokenisation and morpho-syntactic annotation. The contribution also reports preliminary evaluation results from the first training cycle of a UD-compliant syntactic parser, providing insights into the feasibility and future directions of large-scale syntactic annotation for Sardinian.
Keywords: Universal Dependencies; Sardinian; Less resourced languages; POS-tagging

1. INTRODUZIONE

In questo contributo intendiamo presentare una prima proposta per la creazione di una *treebank* per il sardo secondo il sistema di annotazione delle Universal Dependencies (UD). Nonostante esistano infatti diverse treebank per le lingue romanze, non ne esistono per il sardo e, più in generale, sono ancora poche quelle per lingue romanze minoritarie, in particolare se prive di standardizzazione. Descriveremo quindi le prime fasi di una creazione di una treebank per il sardo, evidenziando le principali criticità incontrate e delineando soluzioni possibilmente estensibili anche ad altre varietà di minoranza.

2. LA LINGUA SARDA E LE SUE VARIETÀ

Il sardo è una lingua romanza parlata nell'isola di Sardegna tradizionalmente suddivisa in quattro varietà: campidanese, logudorese, nuorese e arborense (Virdis, 1988). Tali varietà sono però riconducibili a due macrovarietà divise da un fascio di isoglosse in molti casi sovrapponibili. Le due macrovarietà vengono classificate come varietà distinte in ISO 639-3 con le denominazioni di campidanese (src), varietà del centro-sud, e logudorese (sro), varietà del centro-nord. Altre due varietà che ISO 639-3 identifica come sarde (gallurese e sassarese) non sono in realtà ritenute in letteratura come appartenenti al sardo, ma piuttosto affini a varietà corse e toscane e non sono state pertanto considerate ai fini della costituzione della nostra Treebank.
Dal 2006 esiste inoltre una varietà scritta per i documenti in uscita dall'Amministrazione regionale, denominata Limba Sarda Comuna (LSC) elaborata dalla Regione autonoma della Sardegna. La stessa Regione, nel 2022 ha promulgato, nell'ambito della Certificazione sperimentale della lingua sarda, dei criteri ortografici orientativi per la lingua sarda.

3. STATO DELL'ARTE

La scarsità di risorse annotate è un problema comune alle varietà di minoranza in Italia. Segnaliamo comunque che per il ligure[1], il napoletano[2] e il siciliano[3] sono state realizzate e rese disponibili nel repository di UD delle treebank.
Non esistono al momento corpora di riferimento in lingua sarda, nemmeno non annotati. Risulta attualmente in preparazione un corpus di parlato di emigranti sardi (Mura et al., 2023) sul quale è stato creato un *parts-of-speech* (POS) *tagger* automatico per il sardo che utilizza un *language model* neurale basato su BERT (Carta et al., 2025). Esistono inoltre alcune risorse per le varietà storiche: segnaliamo l'archivio testuale ATLISOR (Fortunato & Ravani, 2015) un corpus non annotato che raccoglie testi di età medioevale, e il corpus di testi sardi di epoca moderna EModSar (Puddu & Talamo, 2020), attualmente annotato per POS. Primi tentativi di annotazione sia per POS che sintattica erano stati fatti per il sardo moderno da Puddu & Stein (2018) tramite il software Arborator (Gerdes, 2013).

4. LA CREAZIONE DEL CORPUS

La prima scelta che abbiamo dovuto operare nella ideazione della treebank è stata quella della varietà di sardo da utilizzare per il corpus di partenza. Ciò che qui interessa notare è che le differenze tra le varietà sono perlopiù di tipo fonetico e lessicale, mentre gli studiosi concordano sulla maggiore unità del sardo a livello morfosintattico. A ciò si aggiunga che la discussione sulla effettiva esistenza di due macrovarietà è piuttosto articolata, dato che, al confine tra le due varietà, è possibile individuare un'area mediana che presenta tratti di entrambe le macrovarietà (Hajek & Goebl, 2021; Loporcaro & Putzu, 2024).
La nostra ipotesi di lavoro è pertanto che sia possibile creare un annotatore a livello morfosintattico e un'unica treebank per tutte le varietà di sardo.
Per quanto riguarda la scelta dei testi, nella creazione di corpora di lingue minoritarie si fa spesso largo uso del web, anche quando, come nel caso del sardo, siano disponibili testi a stampa. Il sardo è documentato tra le lingue utilizzate sul web da W3CTechs, anche se, secondo una ricerca condotta da Soria et al. (2018), i parlanti non sono solitamente a conoscenza di media e servizi in sardo, anche quando questi siano disponibili.
Uno dei problemi nel caso di lingue non standardizzate, è quello di identificare la subvarietà alla quale i testi appartengono in modo da creare un corpus per quanto possibile bilanciato.
A questo proposito, le pagine di Wikipedia in sardo risultano particolarmente informative, dato che indicano la varietà nella quale sono redatte: limba sarda comuna (LSC), logudorese, campidanese e nuorese.
Per la creazione del corpus sono stati scelti tre testi: due pagine di Wikipedia (una in LSC e una in campidanese) e una fiaba in dorgalese, subvarietà del nuorese orientale e meridionale per un totale di 3862 token.

5. ANNOTAZIONE

Si è partiti da un'annotazione automatica dei tre testi effettuata tramite il parser Stanford Stanza (Qi et al. 2020) versione 1.10.1, usando il modello per l'italiano, addestrato su quattro treebank UD: ISDT, VIT, PoSTWITA, e TWITTIRO. Il modello–ha segmentato in frasi, tokenizzato, lemmatizzato e annotato per *parts of speech* (POS), caratteristiche morfologiche e dipendenze sintattiche. In assenza di modelli pre-addestrati per il sardo, si è scelto di utilizzare un modello disponibile per l'italiano in virtù delle similarità morfosintattiche tra le due lingue, al fine di partire da una versione pre-annotata dei testi che, pur non essendo accurata dal punto di vista dell'annotazione linguistica, rispecchiasse le specifiche del formato CoNLL-U.
Due parlanti native e linguiste hanno poi corretto in forma collaborativa l'output dell'annotazione, utilizzando la piattaforma Arborator-Grew (Guibon, 2020). La correzione è stata condotta in maniera indipendente e poi controllata per verificarne l'uniformità. Per l'attuale proposta, sono state corrette 65 frasi per un totale di 1280 token. Tokenizzazione e segmentazione in frasi non hanno avuto necessità di particolari interventi.

[1] https://github.com/UniversalDependencies/UD_Ligurian-GLT
[2] https://github.com/UniversalDependencies/UD_Neapolitan-RB
[3] https://github.com/UniversalDependencies/UD_Sicilian-STB

La lemmatizzazione ha invece richiesto una fase di normalizzazione. La variabilità fonetica tra le diverse varietà ha imposto il ricorso ad una normalizzazione prevalentemente ortografica, sulla base dei criteri proposti per la LSC e per la Certificazione provvisoria sperimentale della conoscenza del sardo. A livello morfologico sono state operate delle normalizzazioni, quali la riduzione delle forme di infinito in *-are* e l'uscita dei nomi maschili in *-e*. Sottolineiamo però che la normalizzazione ha riguardato unicamente il lemma e non la forma ed è fondamentale sia per la creazione di un lemmario unico, sia per future ricerche di tipo sintattico estese a tutto il sardo.
Per quanto riguarda il *POS-tagging*, il *tagset* utilizzato (cfr. tabella 1) è Universal Parts of Speech (UPOS) di Universal Dependencies (de Marneffe et al., 2021), standard ampiamente consolidato e utilizzato anche per lingue romanze di minoranza come l'occitano. Al momento, non è stata corretta l'annotazione per caratteristiche morfologiche.

Open class	Closed class	Other
ADJ	ADP	PUNCT
ADV	AUX	SYM
INTJ	CCONJ	
NOUN	DET	
PROPN	NUM	
VERB	PART	
	PRON	
	SCONJ	

Tabella 1 Il tagset per POS utilizzato nella Treebank

Anche per quanto riguarda le relazioni sintattiche, sono stati utilizzati i tipi definiti nella versione 2 di UD, indicati nella tabella 2.

	Nominals	**Clauses**	**Modifier words**	**Function words**
Core arguments	nsubj ob iobj	csubj ccomp xcomp		
Non-core dependents	obl vocative expl dislocated	Advcl	Advmod discourse	aux cop mark
Nominal dependents	nmod appos nummod	acl	Amod	det clf case
Coordination	**Headless**	**Loose**	**Special**	**Other**
conj cc	fixed flat compound	list parataxis	Orphan goeswith reparandum	punt root dep

Tabella 2. I tipi di relazione sintattica marcati nella Treebank

6. ADDESTRAMENTO DI UNA NLP PIPELINE COMPLETA PER IL SARDO

Le 65 frasi sono state utilizzate per addestrare una NLP pipeline completa di Stanford Stanza per il sardo, in maniera da poter annotare in maniera automatica ulteriori frasi e poi correggerne l'output. La pipeline comprende modelli per i seguenti *task* di *natural language processing*: tokenizzazione e segmentazione in frasi, *multi-word token expansion* (mwt), lemmatizzazione, annotazione per *part-of-speech (POS)* e caratteristiche morfologiche e annotazione sintattica. Stanford Stanza richiede l'utilizzo di *word embeddings* per gli ultimi due task. A tale scopo, abbiamo utilizzato nell'addestramento i *word*

vectors per il sardo forniti dal progetto Fasttext[4] (Bojanowski, Grave, Joulin e Mikolov 2017) ed estratti da Wikipedia.
Il set per l'addestramento è ripartito in *train*, *dev* e *test* come descritto nella seguente Tabella numero 3 (Set per l'addestramento).

Set	Token	Frasi
Train	911	52
Dev	110	6
Test	259	7

Tabella 3 – Set per l'addestramento

Come di consueto, il parser addestra il modello sulle frasi del set di *train*, aggiustando poi automaticamente gli iperparametri sul set di *dev*. Infine, una valutazione della bontà del modello è condotta sul set di *test*, utilizzando come riferimento le misure utilizzate nel task di valutazione CoNLL-U 2018 (Zeman et al., 2018).
Di seguito, discutiamo i risultati di questa valutazione per ciascun modello della pipeline.
Modello per la tokenizzazione e la segmentazione in frasi. Questo modello combina i task di divisione in token e in frasi. Il modello dimostra ottime performance nella tokenizzazione (F1 = 96.79), ma purtroppo le sue capacità di determinare l'inizio e la fine di una frase (*sentence-splitting*) risultano insufficienti (F1 = 58.82). La scarsa efficacia del sentence splitting è probabilmente dovuta anche alle caratteristiche di uno dei testi utilizzati per il training che presenta molte istanze di discorso diretto e numerosi incisi.
Modello per la *multi-word token expansion*. Il modello 'espande' (*expansion*) i token multi-parola, sia legati che intervallati da spazi bianchi, nei token costituenti, ad es. *Art Nouveau* o *Intamis chi*. Nonostante gli ottimi risultati raggiunti (F1 = 99.09), il modello non è probabilmente valutabile a causa dell'esiguo numero di token multi-parola presenti nel set di addestramento, soltanto quattro.
Modello per la lemmatizzazione. Il modello presenta buone performance (F1 = 87.27) nel ricondurre la forma al lemma.
Modello per l'annotazione per parts of speech e caratteristiche morfologiche. Il modello annota i token per UPOS e per caratteristiche morfologiche, descrivendo le categorie nominali, verbali e aggettivali del token. Nella parte relativa alle UPOS, il modello raggiunge performance discrete (78.18). I risultati sono anche discreti per l'annotazione morfologica (77.27): dobbiamo però notare che tale annotazione al momento non è stata corretta manualmente, ma riflette ancora quella del modello italiano (vedi sopra).
Modello per l'annotazione sintattica. Il modello annota i token per dipendenze sintattiche, individuando le relazioni di testa e dipendente e descrivendo la relazione sintattica. Dato l'esiguo set di annotazione, i risultati non sono purtroppo soddisfacenti, configurando questo modello come il tallone d'Achille della nostra *NLP* pipeline. Lo UAS (*Unlabeled Attachment Score*) è del 63.64%, ovvero il modello ha difficoltà nell'identificare la corretta testa sintattica, a prescindere dalla relazione sintattica, in almeno un token su tre; se prendiamo in considerazione anche la relazione sintattica (LAS: *Labeled Attachment Score*), il risultato scende a 50.91%, ovvero una volta su due il modello sbaglia ad annotare la relazione sintattica. Escludendo le parole funzionali e considerando solo aggettivi, nomi e verbi, i risultati sono ancora più deludenti: il CLAS (*Content-word Labeled Attachment Score*) è infatti del 34.43%. Infine, il MLAS (*Morpho-Syntactic Labeled Attachment Score*), una misura che combina parte del discorso, dipendenza sintattica e caratteristiche morfologiche, è del 29.51%.

7. CONCLUSIONI E PROSPETTIVE FUTURE

La creazione di risorse elettroniche per lingue di minoranza non standardizzate rappresenta una sfida in quanto, innanzi tutto, impone delle scelte nelle varietà da utilizzare sia per il corpus di partenza che per l'addestramento. Nel caso del sardo, abbiamo deciso di utilizzare diverse subvarietà per il corpus di partenza, normalizzando le forme a livello di lemma. La scelta appare per ora corretta, in quanto il modello presenta buone performance nel ricondurre forme appartenenti a subvarietà distinte ad un unico lemma. Nel prossimo futuro ci proponiamo di estendere l'addestramento del modello al maggior numero possibile di subvarietà.

[4] https://fasttext.cc/docs/en/crawl-vectors.html

Anche a livello di POS-tagging, il modello dimostra livelli di correttezza discreti, che potranno essere sicuramente migliorati tramite l'aggiunta delle feature morfologiche all'annotazione.
Il modello risulta decisamente meno performante nell'analisi sintattica. Uno dei motivi, come detto, potrebbe essere l'esiguo numero di frasi sinora annotate, per cui ci riserviamo di rivalutare gli esiti una volta incrementato tale numero. Condurremo inoltre un'analisi di tipo qualitativo sulla tipologia degli errori, per verificare se una percentuale non possa essere ricondotta a particolari strutture sintattiche del sardo non riconosciute dall'annotatore. L'obiettivo è quello di pubblicare la treebank in una release ufficiale di UD, presumibilmente a maggio 2026.

BIBLIOGRAFIA

Bojanowski, P., Grave, E., Joulin, A., Mikolov, T. (2017) Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5 (pp. 135-146). https://doi.org/10.48550/arXiv.1607.04606.

Carta, S. M., Concas, F., Fenu, G., Giuliani, A., Manca, M. M., Mura, P., & Pisano, S. (2025). A BERT-based Approach for Part-of-Speech Tagging in the Low-Resource Context of Sardinian. In C. Bosco, E. Jezek, M. Polignano, & M. Sanguinetti (A c. Di), *CLiC-it 2025 Italian Conference on Computational Linguistic*. https://aclanthology.org/2025.clicit-1.18/.

de Marneffe, M.-C., Manning, C., Nivre, J., & Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, *47*(2), 255–308. https://doi.org/10.1162/coli_a_00402.

Fortunato, M., & Ravani, S. (2015). L'informatica al servizio della filologia e della linguistica sarda: Il corpus ATLiSOr (archivio testuale della linguasarda delle origini). *Bollettino di studi sardi*, *8*, 53–90.

Gerdes, K. (2013). Collaborative dependency annotation. *Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013)*, *88–97*. https://aclanthology.org/W13-3711/.

Guibon, G. (2020). When Collaborative Treebank Curation Meets Graph Grammars. Arborator With a Grew Back-End. *Proceedings of the 12th Conference on Language Resources and Evaluation*, 5291–5300. https://aclanthology.org/2020.lrec-1.651/.

Hajek, S., & Goebl, H. (2021). Le strutture profonde del dominio linguistico sardo: Un'analisi dialettometrica. In *Actes du XXIXe Congrès international de linguistique et de philologie romanes, Copenhague 1er–6 juillet 2019. Section 7. Dialettologia e geolinguistica medievale e moderna (Europa e fuori dall'Europa)* (Vol. 2, pp. 979–991). Société de linguistique romane.

Loporcaro, M., & Putzu, I. (2024). Sardinian. In *Oxford Encyclopedia of Romance Linguistics* (pp. 1–42). Oxford University Press. https://dx.doi.org/10.1093/acrefore/9780199384655.013.725.

Mura, P., Carta, S., Giuliani, A., & Manca, M. (2023). *The corpus of Sardinian emigrants:a tool for a quantitative approach to contact phenomena,*. MiLES: Minority Languages in European Societies - International Conference-Turin, Torino/Bard.

Peng, Q., Qi, P., Zhang, Yuhao, Zhang, Yuhui, Bolton, J., & Manning, C. (s.d.). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. *ACL2020 System Demonstration*. https://doi.org/10.18653/v1/2020.acl-demos.14.

Puddu, N., & Stein, A. (2018). Word-level and higher level annotation of the Sardinian Medieval Corpus. *Proceedings of the Second Workshop on Corpus-Based Research in the Humanities*.

Puddu, N., & Talamo, L. (2020). EModSar: A Corpus of Early Modern Sardinian Texts. In *Atti del IX Convegno Annuale dell'Associazione per l'Informatica Umanistica e la Cultura Digitale (AIUCD). La svolta inevitabile: Sfide e prospettive per l'Informatica Umanistica* (pp. 210–215). Associazione per l'Informatica Umanistica e la Cultura Digitale. https://doi.org/10.6092/unibo/amsacta/6316. https://doi.org/10.6092/unibo%2Famsacta%2F6316.

Regione autonoma della Sardegna. (2006). *Limba sarda comuna. Norme linguistiche di riferimento a carattere sperimentaleper la lingua scritta dell'Amministrazione regionale*. https://www.regione.sardegna.it/documenti/1_72_20060418160308.pdf.

Regione autonoma della Sardegna. (2022). *Certificazione provvisoria sperimentale della conoscenza delle lingue di minoranza storica parlate in Sardegna. Criteri ortografici orientativi per la lingua sarda*. https://www.regione.sardegna.it/documenti/1_38_20220721131659.pdf.

Soria, C., Quochi, V., & Russo, I. (2018). The DLDP Survey on Digital Use and Usability of EU Regional and Minority Languages. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*. LREC 2018. https://aclanthology.org/L18-1656/.

Virdis, M. (1988). Sardisch: Areallinguistik. In G. Holtus, M. Metzeltin, & C. Schmitt (A c. Di), *Lexikon der Romanistischen Linguistik: IV* (pp. 897–913). Niemeyer.

Zeman, D., Hajič, J., Straka, M., Nguyen, D. Q., Potthast, M., Stepanov, E., Di Fabbrizio, G., Promrit, N., & Popel, M. (2018). CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 1–21. https://doi.org/10.18653/v1/K18-2001.

A Learning-by-Doing Approach to Spoken Data Collection: The Case of the LPSP Corpus

CLAUDIA ROBERTA COMBEI[1], GAIA ELEONORA DI RAIMONDO[2], SOFIA MAESTRI[3], BENEDETTA ROMANAZZI[4], ELENA SCOTTI[5]

[1]UNIVERSITÀ DEGLI STUDI DI ROMA "TOR VERGATA", ITALY – CLAUDIA.ROBERTA.COMBEI@UNIROMA2.IT
[2]UNIVERSITÀ DI PAVIA, ITALY - GAIAELEONORA.DIRAIMONDO01@UNIVERSITADIPAVIA.IT
[3]UNIVERSITÀ DI PAVIA, ITALY - SOFIA.MAESTRI01@UNIVERSITADIPAVIA.IT
[4]UNIVERSITÀ DI PAVIA, ITALY - BENEDETTA.ROMANAZZI01@UNIVERSITADIPAVIA.IT
[5]UNIVERSITÀ DI PAVIA, ITALY - ELENA.SCOTTI01@UNIVERSITADIPAVIA.IT

ABSTRACT (ENGLISH)

The paper presents an experiential learning model implemented in a 36-hour, 7-day intensive MA course in Linguistics at Collegio Ghislieri in Pavia. The course, named "Laboratory Phonetics and Speech Processing", produced the LPSP corpus, comprising two subcorpora: LPSP-IT (L1 Italian) and LPSP-ENG (L2 English). Under instructor supervision, students completed the full speech corpus construction cycle: task design, drafting and administering of informed consent and questionnaire, booth recording, speech segmentation and cleaning, data curation, documentation, licensing, and archiving. Ethics was embedded at all stages (consent in plain language, attention to sociodemographic data included, anonymization, controlled sharing). Recordings were collected in January 2026 in a phonetics laboratory, using a Blue Yeti Pro microphone at a sampling frequency of 44.1 kHz. Each of the 18 speakers completed a reading task (texts created for phoneme coverage; one speaker at a time; mono; cardioid mode; two attempts) and a find-the-difference spontaneous dialogue (paired speakers; stereo; bidirectional mode). Cleaning and segmentation resulted in 90 .wav files (total: 2 hours and 47 minutes), 45 for LPSP ENG and 45 for LPSP IT. Speakers were aged 22-27 years (M = 24.44, SD = 1.50). All were native speakers of Italian and comprised 61.11% women, 33.33% men, and 5.55% not disclosing their gender. Most participants resided in Northern Italy and reported a high level of education. This experience illustrates how attention to the technical dimensions of spoken-data collection, combined with reproducibility and ethical reflection, can transform students into competent data curators for their theses and early-stage linguistic research. The LPSP Corpus (audio recordings, stimuli, and sociodemographic information) is deposited on OSF under a CC BY-NC-SA 4.0 license and is available upon request, enabling reuse and replication of both the speech resource and the associated teaching model.
Keywords: spoken data; corpus collection; speech corpus; experiential learning

ABSTRACT (ITALIANO)

Un approccio di apprendimento esperienziale alla raccolta dei dati di parlato: il caso del corpus LPSP.
Il lavoro presenta un modello di apprendimento esperienziale implementato in un corso magistrale di linguistica (7 giorni, intensivo, 36 ore) tenuto al Collegio Ghislieri di Pavia. Nel corso, intitolato "Laboratory Phonetics and Speech Processing", è stato creato il corpus di parlato LPSP, composto da due sottocorpora: LPSP-IT (italiano L1) e LPSP-ENG (inglese L2). Sotto la supervisione della docente, i/le partecipanti hanno completato l'intero ciclo di costruzione del corpus: progettazione e somministrazione di consensi informati e questionari, registrazione in cabina, pulizia e segmentazione, documentazione, licenza, gestione e archiviazione dati. Gli aspetti etici sono stati integrati in tutte le fasi (consenso informato chiaro, attenzione ai dati sociodemografici, anonimizzazione, condivisione controllata). Le registrazioni sono state fatte a gennaio 2026 in un laboratorio di fonetica, con un microfono Blue Yeti Pro (frequenza di campionamento: 44,1 kHz). I 18 parlanti hanno svolto un compito di lettura (testi che rispettano esigenze fonologiche; mono; modalità cardioide; due tentativi) e un dialogo spontaneo di tipo trova le differenze (stereo; modalità bidirezionale). Le operazioni di pulizia e segmentazione hanno prodotto 90 file .wav (2 ore e 47 minuti), 45 per LPSP-IT e 45 per LPSP-ENG. I 18 parlanti madrelingua italiano avevano un'età compresa tra 22 e 27 anni (M = 24,44, DS = 1,50), 61,11% erano donne, 33,33% uomini e 5,55% di genere non dichiarato. La maggioranza risiedeva nel Nord Italia e aveva un livello alto di istruzione. Questa esperienza didattica mostra come l'attenzione all'etica, alla riproducibilità e agli aspetti tecnici e teorici della raccolta dei corpora orali possa trasformare gli/le studenti/esse in curatori/trici competenti di dati linguistici per tesi e altre ricerche. Il corpus LPSP è depositato su OSF con

licenza CC BY-NC-SA 4.0 ed è disponibile su richiesta per consentire il riuso dei dati e la riproducibilità del modello didattico.
Parole chiave: dati di parlato; raccolta di corpora; corpus di parlato; apprendimento esperienziale

1. INTRODUCTION[1]

This paper describes a teaching and research initiative carried out within the 36-hour Master's degree in Linguistics course "Laboratory Phonetics and Speech Processing" at Collegio Ghislieri in Pavia. The course was designed to bridge theory and practice in experimental phonetics, corpus linguistics, and speech technology, by guiding students through the complete pipeline of speech science research, from task design and ethical approvals to recording, annotation, transcription, cleaning, documentation, analysis, and research dissemination (including the draft of this paper). As a course capstone outcome, the students collaboratively created the LPSP corpus (the name derives from the initials of "Laboratory Phonetics and Speech Processing"), a spoken resource comprising two subcorpora: LPSP-IT for L1 Italian and LPSP-ENG for L2 English, both containing data uttered by native speakers of Italian. The corpus can be reused for future research (including their own dissertations) and teaching. The project provided a structured experiential learning environment in which students assumed rotating, well-defined roles (e.g., task design, fieldwork, audio engineering, data management), thereby acquiring both methodological rigor and good data stewardship.

The rationale for this course is twofold. From a teaching and learning perspective, the instructor began from the premise (supported by their own experience) that linguistics students often work with pre-existing corpora and other linguistic resources without ever engaging directly with the processes of data elicitation, processing, documentation, and curation (Ma et al., 2023). This lack of exposure can mask issues of data provenance, representativeness, ethical considerations, measurement validity, and reproducibility, which typically surface only later during their Master's degree theses, at a stage when meaningful correction is often no longer feasible. By assuming responsibility for each stage of corpus creation, students confronted the practical challenges inherent in real-world spoken data collection: designing elicitation tasks that simultaneously support segmental and suprasegmental analyses, standardizing recording conditions, and ensuring that file naming, version control, and documentation aligned with typical linguistic research requirements (Wieczorkowska, 2025).

From a research-oriented perspective, the project provided experience with professional tools and techniques (e.g., Praat, forced alignment pipelines) and speech processing workflows (e.g., segmentation, feature extraction, signal inspection), while foregrounding transparent documentation practices to facilitate accurate interpretation by other potential corpus users. The combination of these two dimensions allowed students to gain both practical and technical insight into the methods that underpin research in speech science (Mauranen, 2004).

The course, and subsequently the corpus construction, integrated ethical aspects across all stages. Rather than treating ethics as a one-off compliance exercise, students engaged with it iteratively by drafting and refining informed consent forms in plain language, designing sociodemographic questionnaires that are informative yet proportionate (neither overly long nor unduly intrusive), implementing anonymization for each subcorpus, and discussing potential risks related to identifiability, data sharing, and re-use. Ethical choices informed methodological ones, for example, deciding which sociodemographic variables to collect (and which to omit), how to justify them in relation to possible research questions, and how to store, license, and distribute data responsibly (Cheng et al., 2024). This ethical reflection at all steps aimed to improve the students' understanding of privacy, diversity, potential bias, and the scope of permissible reuse, all aspects that they can carry into their own future research (e.g., their Master's dissertations).

Therefore, LPSP corpus aimed to function as both an instructional scaffold and a research contribution. The purpose of the activity was to help students become competent linguistic data curators (not only users) while producing a reusable resource for the research community useful for investigating L1 Italian and L2 English.

[1] This paper is the result of a collaborative effort by all the authors, who share responsibility for its content. For the purposes of the Italian academic system, we specify the contributions to the first draft, as follows: Sections 1 and 5 were written by C.R. Combei, Section 2 by B. Romanazzi and C.R. Combei, Section 3 by G.E. Di Raimondo and E. Scotti, and Section 4 by S. Maestri. The data visualization, revision, editing, and preparation of the final published version of the paper were carried out by C.R. Combei.

The remainder of the paper details the corpus design and data collection protocol (Section 2), summarizes corpus composition and participant profiles (Section 3), and describes data availability, licensing, and intended uses (Section 4), with the aim of allowing replication of both the corpus and the teaching model. Concluding remarks and future steps are briefly outlined at the end of the paper (Section 5).

2. CORPUS DESIGN AND DATA COLLECTION PROTOCOL

As part of the coursework, students designed and compiled a spoken corpus, namely the LPSP corpus. This resource comprises two subcorpora: an L1-Italian subcorpus (LPSP-IT) and an L2-English subcorpus (LPSP-ENG). The students were divided into two groups, each dedicated to one subcorpus. Within each group, students were assigned specific responsibilities, including elicitation task design, drafting informed consent materials, developing and administering the sociolinguistic questionnaire, managing sociolinguistic data, operating the recording workflow, and cleaning and segmenting the audio. For the LPSP corpus 18 native speakers of Italian were recruited (see Section 3 for sociodemographic details). All 18 participants received two informed consent forms (one per subcorpus), detailing the aims of the project, data collection and storage procedures, participant rights, and data sharing plans (speakers were informed that the corpus would have been shared through the Open Science Framework platform). By signing the informed consent, speakers consented to being recorded and to anonymized data sharing with the research community. Speakers also completed one sociolinguistic questionnaire for each subcorpus to provide background information relevant for analysis and interpretation of the audio material. An anonymization protocol was implemented for both subcorpora: each speaker was assigned two distinct alphanumeric identifiers (one for L1 Italian, one for L2 English), and these identifiers were used to label recordings and link them to sociolinguistic variables. Names were not associated with the audio or the sociolinguistic datasets. Access to the re-identification key (the consent forms and named questionnaires) was restricted to one designated student and to the course instructor (the first author of this paper).

Elicitation took place in January 2026 in the recording booth of the experimental phonetics laboratory at the University of Pavia. Recordings were made with a Blue Yeti Pro microphone using Praat (Boersma, 2001) at a 44.1 kHz sampling frequency. A designated student oversaw the recording process, ensuring consistent file naming with the assigned alphanumeric codes and managing technical parameters (e.g., mono vs. stereo channel; cardioid mode vs. bidirectional polar mode). Data collection occurred over two sessions: the first afternoon was dedicated to the L1-Italian subcorpus and the second to the L2-English subcorpus. Each participant completed four recording sessions in total (two per subcorpus), corresponding to two task types: a reading task (with two attempts) and a find-the-difference task. The choice of the two tasks was dictated by previous research that demonstrated how spontaneous speech differs acoustically from read speech, especially through reduced spectral space, which significantly lowers phoneme and speech recognition accuracy (Nakamura et al., 2008).

For the reading task, speakers read aloud texts specifically constructed to cover the full phoneme inventory of the target language (Italian and British English). Each text was read twice to increase control and reliability, enabling future analyses of the realization of Italian sounds by native Italian speakers and the realization of British English sounds by Italian learners of English. The Italian text had a type–token ratio of 0.77, consisted of 8 sentences, and the mean sentence length was 15.9 words. The lexical items spanned a wide range of frequency per million, which was checked in the CORIS corpus (Rossini Favretti et al., 2002), yielding a balanced complexity profile. The English text had a type-token ratio of 0.65, contained 25 sentences and had a mean sentence length of 8.4 words. Similarly to the Italian text, lexical items for the English text were selected to cover a broad frequency range, which was checked in the British National Corpus (Burnard & Aston, 1998).

The paired task elicited semi-spontaneous dialogue using two closely related images that contained systematic similarities and differences. The English materials were designed around familiar lexical items to ensure accessibility for speakers with lower English proficiency. Paired speakers were given different versions of the image and asked to describe what they saw and identify differences, thereby prompting spontaneous conversation anchored in the visual stimuli.

Both texts and images were created by the students in charge of stimulus design. All stimuli and accompanying materials are available via the Open Science Framework (OSF) repository for the LPSP Corpus: https://osf.io/3vu7e/.

3. CORPUS DESCRIPTION AND STATISTICS

The complete corpus comprises 90 audio files, for a total of 2 hours 47 minutes and 48 seconds. In particular, 45 audio files are part of the LPSP-ENG subcorpus, which contains English-language audio material (for a total of 103 minutes and 18 seconds), while the other 45 audio files regard the LPSP-IT subcorpus with its Italian-language material (for a total of 64 minutes and 30 seconds).
The original recordings of the find-the-differences task and the ones of the reading task of each subcorpus were cleaned by background noises and silence using the freeware software Praat (Boersma, 2001). Specifically, the files of the reading task of both subcorpora were segmented in two different files, as participants made two attempts while recording. Therefore, the initial number of 18 files turned into 36 for the reading task and 9 for the find-the-difference task, for a total of 45 recordings in each subcorpus.
The 18 participants that took part in the corpus recording were all native Italian speakers (L1) aged from 22 to 27 (M = 24.44, S.D.=1.50). The gender distribution in the sample included 61.11% (n = 11) of female speakers, 33.33% (n = 6) of male speakers, and 5.55% (n = 1) who did not prefer to disclose their gender.
Then, 72.22% (n= 13) of the speakers came from Northern Italy, specifically 38.88% (n = 7) from Lombardy, 11.11% (n = 2) from Liguria, 11.11% (n = 2) from Emilia-Romagna, and 11.11% (n = 2) from Piedmont. While the remaining participants, which is 27.77% (n = 5), came from Southern and Central Italy, namely 5.55% (n = 1) from Sicily, 5.55% (n = 1) from Campania, 5.55% (n = 1) from Calabria, 5.55% (n = 1) from Abruzzo, and 5.55% (n = 1) from Umbria. When they were recorded, the 83.33% (n = 15) of speakers resided in Northern Italy, so only 16.66% (n = 3) indicated they resided in the South, in the same region they were born.
Also, at the time of data collection, 55.55% (n = 10), which is more than half of the participants, reported holding a Bachelor's degree as their highest completed level of education, while being enrolled in a Master's degree program in Linguistics. Then, 33.33% (n = 6) were PhD students of Linguistics and reported a Master's degree as their highest completed qualification. A high school diploma was reported by 11.11% (n = 2) of the speakers who were enrolled in a Bachelor's degree program (respectively, one student in Classics and one in Communication and Media Studies) at the time of the recording. The high level of education of the sample reflects the academic recruitment context of the study.
Because the recording protocol included a reading task, information on visual impairments and learning disorders was collected (Ehri, 1997; Miller & Yochum, 1991). According to self-reports, 72.22% of the participants (n = 13) used corrective devices (glasses or contact lenses) for refractive errors such as myopia or astigmatism, whereas 27.78% (n = 5) reported no visual impairments. Furthermore, none of the participants (100%, n = 18) reported having any learning disorders.

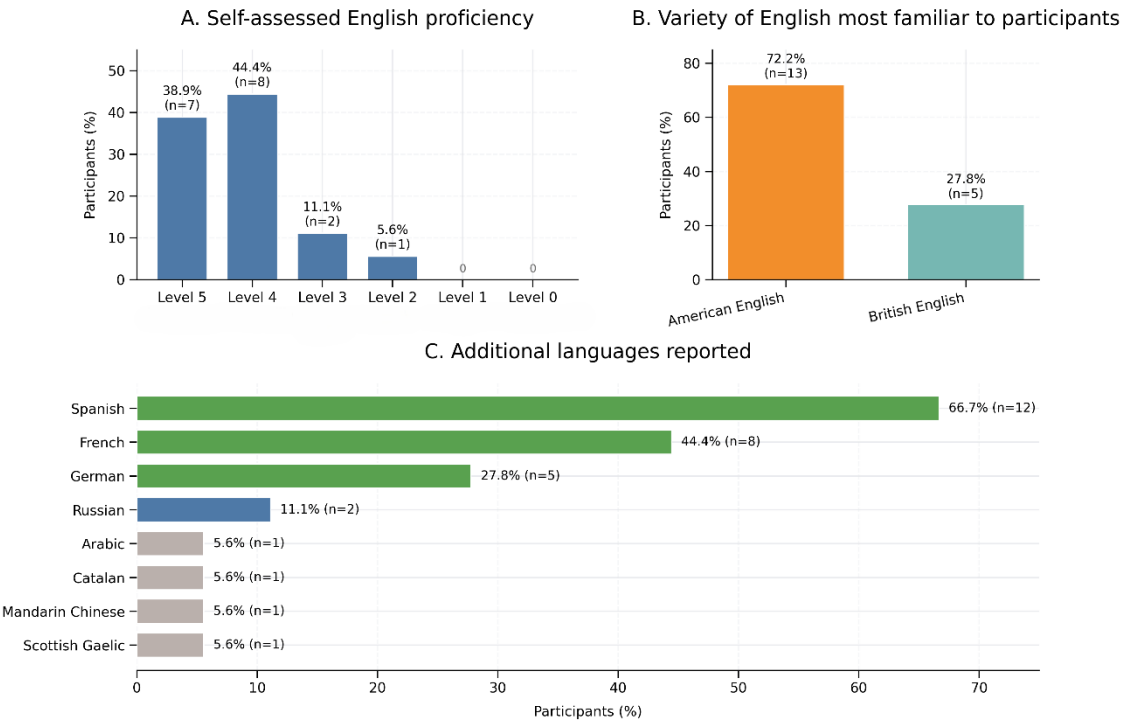


Figure 1. The speakers' L2 linguistic repertoire

The speakers' linguistic repertoire was investigated through a sociolinguistic questionnaire, in which they reported the languages they knew, besides L1 Italian, and self-assessed their proficiency on a 6-point Likert scale (i.e., 0-5). As shown in Figure 1, all participants reported knowledge of English, indeed 38.89% of them (n = 7) reported excellent proficiency (level 5), 44.44% (n = 8) advanced proficiency (level 4), and 11.11% (n = 2) intermediate proficiency (level 3). A further 5.55% (n = 1) indicated low proficiency (level 2). No participants reported very low proficiency (level 1) or no knowledge of English (level 0). In addition, American English was the variety that 72.22% (n = 13) of the participants were most familiar with, compared to British English of which only 27.77% (n = 5) had knowledge of. Moreover, 66.66% (n = 12) of the speakers reported knowledge of Spanish, with varying levels of proficiency. French was reported by 44.44% (n = 8) of the participants and German by 27.77% (n = 5) of them, while only 11.11% (n = 2) reported competence in Russian. The linguistic repertoire also included single instances (5.55%) of Arabic, Mandarin Chinese, Catalan, and Scottish Gaelic.
The sociolinguistic questionnaire also investigated language use across different social domains. In interactions with friends, Italian was reported as the sole language by 66.66% of the participants (n = 12), while 16.66% (n = 3) reported using Italian in combination with English, 5.5% (n = 1) Italian and dialect and 11.11% (n = 2) Italian, English, and dialect. Furthermore, Italian was reported as the only language used at home by 77.77% (n = 14) of speakers, while 16.66% (n = 3) use Italian in combination with dialect. Only 5.5%, which corresponds to one case, uses Italian with Mandarin Chinese and Wu. Regarding interactions with siblings, 77.77% (n = 14) of the participants reported using Italian exclusively, while 11.11% (n = 2) reported using Italian and a dialect; the remaining 11.11% (n = 2) did not respond, probably due to the absence of siblings. Among the 18 participants, 88.88% (n = 16) reported Italian as the L1 of both parents, while 11.11% (n = 2) reported non-Italian parental L1s (Wu in one case, and Italian and Amharic in another). Finally, in interactions with romantic partners, 66.66% (n = 12) of participants reported using Italian exclusively, while 33.33% (n = 6) did not respond to this question, presumably corresponding to participants without a current partner or who did not want to disclose this information.

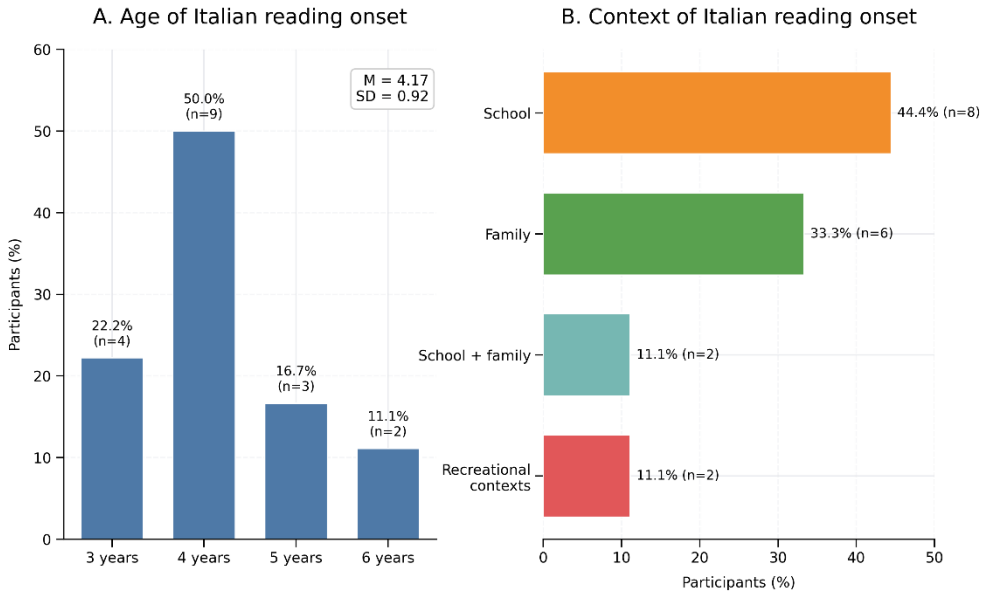


Figure 2. The speakers' L1 Italian reading onset and context

Speakers had to report in the sociolinguistic forms the age and the contexts of Italian reading onset for the LPSP-IT subcorpus (Modiano, 1968; Kamhi-Stein, 2003), whereas for the LPSP - ENG subcorpus the age and contexts of English learning onset and the frequency of spoken English usage (Marian et al., 2007; Unsworth, 2013). Figure 2 presents the data from the sociolinguistic form of the LPSP-IT subcorpus, showing that 50% (n = 9) of participants claimed they had started reading Italian at the age of 4, 22.22% (n = 4) at the age of 3, 16.66% (n = 3) at the age of 5 and only 11.11% (n = 2) at the age of 6 (M = 4.16, SD = 0.92). Regarding the contexts of reading onset, among the total number of participants, 44.44% (n = 8) started to read Italian at school, 33.33% (n = 6) with their family, 11.11% (n = 2) both at school and with their family and 11.11% (n = 2) reported recreational contexts as first onset.

As far as the LPSP-ENG is concerned, as shown in Figure 3, 22.22% (n = 4) started to learn English at the age of 5, 22.22% (n = 4) at the age of 6, 16.66% (n = 3) at the age of 4, 11.11% (n = 2) since birth (although they do not consider themselves native speakers of English), 11.11% (n =2) at the age of 10, 5.55% (n = 1) at the age of 2, 5.55% (n = 1) at the age of 7 and 5.55% (n = 1) at the age of 15 (M = 5.55, SD = 3.55). On the other hand, among 18 participants, 83.33% (n= 15) had learned English at school, 11.11% (n = 2) in a family context and 5.55% (n = 1) in an informal context. In addition, 33.33% (n = 6) of the participants use English a few times in a month, 27,77% (n = 5) every day, 22,22% (n = 4) a few times a week, 11.11% (n = 2) once a week and the 5.55% (n = 1) less than once a month.

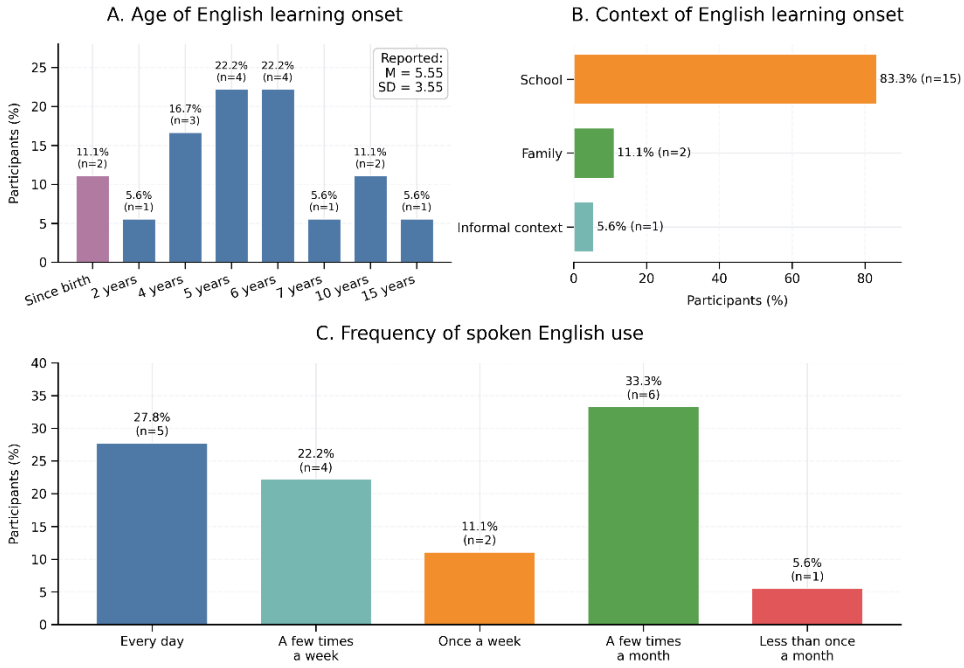


Figure 3. The speakers' L2 English learning onset and context

4. DATA AVAILABILITY AND USABILITY

All corpus anonymized data were uploaded to the OSF repository as a private project and are available for teaching and research purposes, upon reasonable request, by registered users of the platform. The repository was assigned the DOI 10.17605/OSF.IO/3VU7E and is available at the following OSF link: https://osf.io/3vu7e/. The project contains two main folders, LPSP-ENG and LPSP-IT, both of which include the audio files for the read speech task and for the spontaneous speech task (find-the-difference) in .wav format. The name of each audio name consists of an alphanumeric string which is the result of the following information:

- Alphanumeric ID code of the speaker.
- RS (read speech) or FD (find-the-difference), depending on the task performed by the recorded speaker.
- Only for RS, the number of the reading attempt (trial 1 or trial 2).

Folders also include all stimuli in .pdf format and sociodemographic data from the questionnaires in .xlsx format, which report information about the speakers.
All available data are released under the CC BY-NC-SA 4.0 International License and may be downloaded and used for non-commercial research and teaching purposes only. Potential applications include studies of linguistic variation and other sociophonetic aspects, teaching activities in speech processing (e.g., audio segmentation, acoustic analysis, feature extraction), L2 acquisition research, conversation analysis, and other related fields. Use of the data for commercial purposes is strictly prohibited. Any ethical violations will be reported to the relevant authorities and addressed in accordance with applicable regulations.

5. CONCLUSIONS AND FUTURE WORK

In this work we showed how a structured experiential learning approach can train MA students in Linguistics to manage the full speech corpora lifecycle, from the design of elicitation tasks to sharing the data and ethical considerations. The course deliverable, namely the LPSP corpus (with its L1 Italian and L2 English subcorpora), transformed the coursework into a reusable resource for the linguistic community, for teaching and research across phonetics, speech processing, and other related fields. Limitations include the modest sample size of the corpus (18 speakers), single-institution setting, constrained time frame, and a task design that prioritized control over ecological data collection. Future research will aim to broaden the diversity of participants in terms of sociodemographic characteristics, expand the range of spontaneous speech contexts, and incorporate full transcription of all audio materials

alongside multilayer phonetic and phonological annotation, work that is currently being carried out by MA students in Computational Linguistics at the University of Rome "Tor Vergata". All in all, when considered together, we believe that this activity encouraged students to move from being mere users of corpora toward a more reflective and competent engagement with linguistic data collection and curation.

ACKNOWLEDGEMENTS

The first author is thankful to Collegio Ghislieri for hosting the intensive course "Laboratory Phonetics and Speech Processing" and to Chiara Zanchi for promoting it. In addition to the authors of this paper, the following students, listed in alphabetical order, contributed to the collection and processing of the corpus: Matilde Bazzurro, Peter Bellarosa, Anna Giuliana Carlini, Emanuele Maria Ferrero, Francesco Logozzo, Elisa Mattiolo, Assunta Napoletano, Filippo Zana. The authors gratefully acknowledge their efforts, as well as the speakers who generously agreed to be recorded.

REFERENCES

Boersma, P. (2001). Praat, a system for doing phonetics by computer. Glot International, 5(9–10), 341–345.

Burnard, L., & Aston, G. (1998). The BNC handbook: Exploring the British National Corpus. Edinburgh University Press.

Cheng, L., Han, J. & Nasirov, J. (2024). Ethical considerations related to personal data collection and reuse: trust and transparency in language and speech technologies. International Journal of Legal Discourse, 9(2), 217-235. https://doi.org/10.1515/ijld-2024-2010.

Ehri, L. C. (1997). Sight word learning in normal readers and dyslexics. In B. A. Blachman (Ed.), Foundations of reading acquisition and dyslexia: Implications for early intervention (pp. 163–189). Routledge.

Kamhi-Stein, L. D. (2003). Reading in two languages: How attitudes toward home language and beliefs about reading affect the behaviors of "underprepared" L2 college readers. TESOL Quarterly, 37(1), 35–71. https://doi.org/10.2307/3588465.

Ma, Q., Chiu, M. M., Lin, S., & Mendoza, N. B. (2023). Teachers' perceived corpus literacy and their intention to integrate corpora into classroom teaching: A survey study. ReCALL, 35(1), 19–39. https://doi.org/10.1017/S0958344022000180.

Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The language experience and proficiency questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. Journal of Speech, Language, and Hearing Research, 50(4), 940–967. https://doi.org/10.1044/1092-4388(2007/067).

Mauranen, A. (2004). Speech corpora in the classroom. In G. Aston, S. Bernardini & D. Stewart (Ed.), Corpora and Language Learners (pp. 195-211). John Benjamins Publishing Company. https://doi.org/10.1075/scl.17.14mau.

Miller, S. D., & Yochum, N. (1991). Asking students about the nature of their reading difficulties. Journal of Reading Behavior, 23(4), 465–485. https://doi.org/10.1080/10862969109547754.

Modiano, N. (1968). National or mother language in beginning reading: A comparative study. Research in the Teaching of English, 2(1), 32–43. https://doi.org/10.58680/rte196820263.

Nakamura, M., Iwano, K., & Furui, S. (2008). Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance. Computer Speech & Language, 22(2), 171–184. https://doi.org/10.1016/j.csl.2007.07.003.

Rossini Favretti, R., Tamburini, F., & De Santis, C. (2002). CORIS/CODIS: A corpus of written Italian based on a defined and a dynamic model. In A. Wilson, P. Rayson, & T. McEnery (Eds.), A rainbow of corpora: Corpus linguistics and the languages of the world (pp. 27–38). Lincom Europa.

Unsworth, S. (2013). Assessing Age of Onset Effects in (Early) Child L2 Acquisition. Language Acquisition, 20(2), 74–92. https://doi.org/10.1080/10489223.2013.766739.

Wieczorkowska, A. (2025). Methodology for Obtaining High-Quality Speech Corpora. Applied Sciences, 15(4), 1848. https://doi.org/10.3390/app15041848.

Digital Humanities e Letteratura Elettronica in Italia: marginalità e convergenze

GIUSEPPE ARENA[1]
[1]UNIVERSITÀ DI CATANIA, ITALIA – ARENAGIUSEPPE137@GMAIL.COM

ABSTRACT (ITALIANO)

Nel dibattito internazionale sulle Digital Humanities, la letteratura elettronica è spesso considerata un ambito creativo di sperimentazione teorica e metodologica, mentre nel contesto italiano essa occupa una posizione ancora marginale all'interno delle pratiche e delle narrazioni disciplinari. Questo contributo analizza tale scarto, interrogando le condizioni storiche ed epistemologiche che hanno inciso sulla scarsa integrazione della e-lit nelle Digital Humanities in Italia. L'analisi si fonda su una ricognizione critica della letteratura scientifica internazionale, su riferimenti alla tradizione dell'informatica umanistica italiana e su alcuni casi emblematici di riflessione teorica sul testo digitale, che dimostrano possibili convergenze tra le due discipline.
Parole chiave: letteratura elettronica; e-lit italiana; digital humanities; cultura digitale; born-digital

ABSTRACT (ENGLISH)

Digital Humanities and Electronic Literature in Italy: marginality and convergence
In the international debate on Digital Humanities, electronic literature is often considered a creative field of theoretical and methodological experimentation, while in Italy it still occupies a marginal position within disciplinary practices and narratives. This contribution analyzes this gap, questioning the historical and epistemological conditions that have affected the poor integration of e-lit in Digital Humanities in Italy. The analysis is based on a critical review of international scientific literature, references to the tradition of Italian humanistic informatics, and some emblematic cases of theoretical reflection on digital text, which demonstrate possible convergences between the two disciplines.
Keywords: electronic literature; Italian e-lit; digital humanities; digital culture; born-digital

1. INTRODUZIONE

Quando nel 2004 Susan Schreibman, Ray Siemens e John Unsworth pubblicarono *A Companion to Digital Humanities*, l'obiettivo dichiarato era quello di riunire teorici e pionieri provenienti sia dalle scienze umane sia dall'informatica per discutere le Digital Humanities come possibile campo di studi autonomo. In quell'occasione, le DH venivano descritte come un ambito plurale e dinamico, caratterizzato dalla condivisione di presupposti teorici e metodologici comuni, applicabili trasversalmente ai diversi settori delle scienze umane, indipendentemente dall'uso diretto o meno del calcolatore . Questa impostazione rifletteva una concezione ancora fortemente legata alla tradizione delle *Humanities Computing*, intese, secondo la formulazione di Roberto Busa, come l'automazione di ogni possibile analisi dell'espressione umana, con il testo scritto quale nucleo privilegiato dell'indagine . Poco più di un decennio dopo, *A New Companion to Digital Humanities* segnala un cambiamento significativo di prospettiva. Pur continuando a evitare una definizione disciplinare in senso stretto, l'orizzonte si amplia sensibilmente, includendo non solo la modellazione e l'analisi computazionale dell'informatica umanistica, ma anche lo studio culturale delle tecnologie digitali, delle loro possibilità creative e del loro impatto sociale . È in questo contesto che settori fino ad allora marginali, come i videogiochi e la letteratura elettronica, iniziano a essere riconosciuti come aree di interesse legittime per le Digital Humanities.
Nel contesto italiano, al contrario, il rapporto tra DH e letteratura elettronica (detta anche e-literature o e-lit) appare tuttora fragile e disomogeneo. Nonostante l'esistenza di una storia nazionale di pratiche riconducibili alla e-lit e di riflessioni teoriche affini a quelle sviluppate in ambito DH, la letteratura elettronica risulta sistematicamente marginale nelle istituzioni accademiche, nelle narrazioni storiografiche del campo e nelle pratiche di ricerca formalmente riconosciute. Tale marginalità non si traduce, tuttavia, in una semplice assenza: al contrario, essa convive con una serie di convergenze teoriche e metodologiche, spesso non tematizzate come tali.

L'obiettivo di questo contributo è dunque duplice: da un lato, analizzare le ragioni storiche, epistemologiche e istituzionali di tale marginalità, con un focus sulle DH italiane; dall'altro, individuare i punti di convergenza teorica e metodologica tra i due ambiti, mostrando come questo dialogo possa contribuire a ridefinire i confini e le pratiche del campo.

2. LETTERATURA ELETTRONICA COME PRATICA DELLE DH

Nella sezione Creation di A New Companion to Digital Humanities emergono due contributi emblematici di questa apertura internazionale: New Media and Modeling: Games and the Digital Humanities di Steven E. Jones ed Electronic Literature as Digital Humanities di Scott Rettberg. Entrambi affrontano ambiti tradizionalmente percepiti come estranei alla ricerca umanistica, spesso relegati all'intrattenimento o alla sperimentazione artistica, mostrando invece la loro rilevanza metodologica e teorica per le DH. Jones sottolinea, infatti:

> Because games are algorithmic systems that test the player's freedom within programmed formal constraints, and because they combine computing with modes of cultural expression associated with the humanities storytelling, design, aesthetics, social communication – they would seem to be of obvious interest to the digital humanities. (Jones, 2016: 84)

In questa prospettiva, il gameplay può essere interpretato come un processo di modellazione iterativa e ricorsiva, basato su cicli di prova, errore e riformulazione, analogo alle pratiche sperimentali tipiche dei progetti DH (Jones, 2016: 91). A ciò si aggiunge l'attenzione alla piattaforma, intesa come intreccio di hardware, software e contesto socio-culturale, e al ruolo dei videogiochi come casi di studio per la conservazione e curatela di oggetti *born-digital* (Jones, 2016: 89).[1]
Un'argomentazione analoga, ma declinata sul versante strettamente letterario, è proposta da Scott Rettberg, che interpreta la letteratura elettronica come una pratica pienamente interna alle Digital Humanities. Più che come applicazione di strumenti digitali a forme letterarie tradizionali, la e-lit viene intesa come sperimentazione di nuove forme espressive native dell'ambiente digitale. Afferma l'autore: «The creative production of this type of work can itself be understood as a digital humanities practice: not an application of digital tools to a traditional form of humanities research, but rather experiments in the creation of new forms native to the digital environment.» (Rettberg, 2016: 127). .
La pratica creativa diventa così essa stessa una modalità di ricerca DH: le opere di letteratura elettronica sono al contempo artefatti estetici, esperimenti tecnologici documentati e occasioni di riflessione teorica sulla testualità digitale (Rettberg, 2016: 128). La letteratura elettronica comprende, infatti, «opere nate digitali, che non potrebbero vivere al di fuori del medium informatico, poiché ne sfruttano alcune proprietà, utilizzandole come strumenti formali, retorici ed estetici» (Iadevaia, 2021: 7). Secondo la definizione dell'Electronic Literature Organization,[2] si tratta, infatti, di «work with an important literary aspect that takes advantage of the capabilities and contexts provided by the stand-alone or networked computer» (Hayles, 2007). Sono esclusi dal novero tutti quei testi pensati per il supporto cartaceo e poi digitalizzati in un secondo momento, nei quali il passaggio da un'interfaccia all'altra non determina condizioni che rendono l'opera non fruibile o mutata in qualche sua parte e carattere (De Vivo, 2011: 213–214). Il rapporto tra la componente umana e creativa e il testo, che nella tradizione tipografica coincide con il supporto, si arricchisce di un terzo polo, quello del medium tecnologico che determina potenzialità espressive nuove e non conformi alla tradizionale concezione lineare dell'opera e al sistema espressivo testo-centrico. Le forme espressive della letteratura elettronica sono varie (la loro resa finale

[1] Vale la pena sottolineare che il riconoscimento dei videogiochi come oggetto di interesse per le Digital Humanities precede di molto il New Companion del 2016. Già negli anni Novanta la narratologia aveva discusso il rapporto tra gioco e narrazione, dando origine al cosiddetto dibattito ludologia vs. narratologia. Come ricostruisce Aarseth, il termine 'ludologia' fu introdotto da Gonzalo Frasca nel 1999 come proposta metodologica complementare alla narratologia per studiare le strutture specifiche del gioco, e la contrapposizione tra ludologi e narratologi fu in larga misura una costruzione esterna, non una reale frattura teorica (Aarseth, 2023). Sul versante della narratologia computazionale, la voce dedicata del Living Handbook of Narratology (Mani, 2012) definisce questo campo come lo studio della narrativa dal punto di vista del calcolo e dell'elaborazione dell'informazione, il cui ambito include esplicitamente i videogiochi e le opere d'arte computazionali, oltre alla generazione e interpretazione automatica di storie. I videogiochi erano dunque già stabilmente nell'orizzonte delle discipline umanistiche computazionali ben prima del New Companion.
[2] L'ELO (Electronic Literature Organization), nata nel 1999, svolge un ruolo centrale nella canonizzazione e nella documentazione delle opere di e-lit internazionali, attraverso conferenze annuali, collezioni curate e risorse educative. https://eliterature.org/ (cons. 08/02/2026)

dipende dal contesto tecnico d'esecuzione) e mutano con il progresso del medium tecnologico. A differenza delle semplici trasposizioni digitali di testi a stampa, tali prodotti culturali integrano, quindi, interattività, multimedialità, non linearità, mettendo in discussione categorie consolidate come testo, opera e autore (Venerandi, 2024: 22–24). Per tali motivi, i generi della letteratura elettronica sono vari e di difficile definizione; se ne ricordano alcuni: l'Hypertext Literature, la Collaborative Literature, l'Interactive Literature, il MUD, la Generative Literature, la Code Literature, la Web Literature e i videogiochi che, negli ultimi decenni, sono riusciti ad imporsi come mezzo capace di grandi narrazioni, sfruttando l'elemento interattivo e ludico, e che, per tali ragioni, potrebbero essere considerati forme di letteratura elettronica commerciale, vista la grande diffusione presso il pubblico.[3]

3. LETTERATURA ELETTRONICA ITALIANA: GENEALOGIA E RICEZIONE

In Italia la letteratura elettronica affonda le proprie radici nelle neoavanguardie del Novecento.[4] Già Nanni Balestrini, con le sue sperimentazioni combinatorie, è indicato come padre dell'e-lit italiana: la poesia *Tape Mark I* (1961) e il romanzo combinatorio *Tristano* (1966) vengono citati come punti di partenza della storia nazionale della letteratura elettronica.[5] A partire dagli anni Ottanta si sviluppa, inoltre, il filone delle avventure testuali o interactive fiction, la cui paternità italiana è attribuita a Enrico Colombini con la sua Avventura nel Castello, del 1982. Tra i contemporanei si citi, tra tutti, Fabrizio Venerandi che ha pubblicato, nel 2016, Poesie Elettroniche, una raccolta di poesie in formato EPUB3, in cui i versi cambiano nel tempo e reagiscono all'input del lettore (Venerandi, 2016).[6]

Le piattaforme digitali svolgono un ruolo strutturale per la visibilità e lo studio della e-lit. Tra queste, *The Electronic Literature Review*,[7] attivo dal 2012, curato da Daniele Giampà che raccoglie interviste e bibliografia aggiornata; *LEI – Letteratura Elettronica Italia*,[8] la prima comunità italiana dedicata a promozione, ricerca e preservazione del settore. Un'azione formativa è stata svolta dai workshop *Letteratronica* (Argo & Nie Wiem, 2020-2023), laboratori condotti da Fabrizio Venerandi sulla creazione di ipertesti, avventure testuali e videogiochi narrativi.

Le incursioni in ambito accademico sono, invece, testimoniate da convegni dedicati, seppur poco in dialogo con le Digital Humanities. Tra questi: *OLE – Officina di Letteratura Elettronica* (Napoli 2011, con atti a cura di Masucci e Di Rosario), il primo in assoluto a tenersi in Italia e il successivo *OLE.01 Festival* (2014). Di portata internazionale sono state le conferenze *ELO 2022* a Como, sul tema *Education and Electronic Literature*, con un panel specificamente incentrato sul rapporto tra e-lit e Digital Humanities, e il recente convegno *eLIT – il futuro della letteratura, la letteratura del futuro* (Cagliari, 2025), tenutosi presso l'Università degli Studi di Cagliari, in collaborazione con la rivista *Erbafoglio* e LEI, che ha unito riflessione teorica a performance.

Tra i progetti accademici finanziati, il progetto *Interdisciplinary Italy 1900-2020: interart/intermedia* (AHRC, Royal Holloway University of London), diretto da Giuliana Pieri, Clodagh Brook e Florian Mussgnug, pur non essendo centrato esclusivamente sulla letteratura elettronica in senso stretto, esplora l'ibridazione intermediale nel Novecento e contemporaneità italiana, in un più ampio contesto interartistico.[9]

---

[3] A conferma della rilevanza socioculturale ed economica del medium videoludico nel contesto italiano, i dati IIDEA delineano un settore strutturato: 14 milioni di utenti, con una presenza adulta predominante (84% nella fascia 18-64 anni), una pratica consolidata (68% dei giocatori partecipa almeno settimanalmente) e un fatturato complessivo di 2,4 miliardi di euro, con una crescita del 36% dal 2022, con oltre 200 imprese attive. Questi dati suggeriscono che il videogioco si è affermato come industria creativa stabile e come pratica culturale consolidata, rendendo ancora più significativa la sua quasi totale assenza dal dibattito e-lit e DH italiano. https://iideassociation.com/wp-content/uploads/2025/04/IIDEA_I-Videogiochi-in-Italia-nel-2024.pdf (cons. 19/04/2026)

[4] Sullo stretto rapporto tra Avanguardie e lo sviluppo della letteratura elettronica in Italia, si veda (Patti, 2022).

[5] Per un'ampia disamina delle generazioni di e-lit italiana, si veda (Iadevaia, 2021).

[6] All'interno delle Electronic Literature Collections, in particolare nella Collection Volume 4, figurano solo due opere di autori italiani: una riedizione ridotta delle Poesie Elettroniche di Fabrizio Venerandi e le poesie cinetiche in inglese, Ergon/Logos, di Molleindustria. Le informazioni sui metadati di altre opere italiane sono conservate presso il database ELMCIP – Electronic Literature as a Model of Creativity and Innovation in Practice. https://elmcip.net/ (cons. 08/02/2026)

[7] https://electronicliteraturereview.org/ (cons. 08/02/2026)

[8] https://www.letteratura-elettronica.it/LEI/ (cons. 08/02/2026)

[9] https://gtr.ukri.org/projects?ref=AH%2FM008819%2F2 (cons. 08/02/2026)

4. LE RAGIONI STRUTTURALI DELLA MARGINALITÀ

Nonostante questi strumenti ed eventi, la letteratura elettronica resta marginale nel contesto accademico italiano. In un survey del 2010 sulla pubblicazione e distribuzione della letteratura elettronica in Europa, si segnalava come la disciplina non avesse preso piede in Italia e che la disseminazione delle poche esperienze autoriali riportate fosse relegata agli autori stessi e ai loro siti web (Eskelinen & Di Rosario, 2012: 25–28) .[10]

In apertura agli atti dell'OLE, Giovanna Di Rosario evidenziava proprio l'assenza dell'e-lit nelle università italiane, attribuendola da un lato alla scarsa attenzione verso una letteratura non ancora canonizzata e dall'altro alla limitata diffusione di Internet nel contesto italiano all'inizio degli anni Duemila, oltre alla generale esclusione della cultura digitale dalle facoltà di lettere (Di Rosario, 2011: 14).[11]

Negli stessi atti di convegno, Alexandra Saemmer riconduce la marginalizzazione accademica della letteratura elettronica alla sua associazione con la dimensione ludico-interattiva, interpretata come forma di 'distrazione' in senso pascaliano (Saemmer, 2011: 369). Sebbene le avanguardie abbiano storicamente valorizzato la playfulness in opposizione a un'arte seriosa, nel contesto digitale il rischio è che l'attenzione si sposti dal contenuto al funzionamento tecnico, riducendo l'esperienza a mera esplorazione di regole operative. La questione cruciale diventa allora se un'arte interattiva puramente ludica non finisca per assecondare, anziché mettere in discussione, le convenzioni del sistema culturale dominante (Saemmer, 2011: 370).

In un contributo centrale, Scott Rettberg individua invece una triplice sfida alla legittimazione della letteratura elettronica: di ricezione, cognitiva ed economica (Rettberg, 2009). Innanzitutto, il pubblico di potenziali lettori non conosce l'esistenza di queste opere. Sul piano cognitivo, invece, introduce la figura del *configurative reader*, chiamato a negoziare tra l'*hyper attention*, una modalità diffusa, multitasking e orientata alla superficie, tipica delle culture digitali, e dall'altro la *deep attention*, la tradizionale concentrazione profonda e immersiva della lettura lineare. La letteratura elettronica impone infatti una fase preliminare di esplorazione e interazione del testo elettronico, necessaria per accedere a una successiva lettura interpretativa. Questo doppio vincolo rappresenta insieme una potenzialità evolutiva e una barriera per lettori formati su protocolli di ricezione lineari. Sul piano economico, Rettberg descrive il settore come fondato su un'economia del dono; le opere sono, cioè, per lo più create e pubblicate gratuitamente sul Web. Questo ne favorisce certamente la diffusione, ma rende insostenibile la pratica artistica a lungo termine.

5. DH ITALIANE, AIUCD E ASSENZA DELLA E-LIT

L'affermazione delle Digital Humanities in Italia segue una traiettoria specifica che incide profondamente sulla possibilità di accogliere la letteratura elettronica. Come osserva Dino Buzzetti, l'informatica umanistica italiana si è sviluppata prevalentemente all'interno dell'università, con docenti impegnati in progetti disciplinari specifici, piuttosto che in centri di ricerca dotati di forte supporto tecnico, come avvenuto in ambito anglosassone (Buzzetti, 2023: 15–16). A ciò si aggiunge l'assetto dei settori scientifico-disciplinari, che ha reso difficile il riconoscimento di pratiche ibride in ambito istituzionale. La tradizione italiana si è, quindi, caratterizzata per una forte impronta teorica ed epistemologica, come emerge chiaramente dal contributo di Tito Orlandi: «[…] il rapporto tra informatica e discipline umanistiche si può esprimere nella questione se vi sia un modo "informatico" di vedere (anche) le discipline umanistiche [...] Il modo informatico prevede la formalizzazione dei dati [...] e la formalizzazione delle procedure per analizzarli e valutarli» (Orlandi, 1992: 17). Su ciò si innestano i fondamenti concettuali dell'informatica umanistica: formalizzazione, modellizzazione e codifica come processo semiotico (Ciotti, 2023: 28–29). Tali fondamenti sono stati applicati a discipline umanistiche già storicamente affermate, rispetto alle quali l'Italia si è fatta pioniera a livello internazionale soprattutto grazie al proprio patrimonio culturale e bibliotecario: filologia, linguistica, archeologia, archivistica, storia

[10] Le opere, i collettivi artistici e le risorse web lì riportati sono, ad oggi, per lo più inaccessibili o inesistenti. Per approfondire si veda pure la ricognizione del 2016 del GRUPPO GIADA (Gruppo Giada, 2016).

[11] Dal 2011 la situazione italiana è sicuramente migliore. Secondo i dati dell'Italy 2025 Digital Decade Country Report, redatto dalla Commissione Europea, da un punto di vista infrastrutturale, la copertura FTTP (fibre-to-the-premise) è al 70,7%, in crescita, ma ancora non universale. Meno incoraggianti altri dati: solo il 45,8% della popolazione possiede almeno competenze digitali di base (contro un obiettivo UE 2030 dell'80,1%), con lacune riguardanti persone con bassa scolarizzazione, ma anche giovani. https://digital-strategy.ec.europa.eu/en/factpages/italy-2025-digital-decade-country-report (cons. 08/02/2026).

etc. In questo quadro, la e-lit, che nei contesti anglosassoni ha favorito convergenze tra DH, *media studies* ed *electronic art*, risulta difficilmente collocabile. Non riconducibile in modo univoco né alla letteratura comparata né a quella contemporanea, di cui non può essere un sottocampo, né all'informatica, essa resta ai margini dei settori disciplinari tradizionali.
Un ruolo potenzialmente centrale spetterebbe all'Associazione per l'Informatica Umanistica e la Cultura Digitale (AIUCD), principale punto di riferimento nazionale per le DH, che include implicitamente la letteratura elettronica nel proprio orizzonte disciplinare. Lo statuto dell'associazione, infatti, promuove sia la riflessione metodologica e teorica sull'uso delle tecnologie digitali nelle scienze umane, sia «la riflessione sui fondamenti umanistici delle metodologie informatiche e nel campo delle culture di rete»,[12] ambiti che intersecano direttamente la letteratura elettronica in quanto forma specifica di cultura digitale.[13] Eppure, questa apertura fatica a tradursi in pratica consolidata. Da una prima ricognizione qualitativa che ha coperto gli indici degli atti dei convegni dell'associazione per le edizioni 2013-2025 non sono emersi contributi riconducibili alla e-lit. Gli unici casi studio ambigui sono spesso collocabili più propriamente nell'ambito della gamification, della narrazione videoludica generalista o delle applicazioni educative, piuttosto che in quello della letteratura elettronica in senso stretto.[14] Anche la rivista *Umanistica Digitale*, legata all'associazione, mostra una presenza estremamente limitata di contributi affini, prevalentemente orientati verso la conservazione dell'arte digitale o le pratiche narrative dei social network, senza un'esplicita tematizzazione della e-lit.
Una marginalità analoga emerge anche sul piano storiografico. In alcune ricostruzioni recenti dell'informatica umanistica italiana si rintracciano riferimenti puntuali alla letteratura digitale: Marina Buzzoni, ad esempio, inserisce l'esperienza di *Tape Mark I* di Nanni Balestrini esplicitamente all'interno del suo contributo nel terzo volume di *Storia dell'informatica Italiana* (Buzzoni, 2025: 1129). Si tratta, tuttavia, di una presenza isolata. Sebbene sia necessario uno studio più sistematico, una simile assenza si riscontra anche in altri testi universitari sulle DH italiane, la cui attenzione si concentra incidentalmente solo sul primo romanzo ipertestuale mai pubblicato, *Afternoon, a story,* di Michael Joyce.[15]

6. CONVERGENZE TEORICHE E SPAZI DI DIALOGO

Nonostante ciò, le convergenze teoriche e metodologiche tra DH ed e-lit sono evidenti. Lo testimonia bene il volume *Electronic Literature as Digital Humanities,* curato da Dene Grigar e James O'Sullivan. Fin dall'introduzione, dopo aver ripercorso i rapporti dell'Alliance of Digital Humanities Organizations (ADHO) con centri e istituti internazionali di DH, i curatori osservano: «In sum, electronic literature is digital humanities because of our shared philosophy that a computer is not a tool or prosthesis that helps us to accomplish our work; rather, it is the medium in which we work» . Se a livello internazionale l'e-lit è vista come framework metodologico creativo per le DH , in ambito italiano la tensione resta teorica, ma non priva di convergenze. Già negli anni Novanta, Giuseppe Savoca, in una relazione letta a un convegno organizzato dall'Accademia Nazionale dei Lincei e dalla Fondazione IBM Italia, affronta il tema del dialogo tra informatica e letteratura . Partendo dal versante creativo della produzione letteraria, impiega l'espressione 'poesia elettronica' per reinterpretare retrospettivamente le sperimentazioni combinatorie di Nanni Balestrini alla luce del pensiero di Italo Calvino in *Cibernetica e fantasmi*. Punto di arrivo dello studio è quello analitico, delle concordanze, dei dati linguistici formalizzabili e quindi oggetto di studio critico dello studioso di informatica umanistica. Successivamente, Fiormonte, in *Scrittura e filologia nell'era digitale*, ha approfondito la nozione di scrittura elettronica concependo il testo come processo e prodotto, includendo nel suo studio, indirizzato alla filologia e scrittura digitale, disamine sulle esperienze letterarie ipertestuali, MUD e letteratura ergodica, in una prospettiva profetica che anticipa definizioni di

[12] https://www.aiucd.it/associazione/statuto/ (cons. 08/02/2026)
[13] Sulla necessità di una piena integrazione tra IU e CD sulla base di un nuovo fondamento epistemologico comune per garantire alle DH italiane un ruolo autonomo e non subalterno, si veda (Buzzetti et al., 2021)
[14] A titolo esemplificativo, si vedano i contributi della sessione *Contenuti 1 – George Boole* dell'edizione AIUCD 2022: Genovesi; Giungato & Miggiano; Zaterini in (Ciracì et al., 2022). Un'analisi sistematica degli atti AIUCD, della rivista *Umanistica Digitale* e di altre riviste italiane dell'area DH resta auspicabile come sviluppo futuro di questa ricerca
[15] La creazione di quest'opera ipertestuale procedette di pari passo con lo sviluppo della piattaforma di authoring Storyspace. L'opera fu quindi sia un artefatto letterario autonomo, sia un banco di prova per uno strumento poi utilizzato da altri autori e in contesti pedagogici, soggetto a sua volta di studi successivi, ed è diventata esempio di integrazione tra DH ed e-lit (Rettberg, 2016: 129). Il primo ipertesto italiano fu scritto, invece, da Lorenzo Miglioli nel 1993. *Ra-Dio* fu presentato a una conferenza a Reggio Emilia organizzata dal Gruppo '63 e pubblicato da ElettroLibri proprio insieme alla traduzione di *Afternoon, a story* di Michael Joyce. https://elmcip.net/creative-work/ra-dio (cons. 08/02/2026).

e-lit elaborate solo recentemente in ambito internazionale e italiano . Più recentemente, Gino Roncaglia, nella prefazione a *Poesie elettroniche* di Fabrizio Venerandi, ha parlato di primo esempio di letteratura elettronica matura, sottolineando:

> «non tanto l'esito estetico, e non tanto il particolare strumento informatico usato, ma la zona di confine e di interazione fra i due elementi rappresentati dalla regola (che si sostanzia nel codice di marcatura e di programmazione) e dal testo. Lo spazio di possibilità aperto da questa interazione dipende qui in maniera essenziale dagli strumenti usati: strumenti di produzione testuale, in mano all'autore, e strumenti di visualizzazione e lettura, in mano al lettore».

Questa opera nasce, sì, dal tentativo di ridefinire creativamente i limiti del formato ebook, ma la sua scrittura risiede nella codifica del testo tramite linguaggi di marcatura e di programmazione (HTML, CSS, JavaScript), chiamati a definire non solo l'aspetto formale dell'opera, ma anche la sua logica interattiva e interpretativa, proprio come avviene nelle pratiche di codifica in XML-TEI e visualizzazione delle DH.

7. CONCLUSIONI

L'analisi condotta mostra come in Italia esistano reali convergenze tra Digital Humanities e letteratura elettronica, ma come queste non si siano tradotte in una piena integrazione istituzionale e storiografica. La marginalità della e-lit non deriva da una mancanza di produzione o di rilevanza teorica, bensì da una specifica configurazione epistemologica delle DH italiane, storicamente orientate verso l'edizione, l'archivio e il patrimonio, categorie messe in crisi dall'ibridazione e instabilità delle opere di letteratura elettronica.
L'obiettivo di questo contributo non è rivendicare una sua inclusione forzata nelle DH, ma interrogare criticamente i confini del campo e stimolare una riflessione profonda sulle pratiche, i metodi e le istituzioni delle Digital Humanities contemporanee.

BIBLIOGRAFIA

Aarseth, E. (2023). Ludology. In M. J. P. Wolf & B. Perron (A c. di), The Routledge Companion to Video Game Studies (2a ed., pp. 255–259). Routledge. https://doi.org/10.4324/9781003214977-36

Busa, R. (2004). Foreword: Perspectives on the Digital Humanities. In S. Schreibman, R. Siemens, & J. Unsworth (A c. di), A Companion to Digital Humanities (p. XVI–XXI). Blackwell Publishing.https://doi.org/10.1002/9780470999875

Buzzetti, D. (2023). Prefazione. Oltre il limite istituzionale. In F. Ciotti (A c. di), Digital humanities: Metodi, strumenti, saperi (1a edizione, pp. 15–18). Carocci editore.

Buzzetti, D., De Ninno, A., & Fiormonte, D. (2021). Informatica Umanistica e Cultura Digitale. La sfida epistemologica. AIUCD 2021-DHs for Society: E-Quality, Participation, Rights and Values in the

Digital Age: Book of Extended Abstracts of the 10th National Conference, 319–326. https://doi.org/10.6092/unibo/amsacta/6712

Buzzoni, M. (2025). L'informatica umanistica. In M. Agosti & V. Cantoni (A c. di), Storia dell'informatica italiana—Prima parte (1a ed., Vol. 3). Cnr Edizioni. https://doi.org/10.36173/SII- parteprima

Ciotti, F. (2023). Introduzione. La galassia delle Digital Humanities. In F. Ciotti (A c. di), Digital humanities: Metodi, strumenti, saperi (1a edizione, pp. 19–34). Carocci editore.

Ciracì, F., Miglietta, G., & Gatto, C. (A c. di). (2022). AIUCD 2022—Culture digitali. Intersezioni: Filosofia, arti, media. Proceedings della 11a conferenza nazionale. https://doi.org/10.6092/unibo/amsacta/6848

De Vivo, F. (2011). eLiterature: La letteratura nell'era digitale. Definizione, concetto e statuto. In L. Masucci & G. Di Rosario (A c. di), Lavori del Convegno OLE Officina di Letteratura Elettronica, PAN Palazzo delle arti Napoli, January 20-21, 2011 (pp. 212–218). Atelier Multimediale Editrice.

Di Rosario, G. (2011). OLE: Officina di Letteratura Elettronica. In L. Masucci & G. Di Rosario (A c. di), Lavori del Convegno OLE Officina di Letteratura Elettronica, PAN Palazzo delle arti Napoli, January 20-21, 2011 (pp. 13–20). Atelier Multimediale Editrice. Eskelinen, M., & Di Rosario, G. (2012). Electronic literature publishing and distribution in Europe. University of Jyväskylä.

Fiormonte, D. (2003). Scrittura e filologia nell'era digitale. Bollati Boringhieri.

Grigar, D. (2021). Electronic Literature as Digital Humanities: An Introduction. In J. O'Sullivan & D.Grigar (a c. di), Electronic Literature as Digital Humanities: Contexts, Forms, & Practices (1a ed.).Bloomsbury Publishing Inc. https://doi.org/10.5040/9781501363474

Gruppo Giada. (2016). Diffusion, Analysis and Discussion of Electronic Literature in Italy (DADELI). Matlit Revista Do Programa de Doutoramento Em Materialidades Da Literatura, 4(2), 139–155. https://doi.org/10.14195/2182-8830_4-2_7 520 Atti del XV Convegno AIUCD 2026

Hayles, N. K. (2007, gennaio 2). Electronic Literature: What is it? The Electronic Literature Organization. https://eliterature.org/pad/elp.html

Iadevaia, R. (2021). Per una storia della letteratura elettronica. Mimesis Edizioni.

Jones, S. E. (2016). New Media and Modeling: Games and the Digital Humanities. In S. Schreibman, R. G. Siemens, & J. Unsworth (A c. di), A new companion to digital humanities (pp. 84–97). Wiley Blackwell. https://doi.org/10.1002/9781118680605

Mani, I. (2012). Computational Narratology. In P. Hühn & others (A c. di), The Living Handbook of Narratology. Hamburg University Press. http://lhn.sub.uni-hamburg.de/index.php/Computational_Narratology.html

Orlandi, T. (1992). Informatica umanistica: Realizzazioni e prospettive. VV. AA.: Calcolatori e Scienze Umane, Milano, 1–22.

Patti, E. (2022). Opera aperta: Italian electronic literature from the 1960s to the present. Peter Lang.

Rettberg, S. (2009). Communitizing Electronic Literature. Digital Humanities Quarterly, 003(2). https://dhq.digitalhumanities.org/vol/3/2/000046/000046.html

Rettberg, S. (2016). Electronic Literature as Digital Humanities. In S. Schreibman, R. G. Siemens, &

J. Unsworth (A c. di), A new companion to digital humanities (pp. 127–136). Wiley Blackwell. https://doi.org/10.1002/9781118680605

Rettberg, S., & Saum-Pascual, A. (2020). Introduction: Electronic Literature as a Framework for the

Digital Humanities. https://doi.org/10.7273/VYQE-DQ93

Saemmer, A. (2011). Digital arts and literature—Is it just a game? In L. Masucci & G. Di Rosario (A c. di), Lavori del Convegno OLE Officina di Letteratura Elettronica, PAN Palazzo delle arti Napoli, January 20-21, 2011 (pp. 366–377). Atelier Multimediale Editrice.

Savoca, G. (1992). Informatica e letteratura. In Calcolatori e scienze umane. Archeologia e arte, storia e scienze giuridiche e sociali, linguistica, letteratura. Atti del convegno organizzato dall'Accademia Nazionale dei Lincei e dalla Fondazione IBM Italia. Presentazione di E. Presutti, Prefazione di S. Moscati (pp. 289–301). Etas Libri.

Schreibman, S., Siemens, R. G., & Unsworth, J. (A c. di). (2016). A new companion to digital humanities. Wiley Blackwell. https://doi.org/10.1002/9781118680605

Schreibman, S., Siemens, R., & Unsworth, J. (2004). The Digital Humanities and Humanities Computing: An Introduction. In S. Schreibman, R. Siemens, & J. Unsworth (A c. di), A Companion to Digital Humanities (p. XXIII–XXVII). Blackwell Publishing. https://doi.org/10.1002/9780470999875

Venerandi, F. (2016). Poesie Elettroniche. Quintadicopertina, Nazione Indiana.

Venerandi, F. (2024). Manuale di letteratura elettronica. A cura di Vittoria Rubini. Prefazione di Roberta Iadevaia: Vol. 1. Avventure testuali e videogiochi narrativi. Argolibri.

[6/B] Co-creazione, ricerca partecipativa, comunità

Metodi e strumenti di sperimentazione tra co-creazione, Digital Humanities, e ricerca demo-etno-antropologica: la digitalizzazione come pratica relazionale ed etica.

ELEONORA DE LONGIS[1], MATTEO COVA[1], GIOVANNI BERTELLI[1]
[1]IISG

ABSTRACT (ITALIANO)

Il contributo presenta ed esamina gli approcci metodologici adottati nel corso di una sperimentazione sul campo, volta a coinvolgere attivamente una comunità di minoranza nel processo di digitalizzazione del proprio patrimonio culturale. L'articolo si focalizza, in particolare, sul caso di studio della comunità ladina della Val di Fassa. Il metodo co-creativo costituisce qui il fulcro dell'azione, trasformando la digitalizzazione da processo meramente tecnico in una pratica relazionale ed etica basata sul coinvolgimento diretto. Attraverso la *citizen science* e l'impegno empirico, la ricerca integra sensibilità e strumenti tipicamente etnografici per intercettare narrazioni, emozioni e valori identitari che sfuggono ai metodi tradizionali. La sperimentazione con la comunità ladina mira a consolidare un modello di governance comunitaria, dove le scelte tecnologiche sono subordinate alle priorità culturali della minoranza: il risultato finale vuole essere la proposta di un metodo strutturato per una transizione digitale capace di armonizzare ricerca scientifica, innovazione e tutela dell'identità.
Parole chiave: Co-creazione, Coinvolgimento del pubblico, digitalizzazione, patrimonio culturale; minoranze etniche

ABSTRACT (ENGLISH)

The paper presents and examines the methodological approaches adopted during a field experiment aimed at actively involving a minority community in the process of digitalizing its own cultural heritage. Specifically, the article focuses on the case study of the Ladin community of Val di Fassa. Here, the co-creative method serves as the cornerstone of the action, transforming digitalization from a purely technical process into a relational and ethical practice based on direct engagement. Through citizen science and empirical commitment, the research integrates typically ethnographic sensitivities to capture narratives, emotions, and identity values that traditional methods often overlook. The experimentation with the Ladin community aims to consolidate a model of community governance, where technological choices are secondary to the minority's cultural priorities. The ultimate goal is to propose a structured method for a digital transition capable of harmonizing scientific research, innovation, and the protection of identity.
Keywords: Co-creation, Public engagement, Digitisation, Cultural Heritage; Ethnic minorities

1. INTRODUZIONE: SPERIMENTAZIONE DI METODI TRA RICERCA PARTECIPATIVA, *DIGITAL HUMANITIES* ED ETNOGRAFIA

Il contributo propone ed esamina alcune possibilità e approcci metodologici della ricerca finalizzati a supportare il coinvolgimento diretto di comunità di minoranza, in questo caso etnico-linguistiche, nelle azioni di digitalizzazione del proprio patrimonio culturale. In particolare, si presentano metodi e strumenti frutto di una sperimentazione sul campo, testati nel corso della ricerca svolta nel progetto europeo Horizon "DIGICHer: digitisation of cultural heritage of minority communities for equity and renewed engagement" (www.digicher-project.eu), cui ha preso parte un team di ricercatori dell'Istituto Italiano di Studi Germanici.

Nel quadro del *public engagement*, qui inteso come coinvolgimento strutturato dei pubblici come co-ricercatori, la sperimentazione vuole mettere in relazione co-creazione e *citizen science* lungo la filiera: dalla definizione delle priorità alla raccolta dei dati, fino alla restituzione.

Nel contesto più ampio e complesso dei vari obiettivi e sfide affrontate dal progetto (si veda Cova, De Longis 2025), uno degli scopi centrali è stato quello di delineare – e sperimentare – dei modi di partecipazione attiva di una comunità alla transizione digitale, affinché quest'ultima risultasse etica su più livelli, ossia: che rispondesse alla visione della comunità nei confronti del futuro del proprio patrimonio

culturale; che fosse rispettosa, nella riproposizione dei beni digitali, dei valori e dei significati specifici che la comunità associa ad essi; che garantisse, al termine del processo, non solo una riappropriazione del bene stesso da parte della comunità (con la più forte connotazione di 'rimpatrio' dei beni in taluni casi) ma soprattutto un potenziamento del suo ruolo nelle scelte migliori da intraprendere per conservare, gestire, valorizzare e promuovere la propria cultura in digitale (queste e altre tematiche sono affrontate in Linkola-Aikio 2025). La finalità, in sintesi, è quella di ovviare al rischio di una scarsa rappresentazione dei membri della comunità nelle decisioni riguardanti l'intero processo, passando dai modelli *user-centred* tradizionalmente orientati all'accesso e alla fruizione verso una prospettiva *community-governed*, nella quale la comunità esercita una forma di autorità epistemica e progettuale sul patrimonio digitalizzato .
Il percorso si è aperto con una prima fase esplorativa e conoscitiva, nel corso della quale era stata indagata l'attuale situazione della digitalizzazione del patrimonio culturale in tre diverse comunità di minoranza partecipanti al progetto (Sámi della Lapponia, Ladina della Val di Fassa, ebraica in Europa), raccogliendo il generale 'sentimento' da loro espresso verso rischi e opportunità offerti dalla digitalizzazione.
Il presente contributo si focalizza sulla successiva fase di ricerca e analizza, in particolare, quanto svolto presso la comunità ladina della Val di Fassa (Trentino), come esempio cardine e caso di ricerca specifico all'interno del progetto. Nella pratica, si è trattato di attivare un progetto di digitalizzazione secondo i nuovi criteri e paradigmi proposti da DIGICHer.
Il lavoro svolto con il gruppo linguistico ladino fassano si inserisce in quella che è la fase di test della ricerca (*pilot testing*) in cui si è voluta provare l'applicazione, in un contesto reale, di una serie di pratiche co-creative e di *citizen science* ispirate dal design co-produttivo e ancor prima dall'indagine etnografica, proposte come strumenti all'interno di un quadro concettuale più ampio che riguarda tutta la sfera della digitalizzazione. Nel quadro di riferimento confluiscono infatti tutti gli aspetti – etici, giuridici, sociali, economici, tecnologici – che incidono sulla digitalizzazione del bene culturale della comunità, gli ostacoli che da ciascuno ne possono derivare, e una corrispettiva soluzione basata appunto su approcci di *community engagement*. L'obiettivo finale della sperimentazione, in termini essenziali, mira a dimostrare come sia possibile trasformare il processo di digitalizzazione del patrimonio culturale da un'attività puramente tecnica, e di frequente in passato portata avanti con approccio *top-down* dal mondo dell'accademia e della ricerca, a una pratica relazionale ed etica (Echavarria et al., 2022; Grcheva & Oktay Vehbi, 2021; Xanthaki et al., 2017).
Al centro dell'azione sul campo quindi, si pone il gruppo umano, il bene culturale identitario e l'ottica della co-creazione, intesa come buona pratica metodologica adottata dal ricercatore, dall'ente di conservazione del patrimonio culturale e dalle istituzioni, nello sviluppare rapporti virtuosi e progetti partecipativi con la comunità e i suoi membri.
È evidente che nell'atto di mettere al centro delle azioni di ricerca la minoranza e la comprensione della sua prospettiva, il lavoro svolto richiami a sé l'ispirazione, in taluni elementi, di un approccio tipicamente proprio degli studi di tipo demo-etno-antropologico o etnografico. Infatti, nonostante la ricerca sul campo nella fase dei test non abbia come intento quello dello studio etnografico in senso stretto e non abbia l'estensione di uno studio antropologico esaustivo, si trova pure a condividerne in tutto e per tutto gli spazi, i luoghi, i tempi, i contesti di svolgimento, le interazioni umane, e in taluni casi a mutuarne tanto la metodologia, quanto soprattutto gli attuali approcci etici (si vedano in generale Pavanello 2009; O'Reilly 2011; Bonetti 2024). Tipici dell'indagine qualitativa nella pratica etnografica e antropologica sono gli strumenti messi in campo, qui in duplice senso, ovvero per raccogliere dati e per attivare la comunità: la ricerca-azione, i *focus group*, le interviste libere o strutturate, lo *storytelling*. Il risultato diretto dell'applicazione di tali strumenti è normalmente la produzione di testi scritti, note di campo, registrazioni.
La strategia di intervento attraverso i *focus group*, in particolare, si allinea alle necessità e metodologie applicate nei più recenti studi in materia di ricerca etnografica e antropologica dedicata al lavoro con gruppi di persone. Anche i principi e meccanismi di selezione dei partecipanti, finalizzati a generare una discussione articolata e profonda sui temi analizzati, così come la distribuzione dei ruoli dei ricercatori impegnati tra facilitatori e osservatori, rispondono alle medesime metodologie di indagine (Bonetti 2024). Lo stesso può dirsi, dal punto di vista etico e deontologico, per l'impiego del consenso informato come autorizzazione a partecipare alla ricerca – pensato e adattato ai singoli casi e situazioni, per ciascuna comunità – imprescindibile oggi in ogni ricerca di interesse antropologico che preveda raccolta di dati sul campo, successiva analisi e conservazione.

In questa connotazione peculiare, dove coesistono diverse anime della ricerca sul campo, è stato programmato un percorso metodologico che si potesse muovere in questa intersezione di discipline, all'interno di un quadro di riferimento che supporta tale interdisciplinarietà.
L'uso di metodi di *participatory action research* si contrappone a strade più astratte o distaccate come il *survey* tradizionale, concretizzandosi attraverso un impegno empirico diretto con le 'realtà vissute' delle minoranze, proprio dell'etnografia dai lavori fondativi di Malinowski all'inizio del XX secolo in poi (Malinowski 2011). Le attività rientrano pienamente sia nella *community-based research*, in cui si utilizzano metodi che valorizzano la produzione di conoscenza situata, ovvero comprensibile solo all'interno delle cornici culturali specifiche, sia nella cosiddetta '*embodied interaction*', ovvero nella scelta di privilegiare attività in presenza per cogliere esperienze e priorità che non potrebbero essere colte con metodi di ricerca estrattivi o a distanza (si vedano Marcus 2011 e Boccagni 2019).
L'approccio di ricerca che ne risulta si colloca solo in parte nella scia di questi metodi etnografici tradizionali di impegno diretto (osservazione partecipante, interviste, analisi di contesti culturali e materiali d'archivio), pur mantenendo certamente il valore, ad esempio, dell'interazione, visione e analisi diretta dei beni culturali unitamente a membri della comunità, proprietari e conservatori (si prenda l'esempio delle lastre fotografiche dell'archivio Dantone Pascalìn in seguito trattate).
Ugualmente, si mantiene l'interazione viva con rappresentanti della comunità in qualità di interlocutori verso una componente più ampia del gruppo umano che li riconosce come tali, e dunque mediatori della prospettiva della comunità tutta. Di fatto, i test si sono avvalsi di esperti nel campo culturale come esperti viventi e co-progettisti, rientrando dunque a pieno titolo nei metodi della *citizen science*, capaci di importare e integrare nel processo decisionale i sentimenti e le narrazioni collettive.

2. GLI OBIETTIVI DELL'ATTIVITÀ SUL CAMPO E LA CORNICE CARE

Già le fasi preliminari avevano evidenziato l'importanza di porre la base di partenza della progettazione collaborativa su un approccio profondamente conoscitivo ed esplorativo, per delineare, come anticipato, il 'percepito' e il punto di vista peculiare della minoranza verso tutto ciò che riguarda la transizione digitale del proprio patrimonio culturale. Come si vedrà, ciò si è dimostrato essenziale per la definizione dei casi di studio e l'impostazione delle successive attività.
Infatti, oltre ai propositi generali posti dal progetto, le attività di co-creazione sono state misurate e pensate per il raggiungimento di obiettivi particolari, basati sulle necessità specifiche della comunità ladina, a loro volta emerse nella prima fase, in cui si è venuta a sviluppare una comprensione radicata del rapporto della comunità con il proprio patrimonio, raccogliendo appunto prospettive, preoccupazioni e aspettative specifiche della minoranza, tipicamente invisibili, nelle iniziative di digitalizzazione.
Obiettivi specifici delle attività di co-creazione con la comunità ladina sono stati: 1) valutazione di progetti di digitalizzazione basata su valori definiti dai partecipanti stessi, come la rappresentatività culturale e la risonanza emotiva, piuttosto che su criteri puramente tecnici; 2) possibilità di arricchimento partecipativo dei metadati: esplorare come la comunità possa contribuire attivamente all'arricchimento di metadati nel bene digitale, in particolare ponendo la lingua ladina quale elemento centrale di interpretazione e narrazione; 3) possibilità di recupero della memoria collettiva, attraverso la condivisione di conoscenze che possano arricchire ulteriormente il bene; 4) valutazione di scenari futuri riguardo alla gestione, all'accesso e alla responsabilità del proprio patrimonio digitale, ma soprattutto di riusabilità futura dello stesso, garantendo che le decisioni tecniche rimangano allineate con le aspirazioni a lungo termine della comunità. Operativamente, la *citizen science* è stata impostata come contributo partecipativo curato (non come *crowdsourcing* indistinto): raccolta di identificazioni/contestualizzazioni, discussione collettiva e validazione mediata. Questo consente di trasformare il coinvolgimento in produzione di conoscenza verificabile, mantenendo controllo comunitario e responsabilità.
Obiettivi specifici sono quindi andati delineandosi nel tempo, man mano che hanno preso forma le attività di co-creazione, ovvero laboratori estemporanei di confronto e ricerca, tenuti presso l'Istitut Cultural Ladin "Majon de Fascegn", con l'intenzione di lavorare a stretto contatto con la comunità, superando il metodo estrattivo e potenziando quello collaborativo.
Dal punto di vista operativo, come è stato quindi strutturato il coinvolgimento e il rapporto di ricerca con la comunità?
Se la minoranza ladina è stata coinvolta non come destinataria finale di un bene digitale, ma come soggetto collettivo competente e legittimato a intervenire nei processi decisionali relativi al proprio

patrimonio culturale, allora tale impostazione ha comportato il riconoscimento di un ruolo attivo nella definizione delle priorità, delle modalità di intervento e delle finalità della digitalizzazione.
Per rendere operativo tale ruolo, il coinvolgimento è stato impostato attraverso un campione intenzionale dei membri (tra 15 e 20 nel corso delle diverse fasi), volto a rappresentare la pluralità interna della comunità ladina, includendo soggetti riconducibili a diversi settori della collettività: società civile, istituzioni culturali, educazione e ambiti produttivi (corrispondenti alle dimensioni della quadrupla elica come definite da Carayannis 2009). Tale configurazione ha permesso la costruzione di un soggetto collettivo eterogeneo e competente, capace di esprimere prospettive e responsabilità differenziate rispetto al patrimonio culturale. L'Istitut Cultural Ladin ha svolto un ruolo centrale di mediazione, contribuendo a garantire legittimità, rappresentatività e fiducia nel processo di coinvolgimento.
Riguardo al riconoscimento del ruolo attivo della minoranza in tal senso, è importante sottolineare un elemento, sul piano etico, trasversale a tutte le iniziative di ricerca intraprese: DIGICHer ha assunto come riferimento i principi *CARE for Indigenous Data Governance* , reinterpretandoli come una cornice potenzialmente trasferibile e adattabile al contesto di una minoranza linguistica europea, cioè anche al di fuori della realtà della ricerca sui dati delle comunità indigene da cui sono scaturiti. Nel caso specifico, ad esempio, il potenziamento della comunità come soggetto decisore rende infatti operativi, in particolare, i principi di *Authority to Control* e *Collective Benefit*, traducendo i CARE da enunciazione normativa a pratica di governance del patrimonio digitalizzato. La digitalizzazione viene così concepita come processo relazionale, nel quale le scelte tecniche risultano subordinate a priorità culturali, valoriali e comunitarie.
Oltre a questa estensione, il progetto sfrutta il quadro CARE in modo innovativo in altre due direzioni. La prima risiede nell'applicazione dei principi. Costituiti come *framework* per la gestione dei dati indigeni, essi trovano normalmente attuazione o inserimento in strumenti di tipo giuridico o affine (linee guida, standard). DIGICHer si ispira ai principi CARE anche come base etica per la costruzione dell'interazione umana con le comunità di minoranza, progettando una metodologia o protocollo di azione partecipativa – dedicata al caso specifico di progetti che abbiano come *focus* il patrimonio culturale – che nel suo essere tale ambisce a certificarne la presenza e il rispetto. La seconda, come si espliciterà più avanti, è quella conseguente di utilizzare i principi CARE stessi come metro di valutazione dei metodi messi in campo.
Su queste basi, un primo momento metodologicamente qualificante si è tradotto nel coinvolgimento della comunità nella scelta, anzitutto, del bene culturale su cui concentrare l'azione di digitalizzazione. Questa scelta risponde ancora all'esigenza di scardinare strutturalmente la logica 'dall'alto' che caratterizza molti progetti di digitalizzazione, evitando che la partecipazione venga percepita come una concessione istituzionale, ma piuttosto come un qualcosa di legittimo e radicato nei significati attribuiti al patrimonio dalla comunità stessa. Di conseguenza, è stata avviata una valutazione comparativa dei beni culturali proposti dall'Istitut Cultural Ladin come possibili candidati alla digitalizzazione. Attraverso strumenti partecipativi, il digitale è stato trattato come oggetto di riflessione e negoziazione culturale, facendo emergere aspettative, timori e ambivalenze: la selezione del bene è avvenuta, quindi, sulla base di criteri discussi collettivamente, quali la rilevanza culturale, la rappresentatività comunitaria e il potenziale di riuso, collocando le valutazioni tecniche – inclusi *Open Access* e principi FAIR – a valle di una scelta culturale e comunitaria. Sul piano degli output digitali, questo si tradurrebbe nella costruzione di un'archiviazione virtuale della collezione (collezione digitale consultabile e citabile, con identificativi stabili e metadati tracciabili), in cui l'accesso e il riuso sono orientati all'Open Access quando culturalmente appropriato, prevedendo se necessario livelli di accesso e condizioni d'uso coerenti con la governance comunitaria.

3. LE FOTOGRAFIE DI DANTONE PASCALÌN E I RISULTATI DI METODO

La preferenza è ricaduta sulla digitalizzazione dell'archivio fotografico di Franz Dantone Pascalìn (1839–1909), figura fondamentale nella storia della Val di Fassa e ampiamente riconosciuto come il primo fotografo delle Dolomiti. L'archivio, oggi di proprietà degli eredi del Sig. Franco Dezulian e in parte già depositato presso l'Istitut Cultural Ladin, era in gran parte inedito e conservato in condizioni che rendevano necessario e urgente un intervento di conservazione strutturato. La collezione è composta da ben 926 negativi su lastra di vetro e 160 stampe fotografiche all'albumina. Gli scatti documentano paesaggi, trasformazioni, riti, feste, mestieri, e altri aspetti quotidiani della vita della valle tra la fine del secolo XIX e l'inizio del XX. Le lastre erano conservate in scatole originali e in tre casse di legno d'epoca, materiali che testimoniano il loro valore storico, ma che evidenziavano anche la fragilità del loro stato di conservazione. Il restauro delle lastre, grazie alla collaborazione tra proprietari e Istitut Cultural Ladin, ha

rappresentato molto più di un semplice sforzo tecnico: per la comunità significa un processo di recupero della memoria collettiva attraverso immagini che documentano vividamente la società ladina a cavallo tra due secoli, il profondo legame con il paesaggio montano, la vita rurale, le famiglie, le tradizioni, l'alpinismo, i primi sviluppi del turismo nelle Dolomiti e, come filo conduttore, la stessa lingua ladina, elemento fondamentale dell'identità locale.
Già da questa fase di confronto interno alla comunità, sono emersi nodi critici rilevanti, in particolare in merito ai rischi percepiti di decontestualizzazione e standardizzazione del patrimonio culturale in ambiente digitale, nonché le tensioni tra istanze di apertura dei contenuti e necessità di controllo comunitario. Sono inoltre emerse numerose preoccupazioni relative agli usi futuri del bene digitale, soprattutto in ambito educativo, e al rischio di semplificazione eccessiva di materiali culturalmente complessi.
Il secondo stadio di attività ha segnato il passaggio a una fase generativa, orientata alla riflessione su come l'archivio Dantone Pascalìn potesse essere arricchito, valorizzato, utilizzato e governato nel tempo a venire. L'arricchimento è stato qui inteso come pratica interpretativa e comunitaria, e non come mera produzione di metadati: la comunità ha discusso l'inclusione di narrazioni, memorie orali e saperi situati, rafforzando il legame tra patrimonio digitalizzato e identità comunitaria. Tale riflessione ha portato anche ad una rielaborazione comunitaria del principio di *reusability* compreso nei principi FAIR, interpretato come possibilità di riuso contestualizzato e non estrattivo. In questo senso, va notato come i principi FAIR siano stati bilanciati nella cornice CARE – secondo il naturale rapporto esistente tra i due sistemi di principi – evidenziando come la sostenibilità tecnica debba comunque essere negoziata in relazione a valori, finalità e responsabilità comunitarie. La discussione sugli usi futuri del bene digitale ha privilegiato, di fatto, proposte con una forte valenza educativa, culturale e intergenerazionale, sottolineando la necessità di modelli di governance capaci di garantire continuità, controllo comunitario e responsabilità nel lungo periodo.
Al termine del processo, il risultato che si vuole produrre, dal punto di vista della comunità, è una proposta concreta per l'utilizzo e la valorizzazione degli scatti di Dantone Pascalìn in quanto risorsa digitale della comunità stessa, che sia frutto del lavoro di progettazione partecipata e arricchimento, convogliando i desideri della minoranza ladina in fatto di utilità, riuso, potenzialità.
Dalla prospettiva interna dell'azione di ricerca, invece, il risultato è quello di poter validare un metodo di co-creazione e approccio etnografico con una minoranza sul tema della transizione digitale dei beni culturali, che garantisca il rispetto dei valori identitari e di conseguenza assicuri adesione ai principi CARE grazie alla sua natura di processo collaborativo.
Attraverso l'approvazione, da parte della comunità coinvolta, della proposta finale per la valorizzazione digitale dell'archivio, e grazie ad un sistema di raccolta strutturata di feedback (ancora con strumenti come intervista, *survey* o gruppi di discussione) presso i membri e i rappresentanti delle istituzioni culturali e politiche della minoranza ladina sulla qualità ed efficacia della ricerca partecipativa e dell'*engagement* messi in campo, si intende validare infine il metodo proposto.
Il livello di coinvolgimento e partecipazione della comunità può di fatto essere misurato attraverso uno strumento come la scala di Arnstein per la partecipazione (Arnstein 1969), ancora valida e preferibile al più recente modello del 'cubo della partecipazione' di Fung (Fung 2006), poiché centrata proprio a comprendere il livello di inclusione nei processi decisionali di gruppi marginalizzati o minoritari della società. La scala misura, attraverso una serie di parametri, il grado di partecipazione dei cittadini alle decisioni pubbliche, dal più basso livello di totale esclusione a quello più alto di controllo diretto. Questo metodo di valutazione vuole essere ulteriormente affiancato e incrociato con un confronto sul grado di presenza dei principi CARE. Come anticipato, infatti, la cornice dei CARE fornisce al progetto anche una sorta di griglia concettuale per la valutazione e validazione delle attività intraprese: ciò che si cerca di ottenere è il raggiungimento del livello più alto di coinvolgimento che permetta al tempo stesso anche il più alto grado di presenza dei principi CARE.
Il processo di la validazione, tutt'ora in corso d'opera, restituisce già risultati significativi. I dati raccolti, tra produzione scritta, note di campo, materiale registrato, in fase di analisi e interpretazione, indicano già alcuni spunti di riflessione.
Anzitutto, sulla scia delle attività intrattenute con la comunità ladina, si sta dimostrando l'efficacia di un approccio che riesce a colmare in più passaggi il cosiddetto "gap di coinvolgimento" e affrontare le questioni etiche, riportando alla minoranza l'autorità decisionale sul proprio patrimonio e armonizzando tra loro le necessità di ricerca scientifica, promozione culturale, istanze identitarie. Alla produzione dei dati si affianca una restituzione attraverso narrazioni digitali (percorsi tematici e micro-storie *place-based*

collegate alle fotografie), co-progettate con i partecipanti. In questo senso la restituzione funziona anche come comunicazione scientifica innovativa: contenuti bilingui, accessibili e riutilizzabili per pubblici diversi (scuole, territorio, comunità scientifica). Al centro delle azioni si pone il fulcro attorno al quale ruotano i diversi elementi, costituito appunto dal metodo stesso, il vero punto di contatto tra il bene culturale in sé, l'occhio esterno dello studioso, i desideri interni della collettività, il sapere comunitario, la prospettiva del digitale. Così, il ruolo attribuito ai membri della comunità di esperti viventi e paraetnografi (Marcus 2011) ha configurato i partecipanti non solo come informatori ma come contributori attivi al processo di progettazione – allineandosi a quanto già enunciato dai metodi della *citizen science* e alla ricerca comunitaria – apportando direttamente narrazioni collettive e conoscenze esperienziali.
Le attività hanno rivelato un'intersezione produttiva tra metodologie basate sulla co-creazione e sensibilità tradizionalmente associate alla ricerca etnografica e demo-etno-antropologica, mentre l'adozione dei principi della *participatory action research* ha configurato i laboratori come processi di impegno empirico diretto con le realtà vissute delle comunità di minoranza.
Proprio grazie a questa scelta metodologica è stato possibile cogliere un intero panorama emotivo, espanso, che circonda e sovrasta ciò che può sembrare dall'esterno un processo puramente tecnico come il portare in digitale una fotografia, un reperto, una tradizione orale o altro. Si tratta di un panorama composto da sensazioni, esperienze, priorità, preoccupazioni, a cui non sarebbe stato possibile accedere altrimenti, e che non sarebbe stato possibile comprendere adeguatamente attraverso strumenti di ricerca remoti o decontestualizzati.
Questa intersezione di metodi è, al momento, una delle possibili proposte e delle possibili strade percorribili in futuro dalla ricerca in questo campo. L'approdo finale vuole essere un programma metodologico strutturato, offerto non solo al mondo accademico, ma alle comunità di minoranza stesse e ai relativi enti di conservazione del patrimonio culturale per sperimentare modelli virtuosi di interazione tra i vari soggetti interessati dal fenomeno della transizione digitale, delle scienze legate alle *digital humanities* e, più in generale, allo studio del rapporto tra persone, identità e cultura.

BIBLIOGRAFIA

Arnstein, S. R. (1969). A Ladder of Citizen Participation. *Journal of the American Institute of Planners*, *35 (4)*, 216-24.

Boccagni, P. (2019). Multi-sited ethnography. In *SAGE Research Method Foundations*. https://doi.org/10.4135/9781526421036842870

Bonetti, R., & Natali, C. (2024). *La pratica della ricerca antropologica: Strumenti e metodologie*. Carocci.

Carayannis, E. G., & Campbell, D. F. J. (2009). "Mode 3" and "quadruple helix": Toward a 21st century fractal innovation ecosystem. *International Journal of Technology Management, 46*(3–4), 201. https://doi.org/10.1504/IJTM.2009.023374

Carroll, S. R., Garba, I., Figueroa-Rodríguez, O. L., Holbrook, J., Lovett, R., Materechera, S., Parsons, M., Raseroka, K., Rodriguez-Lonebear, D., Rowe, R., Sara, R., Walker, J. D., Anderson, J., & Hudson, M. (2020). The CARE Principles for Indigenous Data Governance. *Data Science Journal, 19*, Article 43. https://doi.org/10.5334/dsj-2020-043

Carroll, S. R., Garba, I., Plevel, R., Small-Rodriguez, D., Hiratsuka, V. Y., Hudson, M., & Garrison, N. A. (2022). Using Indigenous Standards to Implement the CARE Principles: Setting expectations through Tribal research codes. *Frontiers in Genetics, 13*, Article 823309. https://doi.org/10.3389/fgene.2022.823309

Cova, M., & De Longis, E. (2025). Il patrimonio culturale digitale delle minoranze etniche: Il progetto DIGICHer tra le comunità Sámi, ladine ed ebraiche in Europa. In S. Rebora, M. Rospocher, & S. Bazzaco (Eds.), *Diversità, equità e inclusione: Sfide e opportunità per l'informatica umanistica nell'era dell'intelligenza artificiale. Proceedings del XIV Convegno Annuale AIUCD2025* (pp. 72–77). AIUCD. https://amsacta.unibo.it/id/eprint/8380/

Echavarria, K. R., Samaroudi, M., Dibble, L., Silverton, E., & Dixon, S. (2022). Creative experiences for engaging communities with cultural heritage through place-based narratives. *Journal on Computing and Cultural Heritage, 15*(2), 1–19. https://doi.org/10.1145/3479007

Fung, A. (2006). Varieties of Participation in Complex Governance. *Public Administration Review*, 66, 66-75. https://doi.org/10.1111/j.1540-6210.2006.00667.x

Grcheva, O., & Oktay Vehbi, B. (2021). From public participation to co-creation in the cultural heritage management decision-making process. *Sustainability, 13*(16), Article 9321. https://doi.org/10.3390/su13169321

Linkola-Aikio, I.-A., Keskitalo, P., Ballardini, R., & Sarantou, M. (Eds.). (2025). *Digital Indigenous cultural heritage*. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-76941-2

Malinowski, B. (2011). *Argonauti del Pacifico occidentale*. Bollati Boringhieri.

Marcus, G. (2011). Multi-sited ethnography: Five or six things I know about it now. In S. Coleman & P. Hellerman (Eds.), *Multi-sited ethnography*. Routledge.

O'Reilly, K. (2011). *Ethnographic methods*. Routledge.

Pavanello, M. (2009). *Fare antropologia: Metodi per la ricerca etnografica*. Zanichelli.

Stickdorn, M., Hormess, M., Lawrence, A., & Schneider, J. (Eds.). (2021). *This is service design doing: Applying service design thinking in the real world: A practitioners' handbook* (10th reprint). O'Reilly Media.

Stickdorn, M., & Schneider, J. (Eds.). (2016). *This is service design thinking: Basics, tools, cases* (Paperback ed., 6th printing). BIS Publishers.

Xanthaki, A., Valkonen, S., Heinämäki, L., & Nuorgam, P. K. (Eds.). (2017). *Indigenous peoples' cultural heritage: Rights, debates, challenges*. Brill | Nijhoff. https://doi.org/10.1163/9789004342194

Integrare strumenti digitali e conoscenze della comunità per una progettazione urbana resiliente al clima: l'iniziativa Dundrum by Design

CHIARA COCCO[1], MATTIA LEONE[2], ANTONIO SAVINO[2], MIRUNA POPA[1], SARA TEDESCO[2]
[1] SCHOOL OF ARCHITECTURE, PLANNING AND ENVIRONMENTAL POLICY, UNIVERSITY COLLEGE DUBLIN, IRLANDA
[2] DIPARTIMENTO DI ARCHITETTURA, UNIVERSITÀ DEGLI STUDI DI NAPOLI FEDERICO II, ITALIA

ABSTRACT (ITALIANO)

Le città sono chiamate a sviluppare approcci di pianificazione urbana fondati su dati e conoscenze locali, improntati all'inclusività e orientati ai principi della giustizia climatica. Tuttavia, la pianificazione urbana ha storicamente faticato a integrare in modo efficace la partecipazione dei cittadini nelle prime fasi del processo progettuale. Dundrum by Design (DbD), un'iniziativa collaborativa in corso a Dundrum (Dublino, Irlanda), sperimenta un approccio partecipativo integrato che combina piattaforme digitali, mappature partecipate, strumenti di geodesign e apprendimento esperienziale per supportare processi di progettazione urbana orientati alla resilienza climatica sin dalle fasi iniziali. Sviluppato nell'ambito del progetto europeo PROBONO e del Dublin Living Lab, DbD coinvolge università, autorità locali e comunità attraverso un workflow sequenziale che integra raccolta di conoscenze locali, attività di co-creazione, simulazioni climatiche e valutazione partecipata di scenari progettuali. La metodologia si avvale di piattaforme digitali, attività ludico-didattiche e metodi computazionali di valutazione climatica degli scenari proposti. I risultati preliminari evidenziano la capacità dell'approccio di organizzare in modo coerente contributi eterogenei, rafforzare la continuità tra le fasi partecipative e generare un crescente livello di partecipazione comunitaria. Le simulazioni climatiche e gli scenari sviluppati forniscono alle autorità locali strumenti integrati per supportare decisioni informate e politiche climatiche contestualizzate. DbD dimostra come un approccio tecnologico multifattoriale e pedagogico possa migliorare la qualità e la legittimità dei processi di progettazione urbana e offrire spunti trasferibili ad altri contesti impegnati nella transizione climatica.
Parole chiave: co-creazione; geodesign; strumenti digitali; living labs; trasformazione urbana; resilienza climatica

ABSTRACT (ENGLISH)

Integrating digital tools and community knowledge for climate-sensitive urban design: the Dundrum by Design initiative.
Cities are called upon to develop urban planning approaches grounded in data and local knowledge, informed by inclusivity, and oriented towards the principles of climate justice. However, urban planning has historically struggled to effectively integrate citizen participation in the early stages of the design process. *Dundrum by Design* (DbD), an ongoing collaborative initiative in Dundrum (Dublin), is experimenting with an integrated participatory approach that combines digital platforms, participatory mapping, geodesign tools, and experiential learning to support climate-resilient urban design processes from the earliest stages. Developed as part of the European PROBONO project and the Dublin Living Lab, DbD involves universities, local authorities, and communities through a sequential workflow that integrates local knowledge gathering, co-design activities, climate simulations, and participatory evaluation of design scenarios. The methodology utilizes digital platforms, educational games, and computational methods for assessing the climate impact of the proposed scenarios. Preliminary results highlight the approach's ability to coherently organize heterogeneous contributions, strengthen continuity between participatory phases, and generate a growing level of community participation. The climate simulations and scenarios developed provide local authorities with integrated tools to support informed decisions and contextualized climate policies. DbD demonstrates how a multifactorial and pedagogical technological approach can improve the quality and legitimacy of urban design processes and offer insights that can be transferred to other contexts engaged in climate transition.
Keywords: co-design; geodesign; digital tools; living labs; climate-resilience; urban transformation

1. INTRODUZIONE

Le città impegnate in processi di transizione climatica sono sottoposte a una pressione crescente affinché adottino processi di pianificazione basati su dati empirici e processi socialmente inclusivi. Per decenni, la pianificazione urbana ha faticato a colmare il divario tra una significativa partecipazione dei cittadini e i processi di progettazione urbana (Hudson-Smith e Shakeri, 2022; Sondou et al., 2025). In una società sempre più urbanizzata e digitalmente avanzata, emerge una rinnovata esigenza di metodologie volte a promuovere lo sviluppo di città resilienti e sostenibili, con un approccio incentrato sulle persone (Kavouras et al., 2025). Sebbene gli strumenti digitali partecipativi abbiano ampliato le opportunità di coinvolgimento della comunità, essi sono spesso utilizzati in modo isolato, in fasi avanzate del ciclo di pianificazione o senza meccanismi chiari per tradurre la conoscenza e il contributo dei cittadini in risultati concreti in termini di progettazione e politiche. Promuovere la co-costruzione del sapere e la co-creazione con le comunità fin dalle prime fasi del processo progettuale alla scala urbana può favorire lo sviluppo di quartieri verdi e sostenibili (Buijs et al., 2024; Kavouras et al., 2025). Il presente contributo analizza come un approccio integrato, che combini più piattaforme digitali con forme di partecipazione attiva dei cittadini e dei portatori di interesse locali, possa supportare un processo di progettazione urbana orientato alla resilienza climatica sin dalle prime fasi.

La ricerca si basa su *Dundrum by Design* (DbD), un'iniziativa collaborativa avviata nel giugno 2025 e attualmente in corso, sviluppata a Dundrum, un centro suburbano di Dublino che sta attraversando una significativa trasformazione connessa alla riconfigurazione della mobilità, al miglioramento degli spazi pubblici e al perseguimento degli obiettivi di adattamento e mitigazione climatica. DbD è realizzato attraverso una partnership tra l'*University College Dublin* (UCD), l'Università degli Studi di Napoli Federico II (UNINA), il *Dún Laoghaire–Rathdown County Council* (DLRCoCo) e le comunità locali. Questa collaborazione si inserisce nel progetto PROBONO finanziato dall'UE, che adotta l'approccio del Living Lab (Voytenko et al., 2016) per combinare innovazioni tecniche e sociali avanzate a supporto della transizione climatica. L'iniziativa è allineata ai quadri di pianificazione previsti dalla normativa vigente, tra cui il Dundrum Local Area Plan (2023), e sperimenta modalità di partecipazione alternative che superino i metodi di consultazione convenzionali.

Attraverso il PROBONO Dublin Living Lab, DbD consente di testare e valutare metodi e strumenti innovativi, come il geodesign, per coinvolgere un'ampia gamma di portatori di interesse, comprese le comunità locali, all'interno del processo di transizione climatica. L'attenzione è rivolta alle vie storiche di Dundrum, in particolare *Main Street* e *Taney Cross*, attualmente interessate da trasformazioni significative. La visione degli enti locali (DLRCoCo) è quella di rendere Dundrum un centro urbano più accessibile ai pedoni, con spazi pubblici e parchi attivi e accoglienti, valorizzandone al contempo il carattere storico e sostenendo un'economia locale dinamica.

Attraverso l'impiego di più piattaforme digitali, DbD invita inoltre la comunità a segnalare ulteriori aree che necessitano di interventi. I contributi dei cittadini durante tutto il processo partecipato sono condivisi con DLRCoCo e verranno presi in considerazione in particolare nell'ambito del progetto *Dundrum Living Streets*, un'iniziativa dell'amminstrazione locale volta a promuovere la mobilità sostenibile e il miglioramento dello spazio pubblico.

L'approccio di DbD utilizza piattaforme digitali al fine di semplificare e diversificare i processi di partecipazione. Tali piattaforme consentono di co-costruire il sapere con le comunità locali, rappresentare i dati e la conoscenza in maniera interattiva, simulare diversi scenari climatici e integrare interventi progettuali informati e elaborati sulla base dei contributi della comunità. Le simulazioni climatiche forniscono a DLRCoCo soluzioni contestualizzate basate su dati empirici e proiezioni future. Queste possono essere utilizzate sia per valutare misure di resilienza al cambiamento climatico, sia per testare diversi scenari progettuali.

Il progetto DbD si basa inoltre su precedenti workshop di geodesign condotti dall'UCD (Cocco et al., 2025), incorporandone insegnamenti e metodologie, e si allinea parallelamente a quadri più ampi, come il *New European Bauhaus*[1]. Il contributo suggerisce che l'adozione di un approccio tecnologico multifattoriale nei processi di progettazione urbana possa supportare le amministrazioni locali e le comunità nell'immaginare scenari futuri e nel promuovere una visione olistica e sistemica di Dundrum.

[1] https://new-european-bauhaus.europa.eu/

2. METODOLOGIA

Anziché concepire la partecipazione come un singolo evento consultivo, DbD è strutturato come un processo partecipativo sequenziale, all'interno del quale dati, idee e decisioni sono progressivamente raffinati attraverso strumenti e attività interconnessi.
DbD mira a riunire diversi stakeholder in un processo collaborativo e a promuovere l'assunzione di un ruolo attivo da parte dei cittadini all'interno del processo di progettazione urbana. La metodologia proposta incoraggia diverse forme di partecipazione e raccolta dati, integrando conoscenze locali e analisi esperte.
Il processo inizia con una fase preliminare volta all'identificazione degli *stakeholder* e delle condizioni climatiche locali (fase preparatoria), seguita da una fase di raccolta delle conoscenze, in cui residenti, giovani e studenti condividono prospettive attraverso sessioni di mappatura, giochi e discussioni. Il flusso di lavoro prosegue quindi con workshop operativi che utilizzano metodi e strumenti di geodesign, attività ludico-didattiche nelle scuole e votazioni comunitarie per definire proposte condivise (fase di workshop). Nella fase finale (*post-processing*), tali idee vengono ulteriormente raffinate attraverso simulazioni climatiche e raccolte in un report che riflette la visione collettiva, includendo un forte contributo dei partecipanti più giovani.
Nel corso del processo DbD ha fatto ricorso a tre principali piattaforme digitali, descritte di seguito, per garantire un coinvolgimento sostanziale dei diversi *stakeholder* e della più ampia comunità di Dundrum.
La prima fase ha riguardato la produzione di conoscenza a partire dal sapere delle comunità, supportata attraverso la piattaforma *Meitheal Mapping*[2] (vd. Fig.1). È stata condotta mappatura partecipata online che ha consentito ai residenti di contribuire con osservazioni, priorità e criticità georeferenziate relative a sistemi urbani, quali infrastrutture verdi, mobilità, patrimonio e spazi pubblici. Questa fase ha permesso di costruire una base di conoscenze condivisa, fondata sull'esperienza locale.

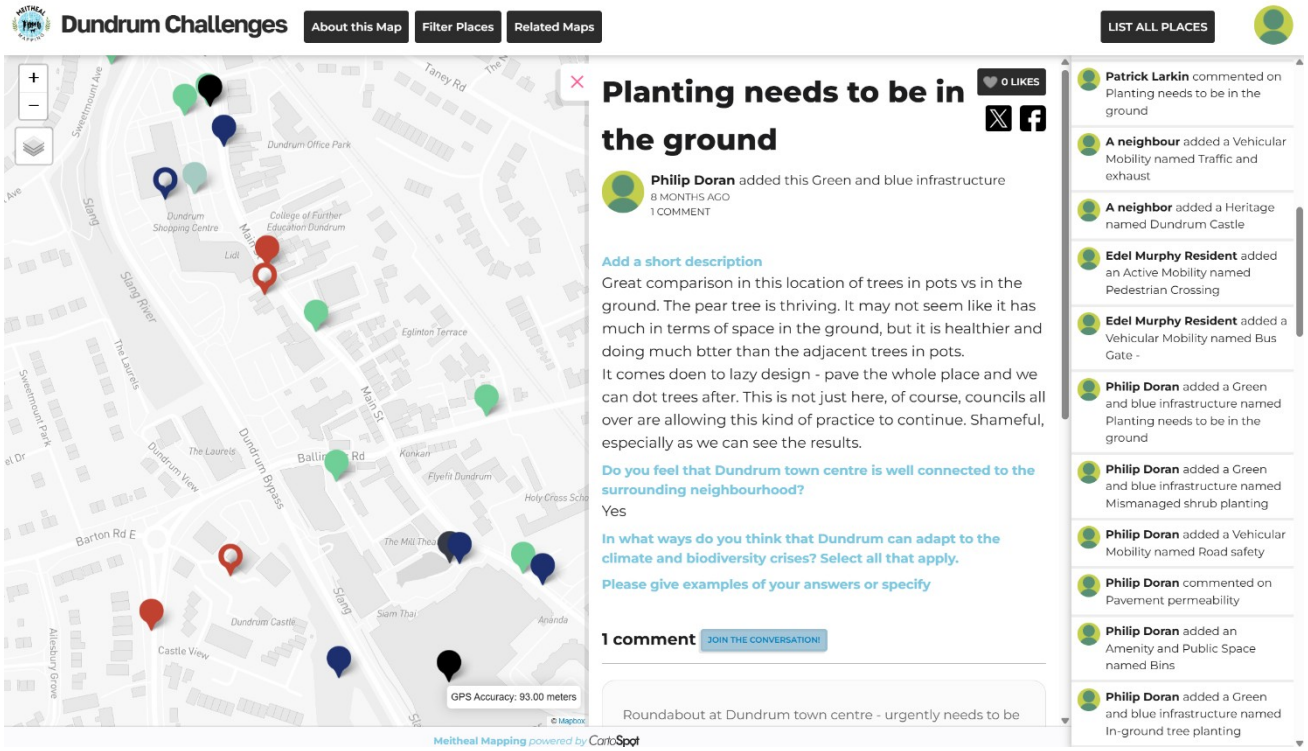


Figura 1. La piattaforma online *Meitheal Mapping* è stata utilizzata durante la fase di raccolta delle conoscenze

I contributi raccolti tramite *Meitheal Mapping* sono stati utilizzati come punto di partenza durante la fase di workshop svoltasi a settembre 2025, dove si è sperimentato l'uso di *GeodesignHub*[3] come ambiente collaborativo di supporto decisionale. In questo contesto, membri della comunità, studenti e altri

[2] https://meitheal.cartospot.com/
[3] https://www.geodesignhub.com/#collab

stakeholder hanno sviluppato proposte progettuali, intrapreso processi negoziali e costituito coalizioni focalizzate su priorità condivise. Nella fase di *post-processing* dei risultati del workshop sono state create due mappe: la prima evidenzia le priorità condivise su cui i partecipanti hanno raggiunto un accordo (*Community Desires Map*); la seconda, contenente 18 proposte progettuali su cui non è stato raggiunto un consenso, è stata pubblicata sulla piattaforma di consultazione online del DLRCoCo, *Citizen Space*[4], ed è stata utilizzata per una votazione pubblica. Questo ulteriore passaggio ha contribuito a rafforzare la trasparenza dell'intero processo e di estendere la partecipazione oltre ai soli partecipanti del workshop. Per raccogliere contributi da parte dei giovani di Dundrum è stata impiegata una combinazione di *serious games* digitali e analogici. Minecraft è stato utilizzato come strumento principale per stimolare l'interesse dei giovani cittadini verso il futuro di Dundrum, consentendo la co-creazione di spazi e proposte dal loro punto di vista (Poplin et al., 2022). Lo strumento, intuitivo e attrattivo per i più giovani, ha talvolta coinvolto anche gli adulti (vd, Fig.2). Il suo utilizzo durante le prime sessioni di partecipazione, organizzate presso la biblioteca del quartiere, ha contribuito indirettamente alle attività di mappatura partecipata, poiché i genitori, in attesa che i figli completassero le attività con Minecraft, hanno inserito almeno un contributo sulla piattaforma *Meitheal Mapping*. È attualmente attivo un server di Minecraft per raccogliere contributi da giovani di età compresa tra gli 8 e i 18 anni, e sono già stati svolti cinque workshop nelle scuole locali che hanno coinvolto oltre 160 studenti.



Figura 2. Attività di progettazione partecipata rivolte ai giovani. L'iniziativa ha integrato strumenti analogici, come i giochi di carte (sinistra), e digitali, attraverso Minecraft (destra), per raccogliere visioni sul futuro di Dundrum. Il coinvolgimento attivo di alcuni genitori ha favorito il dialogo intergenerazionale sulle preferenze urbane e i bisogni della comunità.

Parallelamente, gli studenti di architettura hanno tradotto le indicazioni emerse dalla comunità in proposte di progettazione urbana. Due giornate dedicate all'occupazione temporanea dello spazio pubblico e curate dagli studenti hanno consentito ai residenti di esplorare attivamente e commentare tali proposte, trasformando uno spazio sottoutilizzato attraverso interventi temporanei di *tactical urbanism*. Tutto il processo di coinvolgimento della comunità e di co-progettazione è costantemente supportato da tre canali di comunicazione ufficiali (sito web[5], social media e mailing list) nonché da metodi di coinvolgimento tradizionali (volantini, poster, installazioni, incontri con gruppi comunitari, ecc.). Nel loro complesso, le piattaforme digitali consentono un flusso continuo di dati dalla fase iniziale di costruzione della conoscenza fino alla presentazione finale, integrando progressivamente i contributi raccolti.

[4] https://dlrcoco.citizenspace.com/
[5] https://dundrumbydesign.ie/

A supporto di un processo decisionale orientato a migliorare la resilienza climatica di Dundrum, DbD è affiancato da un'attività analitica continua, condotta da ricercatori dell'UNINA, che si avvale del toolkit dell'*Urban Climate Change Research* Network (UCCRN). Il toolkit integra strumenti GIS e di modellazione computazionale per valutare i rischi climatici, mappare l'effetto isola di calore urbano e le aree soggette a fenomeni di allagamento, analizzare i compromessi e quantificare i co-benefici ambientali, economici e sociali generati (Raven et al., 2025).
I risultati preliminari del progetto DbD sono qui interpretati in una fase iniziale alla luce di modelli consolidati di pianificazione partecipativa digitale (Lee et al., 2022; Ataman et al., 2025; Atzmanstorfer et al., 2025), con particolare riferimento all'efficacia del workflow – in termini di fattibilità, valore percepito, capacità di orientare l'azione e sostenibilità - e all'impatto dei processi partecipativi. Gli esiti del workshop di geodesign e delle attività con *Meitheal Mapping* configurano un'analisi di scenario guidata dalla comunità, che facilita una valutazione contestuale delle condizioni climatiche locali. Le simulazioni, corrispondenti a diverse visioni e proposte progettuali (vd. Fig.3), sono state generate in tempo reale e presentate durante il secondo workshop comunitario, tenutosi a gennaio 2026, e sono attualmente in fase di analisi. Nel loro insieme, questi strumenti possono fornire un sistema integrato per supportare processi decisionali informati e per la valutazione dell'impatto delle varie proposte progettuali.



Figura 3. Analisi climatica e scenari progettuali per Taney Cross. A sinistra, la presentazione durante il secondo workshop comunitario delle simulazioni sull'effetto isola di calore urbano relative allo stato di fatto (Scenario 0), elaborate con gli strumenti UCCRN. A destra, una vista assonometrica di uno degli scenari di trasformazione (Scenario 1), sviluppato integrando le priorità e i bisogni identificati dai residenti durante le varie fasi del progetto.

3. RISULTATI PRELIMINARI

I risultati raccolti finora evidenziano l'efficacia di DbD nell'organizzare in modo efficiente i contributi tra le varie piattaforme, reso possibile dall'approccio tecnologico multifattoriale adottato.
Promuovere un ruolo attivo dei cittadini nella progettazione del proprio futuro ha favorito una più ampia adesione a DbD come iniziativa orientata al cambiamento tangibile. Un risultato evidente di questo approccio è che la mailing list di DbD è cresciuta da 19 a 144 iscrizioni in 6 mesi. Nel frattempo, la prima piattaforma utilizzata per registrare i contributi dei cittadini, *Meitheal Mapping*, ha raccolto circa 121 osservazioni e proposte, successivamente incrementate a 169 dopo il workshop di geodesign tenutosi a fine settembre, includendo gli interventi degli studenti e l'insieme dei contributi dei partecipanti. Sebbene il numero complessivo sia stato successivamente limitato a 92 proposte relative all'iniziativa Living Streets Dundrum, l'evoluzione positiva risulta evidente. La pagina di votazione su *DLR Citizen Space* ha raccolto 69 voti complessivi su 18 idee relative a Living Streets nella prima settimana e più di 400 in un mese. I dati relativi a Minecraft, nell'ambito del coinvolgimento dei giovani, sono attualmente in fase di raccolta e analisi.
I risultati attesi della modellazione climatica comprendono tre principali risultati operativi. La simulazione delle condizioni ambientali attuali fornirà report tecnici e mappe tematiche basate su GIS e modellazione 3D. Lo sviluppo di scenari di uso del suolo illustrerà soluzioni di progettazione urbana resilienti al clima, informate dalle 30 proposte individuate come prioritarie durante il workshop e successivamente integrate

con le idee emerse attraverso il processo di votazione (*Community Desires Map*). La modellazione secondo scenari di pianificazione costituisce la valutazione finale prevista e proporrà interventi climatici, valutando il consumo energetico e l'impronta di carbonio.

4. DISCUSSIONE E CONCLUSIONI

DbD rappresenta una direzione innovativa nel modo in cui i contributi dei cittadini vengono integrati nei processi di progettazione urbana nelle fasi iniziali del dibattito progettuale. Migliorare l'accessibilità e considerare fattori esterni, quali il livello di alfabetizzazione tecnologica, contribuisce a rafforzare la fiducia e le relazioni, favorendo una progettazione sensibile al clima. Semplificando il processo di coinvolgimento e sfruttando i vantaggi delle diverse piattaforme, l'iniziativa offre molteplici momenti strutturati in cui la comunità può fornire contributi.
L'adozione di un approccio pedagogico all'interno della partnership DbD ha contribuito a pratiche innovative e alla creazione di uno spazio di scambio di conoscenze tra studenti e comunità. Tra i principali benefici per tutti i partner – UCD, UNINA, DLRCoCo e la comunità di Dundrum – rientrano lo sviluppo di soluzioni di progettazione urbana informate, fondate su simulazioni realistiche e validate; il rafforzamento della pianificazione a lungo termine; e l'elaborazione di strategie attraverso strumenti di modellazione sviluppati durante i workshop, che consentiranno al DLRCoCo di testare diversi scenari di pianificazione e di definire soluzioni scalabili per un'implementazione più ampia, basate sui contributi della comunità.
I risultati preliminari indicano che l'impianto integrato di DbD rafforza la continuità tra le diverse fasi partecipative, riduce la frammentazione spesso riscontrata nel coinvolgimento digitale e favorisce una comprensione sistemica del cambiamento urbano. L'iniziativa ha inoltre generato un coinvolgimento continuativo della comunità, un archivio in crescita di proposte radicate nel contesto locale e risultati di apprendimento significativi per gli studenti coinvolti come facilitatori e progettisti. Permangono, tuttavia, alcune criticità legate all'accesso digitale, alla rappresentatività e all'interpretazione dei dati qualitativi, ampiamente riconosciute dalla comunità scientifica e dalla letteratura sui processi partecipativi come aspetti che richiedono particolare attenzione e continuo affinamento metodologico. Successive ricerche e metodologie potranno esplorare l'integrazione tra approcci tecnologici e tradizionali per costruire un quadro più olistico, aumentare la partecipazione nelle fasi iniziali e prevedere tempistiche più estese, consentendo a tutti i membri della comunità di contribuire una volta che il progetto acquisisce maggiore credibilità.
Nel complesso, *Dundrum by Design* dimostra come la combinazione di strumenti digitali, apprendimento esperienziale e analisi climatica all'interno di un quadro partecipativo integrato possa migliorare tanto la qualità quanto la legittimità dei processi di progettazione urbana. L'utilizzo combinato di diverse tecnologie digitali offre nuove opportunità per promuovere la co-produzione della conoscenza, la partecipazione e la co-creazione. Pur essendo fortemente contestualizzato, l'approccio risulta adattabile e offre spunti trasferibili a enti locali e istituzioni accademiche impegnati nel sostenere processi di transizione climatica equi e inclusivi.

BIBLIOGRAFIA

Ataman, C. et al. (2025). From insight to action: An integrated assessment framework for digital citizen participation in data-centric urban practices, Cities, 156, p. 105545.

Atzmanstorfer, K. et al. (2025). Critical Success Factors of Participatory Community Planning with Geospatial Digital Participatory Platforms, ISPRS International Journal of Geo-Information, 14(4), p. 153.

Buijs, A. E. et al. (2024) 'Advancing environmental justice in cities through the Mosaic Governance of nature-based solutions', Cities, 147(February), p. 104799. doi: 10.1016/j.cities.2024.104799.

Cocco, C. et al. (2025). Co-creating strategies for heritage reactivation: HeritACT pilot cases in Eleusis, Milan and Ballina, Journal of Cultural Heritage Management and Sustainable Development, ahead-of-print(ahead-of-print). doi: 10.1108/JCHMSD-04-2025-0083/1268634.

Hudson-smith, A. and Shakeri, M. (2022). The Future's Not What It Used To Be: Urban Wormholes, Simulation, Participation, and Planning in the Metaverse, Urban Planning, 7(2), pp. 214–217. doi: 10.17645/up.v7i2.5893.

Kavouras, I. et al. (2025). Empowering Communities Through Gamified Urban Design Solutions, Smart Cities, 8(2), pp. 1–24. doi: 10.3390/smartcities8020044.

Lee, G. Y. et al. (2022). Presenting a comprehensive multi-scale evaluation framework for participatory modelling programs: A scoping review, PLOS ONE, 17(4), p. e0266125.

Poplin, A., de Andrade, B., & de Sena, Í. (2022). Let's discuss our city! Engaging youth in the co-creation of living environments with digital serious geogames and gamified storytelling. Environment and Planning B: Urban Analytics and City Science, 50(4), 1087-1103. https://doi.org/10.1177/23998083221133828 (Original work published 2023)

Raven, J., Leone, M. F., & Nocerino, G. (2025). The Urban Design Climate Workshop: Integrating Mitigation and Adaptation Strategies to Confront Urban Climate Change. In Urban Climate and Urban Design (pp. 167-179). Singapore: Springer Nature Singapore.

Sondou, T. et al. (2025). Urban Planning Through Participatory Democracy: Analysis of Citizen Participation in Urban Planning in Ho (Ghana) and Kpalimé (Togo), Sustainability (Switzerland). doi: 10.3390/su17031161.

Voytenko, Y., et al. (2016). Urban living labs for sustainability and low carbon cities in Europe: Towards a research agenda. Journal of cleaner production, 123, 45-54.

HCI e AI per il Public Engagement: casi di studio per la valorizzazione del patrimonio culturale

SAMUEL ALDO IACOLINA[1], MANUELA ANGIONI[2], VALENTINA MAROTTO[3], FRANCESCA MURA[4], PIERGIORGIO PALLA[5]

[1] CRS4, SAMUEL@CRS4.IT
[2] CRS4, ANGIONI@CRS4.IT
[3] CRS4, VALE@CRS4.IT
[4] CRS4, FRAMURA@CRS4.IT
[5] CRS4, PPALLA@CRS4.IT

ABSTRACT (ITALIANO)

Il contributo presenta i casi d'uso relativi alla valorizzazione del patrimonio culturale attraverso l'integrazione di sistemi di Artificial Intelligence e tecnologie di Human Computer Interaction. Il paper analizza due casi studio principali. Il progetto Sinnos, focalizzato su tecnologie AI-driven per l'analisi dei contenuti e installazioni interattive per la fruizione multimodale. Il secondo caso di studio preso in rassegna è la Casa Museo Cambosu, esempio di musealizzazione immersiva dedicata alla letteratura basata sulla metafora dell'archivio vivente. Un focus specifico, seppur di alto livello, è dedicato alle tecnologie di AI adottate e ai sistemi di accessibilità implementati, con la descrizione di sistemi tattili, interfacce gestuali e modelli generativi per l'adattamento e la fruizione dei contenuti.
Parole chiave: Interazione Uomo Macchina; Intelligenza Artificiale Generativa; Digital Storytelling; Accessibilità Museale; Archivio Digitale;

ABSTRACT (ENGLISH)

HCI and Artificial Intelligence for Cultural Heritage Enhancement: The Case of Narrative Museums and Living Archives. This paper presents the use cases emerging from the research activities in the field of cultural heritage valorization through the integration of Artificial Intelligence and Human Computer Interaction. The paper examines two main case studies. The first is the Sinnos project, focused on AI-driven technologies for content analysis and on interactive installations designed for multimodal cultural engagement. The second case study is the Casa Museo Cambosu, an example of immersive musealization dedicated to literature and built around the metaphor of the "living archive." A high-level overview is also provided of the AI technologies adopted and the accessibility systems implemented, including tactile interfaces, gesture-based interaction and generative models for content adaptation and enhanced user experience.
Keywords: Human-Computer Interaction; Generative Artificial Intelligence; Digital Storytelling; Museum Accessibility; Digital Archive;

1. INTRODUZIONE

Le Digital Humanities ricoprono oggi un ruolo cruciale nella definizione di nuove pratiche di ricerca partecipativa, dove il coinvolgimento del pubblico diviene elemento strutturale nella produzione di conoscenza. In questo scenario, il gruppo HCI del CRS4 opera all'intersezione tra tecnologie interattive, intelligenza artificiale e patrimonio culturale, con l'obiettivo di sviluppare ecosistemi digitali che superino la mera fruizione passiva. La trasformazione degli archivi culturali in spazi "vivi", in cui l'interazione gestuale, vocale e tattile abilita nuove forme di storytelling esperienziale, rappresenta il cuore di questa ricerca. Questo contributo illustra un quadro metodologico e applicativo che integra AI generativa e soluzioni di hardware affordability per creare installazioni sostenibili e accessibili, rispondendo alla necessità di rendere il patrimonio culturale sardo fruibile a un pubblico vasto ed eterogeneo attraverso casi studio reali come il progetto Sinnos e la Casa Museo Cambosu.

L'evoluzione recente dell'Intelligenza Artificiale (AI) sta ridefinendo radicalmente il settore dei beni culturali e artistici (Mordor Intelligence, 2024), trasformandosi da strumento di supporto tecnico a infrastruttura cognitiva trasversale per la produzione, la gestione e la mediazione dei contenuti (AI4Culture, 2024). Secondo le analisi di Gartner, la Generative AI rappresenta oggi il principale motore di innovazione per il public engagement, classificando il suo impatto in aree strategiche quali la creatività multimodale, le interfacce conversazionali avanzate e l'analisi predittiva dei comportamenti dei visitatori

(Gartner, 2025). Il mercato globale riflette questa espansione: la spesa per soluzioni AI nella gestione dei dati culturali è in forte crescita, trainata dalla necessità di creare esperienze personalizzate e scalabili (Gartner, 2025).
Nel contesto museale e archivistico, l'adozione dell'AI si manifesta su due livelli principali. Sul piano operativo e della conservazione, progetti come l'iniziativa nazionale I.PaC coordinata da Cineca dimostrano come tecnologie di riconoscimento ottico (OCR/HCR), estrazione di entità (NER) e object detection possano automatizzare la metadatazione di vasti archivi digitali, rendendo accessibili materiali complessi come manoscritti storici e fondi fotografici (Cineca, (2025). Analogamente, istituzioni internazionali come gli Harvard Art Museums utilizzano la computer vision per la catalogazione automatica (Harvard Art Museums, 2023), mentre progetti di ricerca come ARTEMISIA (Sapienza e CNR) applicano l'analisi predittiva per ottimizzare i flussi di visita.
Sul piano della fruizione e del digital storytelling, l'innovazione è guidata dall'avvento dei modelli multimodali nativi (es. GPT-4o, Gemini), capaci di elaborare e generare simultaneamente testo, immagini e audio. Queste architetture abilitano la creazione di Digital Humans e avatar interattivi che, superando i limiti dei chatbot tradizionali, offrono un'interazione empatica basata su sintesi vocale naturale ed espressioni facciali realistiche, fungendo da guide virtuali evolute (McKinsey & Company, 2024). La generazione di contenuti visivi e audiovisivi ha raggiunto livelli di fotorealismo inediti grazie a modelli come Stable Diffusion (per le immagini) e Sora o Veo (per il video), permettendo la creazione di narrazioni storiche immersive e ricostruzioni ambientali con elevata coerenza fisica e temporale.
Un ambito di applicazione cruciale è quello dell'accessibilità, dove l'AI offre soluzioni per superare le barriere sensoriali e cognitive (ASPHI, LIA, IBM, & IFAB, 2024). Tecnologie come TANGENT e Arte Tattile sfruttano superfici intelligenti per trasformare opere visive in esperienze multisensoriali e tattili per non vedenti (Collettivo Digitale, 2024). Parallelamente, per l'accessibilità cognitiva, progetti come Limpide Letture (ASPHI, IBM) impiegano Large Language Models (LLM) per semplificare testi complessi secondo i criteri dell'approccio Easy-to-Read, rendendo il patrimonio comprensibile a pubblici con difficoltà di apprendimento.
In sintesi, lo stato dell'arte delinea un ecosistema in cui l'AI non sostituisce la curatela umana, ma ne amplifica le potenzialità, abilitando forme di musealizzazione adattiva in cui l'archivio statico diviene un sistema dinamico, capace di dialogare con il visitatore attraverso interfacce naturali, narrazioni generate in tempo reale ed esperienze inclusive (ETT & RAISE Liguria, 2025).

2. IL PROGETTO SINNOS: FRUIZIONE MULTIMODALE ED INTERATTIVA

Questa sezione presenta il progetto Sinnos (Smart INteractive eNvironments Of Sardinia) come caso di studio paradigmatico di valorizzazione territoriale. L'iniziativa, finanziata nel corso del biennio 2021-2022 dall'Assessorato al Turismo e svolto in compartecipazione con Sardegna Ricerche, nasce per ideare e progettare nuove metodologie di valorizzazione culturale. Si è quindi scelto di affrontare la sfida di trasformare la fruizione del patrimonio da passiva a pienamente esperienziale, utilizzando i ricchi contenuti culturali e turistici del portale regionale Sardegna Turismo come base informativa da rielaborare e valorizzare attraverso nuove modalità interattive. L'obiettivo è dare una nuova forma a questo archivio istituzionale, superando la tradizionale consultazione web per approdare a un ecosistema di digital storytelling diffuso. Attraverso l'integrazione di tecnologie di Human-Computer Interaction e Intelligenza Artificiale, i dati di SardegnaTurismo vengono decontestualizzati dai database e ricontestualizzati in ambienti fisici e ibridi, abilitando nuove modalità di heritage sharing che rendono il visitatore co-protagonista nella scoperta delle eccellenze dell'Isola.
Il cuore tecnologico del progetto Sinnos è costituito da un middleware proprietario, un'architettura software avanzata progettata per l'orchestrazione centralizzata e la distribuzione multicanale di contenuti multimediali eterogenei. Questa piattaforma agisce come un collettore intelligente dotato di funzionalità di ricerca semantica e indicizzazione, consentendo l'accesso immediato e strutturato a un vasto repository di asset digitali che include testi storici, produzioni video, archivi fotografici ad alta risoluzione e modelli 3D complessi. La peculiarità del sistema risiede nella sua capacità di manipolare e riadattare dinamicamente questi contenuti — attraverso servizi di content delivery e message brokering in tempo reale — per popolare un ecosistema diversificato di installazioni interattive (figura 1), garantendo una coerenza narrativa trasversale e fruita tramite diverse interfacce utente.



Figura 1. Laboratorio HCI per la valorizzazione culturale

Grazie a questa infrastruttura, i dati alimentano esperienze di fruizione tangibile come il tavolo interattivo, dove un sistema di computer vision opportunamente addestrato riconosce oggetti fisici posizionati dagli utenti (es. riproduzioni di bronzetti o manufatti tessili), attivando automaticamente sulla superficie touch (Iacolina S.A et al., 2020) video tematici e mappe georeferenziate correlate all'oggetto. Il middleware distribuisce inoltre i contenuti verso interfacce di Realtà Virtuale (VR), come nell'applicazione Sardegna Immersiva, che utilizza ricostruzioni fotogrammetriche di siti archeologici navigabili tramite visori o schermi touch. L'interazione naturale è supportata da sistemi di ricerca vocale, che permettono all'utente di interrogare il database geografico tramite comandi in linguaggio naturale (es. "mostrami i nuraghi"), filtrando i POI sulla mappa in tempo reale. Infine, la valorizzazione si estende al supporto cartaceo e mobile tramite la Brochure AR, dove algoritmi di deep learning riconoscono le pagine stampate sovrapponendo modelli 3D e video (es. la tessitura artigianale), e l'applicazione mobile Sardinia, pubblicata negli store Apple e Android, che trasforma lo smartphone in una lente di esplorazione tematica per il territorio. Nell'ambito del progetto, una specifica linea di ricerca è stata dedicata allo sviluppo di un sistema di Proximity Interaction, oggetto di domanda di brevetto (Iacolina S.A et al., 2025), ideato per rispondere alle stringenti esigenze igienico-sanitarie emerse nello scenario post-pandemico ed eliminare la necessità di contatto fisico con le superfici condivise (touchscreen pubblici). La soluzione, attualmente utilizzata in diversi contesti reali e spazi espositivi pubblici, trasforma il dispositivo mobile personale del visitatore in un'interfaccia di controllo remoto: attraverso la scansione di un codice QR o NFC, l'utente stabilisce una connessione sicura con il visualizzatore pubblico (un grande schermo o videowall), acquisendone il comando esclusivo. Il sistema integra un sofisticato gestore di attesa che, tramite l'emissione di token temporanei, regola gli accessi multipli e le sessioni di utilizzo, garantendo un'esperienza ordinata anche in contesti ad alto flusso. Una volta connesso, il visitatore può navigare agilmente tra gallerie di immagini, riprodurre video e interagire con modelli 3D complessi, manipolandoli (rotazione, zoom) sul grande schermo attraverso gesture eseguite sul proprio smartphone, che funge da touchpad virtuale.

3. DAL LABORATORIO AL MUSEO REALE: LA CASA MUSEO SALVATORE CAMBOSU

Il Museo Multimediale Salvatore Cambosu rappresenta l'esito tangibile di un percorso di ricerca in cui le tecnologie di Human-Computer Interaction, uscite dai laboratori del CRS4, hanno trovato piena applicazione in un contesto museale reale, inaugurato nel 2025 a Orotelli. L’intervento ha trasformato la dimora storica dello scrittore in un "archivio vivente", dove si espongono i contenuti rendendo interattivi gli spazi domestici. Riproponendo in chiave digitale la metafora del "favo", cara all'artista Maria Lai, l’archivio digitale viene inteso come struttura reticolare di celle interconnesse che custodiscono la memoria. Il percorso di visita è un continuum narrativo in cui ogni ambiente reagisce alla presenza del visitatore. Nella stanza del telaio, una videoinstallazione sincronizzata anima un telaio sospeso e attiva, tramite sensori in grado di riconoscere le interazioni con oggetti simbolici esposti, il racconto biografico interpretato da un attore su schermi verticali. La cucina recupera la dimensione domestica integrando fuochi virtuali e NFC che, posizionati sui fornelli d'epoca, permettono di scaricare video-ricette tradizionali, fondendo cultura materiale e fruizione digitale. Il fulcro tecnologico dell’allestimento è lo studio (figura 2), in cui la scrivania dello scrittore si trasforma in una sofisticata interfaccia tangibile.



Figura 2. L’interfaccia interattiva dello studio presso Casa Museo Cambosu

Grazie a un innovativo approccio di Interaction Design, che ha permesso di integrare sensori appositamente sviluppati direttamente negli arredi e di mantenerli invisibili al visitatore, l’interazione fisica con oggetti reali — come la macchina da scrivere o i quaderni — attiva narrazioni audiovisive dedicate sui monitor circostanti e proiezioni zenitali sulla superficie del tavolo. In questo modo la figura di Cambosu viene raccontata nelle sue diverse dimensioni: scrittore, giornalista e maestro. Il percorso prosegue nella Mediateca, dotata di touch screen e del sistema interattivo per l'esplorazione tematica dell'archivio multimediale, e nelle sale dedicate al rapporto con Maria Lai, dove sistemi di gamification permettono di ricomporre visivamente le fiabe dell'autrice. La dimensione onirica è protagonista nella camera da letto, trasformata in un ambiente immersivo dove proiezioni sincronizzate su pareti e tendaggi dissolvono la figura dell'attore in elementi naturali, evocando i sogni dello scrittore. Infine, la stanza del pane offre un'esperienza multisensoriale basata sul riconoscimento gestuale: proiezioni zenitali sul tavolo da lavoro mostrano le fasi della panificazione e invitano il visitatore a mimare i gesti dell'impasto per attivare la narrazione visiva e sonora del rito, chiudendo il cerchio tra gesto fisico e contenuto digitale.

4. BENI CULTURALI E AI

La metodologia adottata prevede lo sviluppo di sistemi multimodali che integrano modelli per l'analisi e la generazione di contenuti, superando la divisione tradizionale tra testo, immagine e audio. Negli ultimi anni il lavoro del team si è concentrato per integrare le tecnologie di fruizione culturale con i sistemi di Intelligenza Artificiale, utilizzati non più solo per classificare il patrimonio, ma per analizzare, creare informazioni e generare contenuti in nuovi formati. I materiali d'archivio sono stati processati massivamente attraverso pipeline automatiche per creare sintesi testuali, estrarre keyword e individuare elementi informativi complessi all'interno delle immagini, generando didascalie descrittive e metadati semantici di alta precisione. Tutto questo, unito all'adozione di innovativi sistemi di ricerca semantica e indicizzazione, ha permesso l'implementazione di architetture RAG (Retrieval-Augmented Generation) e di Agenti AI-Driven avanzati. Questi agenti sono in grado di organizzare chatbot conversazionali multilingua che superano la semplice risposta testuale, fornendo all'utente elementi multimediali contestuali, video, immagini storiche e modelli 3D in tempo reale. La capacità generativa si estende alla modalità di creazione di storyboard automatici e sistemi di digital storytelling esperienziale destinati a grandi schermi; soluzioni pensate per essere utilizzate su visualizzatori pubblici in contesti ad alto traffico come aeroporti, fiere ed esposizioni, dove la narrazione si modella in tempo reale sulla varietà dei profili di pubblico, offrendo contenuti coerenti, accessibili e personalizzati. La tecnologia di analisi e di generazione di contenuti è stata ingegnerizzata sfruttando LLM (Large Language Models) e modelli all'avanguardia, come DeepSeek, Gemma, Qwen, Llama, ottimizzati per operare su domini specifici (Meta, 2024). Parallelamente, sono stati creati sistemi RAG e processi di Fine-tuning di modelli vision pre-addestrati, utilizzando tecniche efficienti come LoRA - Low-Rank Adaptation (Stability AI, 2024) per produrre immagini e video stilisticamente coerenti sulla base di dataset di immagini storiche ad-hoc, garantendo

rigore filologico nella generazione visiva (Runway AI, 2024). L'intero ecosistema tecnologico del progetto si basa sul principio chiave di hardware affordability, ovvero la capacità di realizzare soluzioni avanzate, sostenibili e facilmente adottabili dagli enti culturali e dalle imprese del territorio. Dato che l'esecuzione di modelli con miliardi di parametri richiederebbe risorse computazionali ingenti, vengono impiegate metodologie di compressione ed efficientamento quali il pruning, la quantizzazione e la knowledge distillation. Queste tecniche permettono di trasformare Large Language Models (LLM) complessi in versioni altamente efficienti, riducendo l'occupazione di memoria operativa di circa il 75% pur mantenendo oltre il 95% delle capacità originali (DeepSeek-AI, 2024). Tale approccio consente di eseguire localmente pipeline di intelligenza artificiale avanzata su hardware consumer o workstation dotate di una singola scheda grafica. L'infrastruttura di addestramento e inferenza è costruita su due livelli complementari. Il primo step prevede l'utilizzo del cluster HPC del CRS4, dotato di nodi GPU NVIDIA A100, per l'analisi dei contenuti e l'estrapolazione delle informazioni tramite algoritmi di deep learning e modelli LLM operanti su contenuti eterogenei (immagini, video, testi, 3D) e per il fine-tuning e la modellazione multimodale. In un secondo step, l'infrastruttura software e i modelli, oggetto dell'addestramento, possono essere quindi mandati in esecuzione su hardware di costo più basso, come workstation con GPU di 12GB o 16GB oppure board dedicate, come Nvidia Jetson. Questa metodologia consente di sviluppare, ottimizzare e distribuire soluzioni AI avanzate mantenendo costi contenuti e massima scalabilità, rendendo l'intero sistema replicabile e sostenibile per musei, archivi, enti culturali e imprese.

5. ACCESSIBILITÀ, INTERFACCE TANGIBILI, GESTUALI E ADATTIVE

L'accessibilità è declinata in questo ecosistema attraverso tre direttrici complementari che mirano all'inclusione sensoriale e cognitiva. La prima direttrice riguarda l'accessibilità visiva tramite sistemi tangibili. È stato sviluppato un tavolo interattivo per l'esplorazione tattile gerarchica destinato a utenti non vedenti. Il sistema traccia in tempo reale la posizione delle mani su una mappa fisica 3D della Sardegna (Iacolina S.A et al., 2025): un'esplorazione a due mani attiva descrizioni generali del contesto geografico (audio "overview"), mentre l'uso di una singola mano su un punto specifico attiva informazioni di dettaglio ("zoom" informativo), replicando cognitivamente il processo di esplorazione visiva dal generale al particolare. La seconda direttrice affronta l'accessibilità uditiva attraverso la Lingua dei Segni (LIS). Oltre alle app che riconoscono i gesti LIS per navigare i contenuti, è in corso di implementazione un Meta-Human accessibile. Questo avatar digitale di qualità fotorealistica, in grado di operare in tempo reale, inserito in ambienti virtuali o nel metaverso, funge da interprete bidirezionale: è in grado di riconoscere i segni fatti dall'utente tramite webcam e tradurli in parlato/testo per gli udenti, e viceversa tradurre il parlato in animazioni LIS fluide, abbattendo le barriere comunicative. La terza e più innovativa direttrice riguarda l'accessibilità cognitiva o intellettiva. Sfruttando le capacità dei Large Language Models (LLM), è stato integrato un sistema inclusivo capace di adattare la complessità del contenuto in tempo reale. Questo assistente intelligente non si limita a fornire informazioni standard, ma riformula le risposte e le descrizioni museali basandosi sul profilo dell'utente o sulla specifica domanda posta. Il sistema può semplificare il lessico, utilizzare metafore più accessibili e ristrutturare le frasi adattandole alle caratteristiche intrinseche della domanda dall'utente, estrapolate in tempo reale dal sistema, garantendo che il patrimonio culturale sia comprensibile e fruibile anche da persone con disabilità cognitive o difficoltà di apprendimento, realizzando così una personalizzazione radicale dell'esperienza di visita.

6. CONCLUSIONE E LAVORI FUTURI

L'integrazione di HCI e AI nei beni culturali, come dimostrato dai progetti Sinnos e Casa Cambosu, evidenzia come la tecnologia possa amplificare il public engagement, rendendo il patrimonio uno spazio di narrazione attiva e inclusiva. I prototipi presentati validano un approccio che unisce rigore scientifico, sostenibilità hardware e fruizione interattiva. I lavori futuri si concentreranno sull'evoluzione degli agenti AI autonomi, affinché possano orchestrare narrazioni sempre più complesse e personalizzate in tempo reale, mantenendo l'essere umano al centro del processo di validazione e fruizione. Infine, verranno svolti test specifici sulle interfacce e sulle modalità di interazione sviluppate, con attività di valutazione condotte anche in contesti accademici e conferenze dedicate alle tecnologie interattive e all'accessibilità, al fine di consolidare l'efficacia delle soluzioni e garantirne la piena adozione nei contesti museali reali.

RINGRAZIAMENTI

Il progetto SINNOS, svolto in compartecipazione con l'Assessorato al Turismo, Artigianato e Commercio della Regione Autonoma della Sardegna e con l'agenzia regionale Sardegna Ricerche, è stato finanziato con fondi PO FESR Sardegna 2014/2020, linea d'azione 6.8.3a. Si ringrazia lo staff dell'Assessorato al Turismo, Artigianato e Commercio coinvolto nel progetto. Si ringrazia altresì la Fondazione Cambosu di Orotelli, il presidente Piero Marteddu, e il Comune di Orotelli, i sindaci Giovannino Marteddu e Tonino Bosu e lo staff, in particolare Ing. Giuseppe Usai e Ing. Salvatore Falchi. Per realizzare alcuni paragrafi di sintesi da contenuti, successivamente riadattati mediante intervento umano, sono stati utilizzati LLM addestrati ad-hoc sui contributi di ricerca.

BIBLIOGRAFIA

AI4Culture. (2024). Artificial Intelligence for the Cultural Heritage Sector. Europeana Foundation.

ASPHI, LIA, IBM, & IFAB. (2024). Limpide Letture: AI per l'accessibilità cognitiva. Fondazione ASPHI Onlus.

Cineca & Fondazione Scuola dei Beni e delle Attività Culturali. (2025). Nuove tecnologie e mediazione del patrimonio culturale: Il progetto I.PaC. Executive Summary.

Collettivo Digitale. (2024). TANGENT: Superfici tattili interattive per l'accessibilità museale.

DeepSeek-AI. (2024). DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model. arXiv preprint.

ETT & RAISE Liguria. (2025). Arte Tattile: Un'esperienza multisensoriale inclusiva.

Gartner. (2025). Hype Cycle for Artificial Intelligence 2025. Gartner Research.

Gartner. (2025). Top Strategic Technology Trends for 2025: Agentic AI and Generative AI. Gartner Research.

Harvard Art Museums. (2023). AI and Computer Vision for Automatic Cataloging. Harvard Art Museums Digital Lab.

Iacolina S.A., Paddeu G., Zanarini P., Casari C., Brevetto CRS4 (2020). System for detecting interactions with a surface. Domanda di Brevetto Europeo n. 20742856.6 / PCT/IB2020/054720. Deposito Italiano: 102019000007040

Iacolina S.A., Mura. F, Marotto M., Palla P., Angioni M., CRS4 (2025). Sistema di visualizzazione digitale di contenuti multimediali interattivi. Domanda di Brevetto Italiano n. 102025000004008.

Iacolina S.A., Mura. F, Palla P., Angioni M., CRS4 (2025). Sistema e relativo metodo per l'esplorazione tattile manuale o bimanuale di superfici o spazi tridimensionali da parte di un utente. Domanda di Brevetto Italiano n. 102025000003978.

McKinsey & Company. (2024). The Economic Potential of Generative AI: The Next Productivity Frontier.

Meta. (2024). Llama 3: Open Foundation and Instruct Models. Meta AI Research.

Mordor Intelligence. (2024). Digital Art Market Size & Share Analysis - Growth Trends & Forecasts (2024 - 2029).

Runway AI. (2024). Gen-3 Alpha: A New Frontier in Video Generation. Runway Research.

Stability AI. (2024). Stable Diffusion 3: Research Paper. Stability AI.

The Automation of Research in Digital Humanities: Artificial Researchers, Digital Archives, and the Democratization of Scholarly Access

ROBERTO DI QUIRICO
UNIVERSITÀ DEGLI STUDI DI CAGLIARI, DIPARTIMENTO DI INGEGNERIA MECCANICA, CHIMICA E DEI MATERIALI (DIMCM) – DIQUIRICO@UNICA.IT

ABSTRACT (ENGLISH)

Research automation via artificial researchers, defined as AI-driven agents capable of autonomous or semi-autonomous research activities, has shifted from experimental to operational status in high-return scientific domains during 2024-2025. The present study examines the implications for Digital Humanities and social sciences, with a particular focus on digital archives as pivotal sites where automated systems are poised to fundamentally reconfigure access and institutional relationships.
Based on exploratory discussions with senior staff at major European archival institutions and the author's participation in international research automation projects, this analysis identifies two primary integration models: external access via APIs that enable researcher-controlled systems, and internal artificial archivist agents that mediate user interactions. Both models offer significant public engagement benefits by reducing economic, skill-based, and temporal barriers to archival research, thereby expanding potential users beyond specialist academics to include citizen researchers, educators, and community groups.
Nevertheless, significant challenges constrain deployment, including technical limitations in humanities applications, infrastructural heterogeneity that impedes interoperability, and institutional tensions between access democratization and the preservation of symbolic capital invested in physical facilities. This paper contends that, despite these obstacles, archival institutions must engage in anticipatory planning by coordinating digital transformation, developing governance frameworks, and advocating for interoperability standards. Such proactive measures are necessary to shape, rather than merely react to, the transformations redefining scholarly research infrastructures.
Keywords: artificial researchers; research automation; digital archives; public engagement; Digital Humanities

ABSTRACT (ITALIANO)

L'automazione della ricerca nelle Digital Humanities: Ricercatori Artificiali, Archivi Digitali e la democratizzazione dell'accesso alla ricerca
L'automazione della ricerca attraverso ricercatori artificiali—agenti guidati dall'intelligenza artificiale capaci di svolgere attività di ricerca autonome o semi-autonome—è passata da uno stato sperimentale a operativo in ambiti scientifici ad alto ritorno economico nel 2024-2025. Questo contributo esamina le implicazioni per le Digital Humanities e le scienze sociali, concentrandosi sugli archivi digitali come luoghi critici in cui i sistemi automatizzati riconfigureranno radicalmente l'accesso e le relazioni istituzionali. Sulla base di colloqui esplorativi con dirigenti di importanti istituzioni archivistiche europee e del coinvolgimento dell'autore in progetti internazionali di automazione della ricerca, l'analisi identifica due modelli di integrazione: l'accesso esterno tramite API, che abilita sistemi controllati dai ricercatori, e gli agenti archivisti artificiali interni, che mediano le interazioni degli utenti. Entrambi promettono benefici sostanziali per il public engagement, riducendo le barriere economiche, competenziali e temporali alla ricerca archivistica e espandendo i potenziali utenti dagli accademici specialisti ai citizen researchers, agli educatori e ai gruppi comunitari. Tuttavia, sfide significative ne vincolano l'implementazione: limiti tecnici nelle applicazioni umanistiche, eterogeneità infrastrutturale che impedisce l'interoperabilità e tensioni istituzionali tra la democratizzazione dell'accesso e la preservazione del capitale simbolico investito nelle strutture fisiche. Il contributo sostiene che, nonostante queste sfide, le istituzioni archivistiche devono impegnarsi ora in una pianificazione anticipatoria—coordinando la trasformazione digitale, sviluppando quadri di governance e sostenendo standard di interoperabilità—per plasmare anziché reagire semplicemente alle trasformazioni che ridefiniscono le infrastrutture di ricerca.

Parole chiave: ricercatori artificiali; automazione della ricerca; archivi digitali; coinvolgimento pubblico; Digital Humanities

1. INTRODUCTION

In recent years, research automation has transitioned from speculative construct to operational reality. AI-driven agents capable of conducting autonomous or semi-autonomous research activities (hereinafter, Artificial Researchers, or AR), or AI platforms such as AlphaFold, are now ready for deployment across multiple scientific domains. These systems perform extensive literature reviews, identify research gaps, formulate hypotheses, design methodological approaches, and in some cases autonomously generate research outputs (Tang et al., 2025; Beel et al., 2025). The emergence of fully functional automated research platforms, particularly during 2024-2025, marks a critical threshold: research automation is no longer experimental but increasingly operational.

Recent contributions illustrate that Artificial Researchers do not constitute a single architectural model but rather a heterogeneous family of systems characterized by different degrees of autonomy, coordination mechanisms, and roles assigned to human researchers. The *AI co-scientist* framework conceptualizes the Artificial Researcher as a **human-in-the-loop epistemic collaborator**, based on a multi-agent architecture that emphasizes hypothesis generation, structured scientific debate, tournament-based selection, and iterative refinement under explicit human guidance (Gottweis et al., 2025). By contrast, the *ROBIN* system demonstrates a more **integrated and semi-autonomous research pipeline**, in which AI agents not only generate hypotheses but also interpret experimental data and iteratively update research directions within a lab-in-the-loop setting, significantly reducing human cognitive intervention while still relying on physical experimentation conducted by researchers (Ghareeb et al., 2025). Finally, the *AI-Researcher* framework advances the most radical configuration, proposing a **fully autonomous end-to-end Artificial Researcher** capable of literature review, idea generation, algorithmic implementation, experimental validation in silico, and publication-ready manuscript production, coupled with a dedicated benchmark (Scientist-Bench) to evaluate autonomous scientific performance (Tang et al., 2025). Taken together, these cases show that research automation is not converging toward a single model but toward a spectrum of Artificial Researchers whose architectures vary along key dimensions such as autonomy, modularity, validation strategies, and human oversight.

To date, these developments have concentrated in sectors characterized by high economic returns and advanced automation infrastructures. Materials engineering, pharmacology, and computational sciences—particularly machine learning—have witnessed the deployment of sophisticated platforms capable of coordinating robotic laboratories, managing complex experimental workflows, and autonomously generating scientific publications (Tang et al., 2025). In machine learning specifically, where research can be conducted entirely computationally, recent platforms demonstrate capabilities approaching full research autonomy (Beel et al., 2025; Lu et al., 2024).

This study situates these broader transformations within the context of Digital Humanities and social sciences, contending that automated research platforms and artificial researchers will soon become integral components of scholarly research infrastructures in these fields. As primary sources and archival materials increasingly originate in digital formats, the distinction between traditional humanities research and computationally mediated research is progressively diminishing (Chapinal-Heras & Díaz-Sánchez, 2023). Digital archives, in particular, constitute a critical domain where the integration of artificial researchers can fundamentally reconfigure access modalities, expand user demographics, and redefine institutional relationships among archives, researchers, and the broader public.

This contribution focuses specifically on archival contexts to examine how artificial researchers may integrate with existing infrastructures, the potential benefits—particularly in terms of public engagement—such integration offers, and the institutional, technical, and methodological challenges that persist. Importantly, this analysis underscores the necessity of anticipatory planning: archival institutions and research communities must begin preparing for these transformations now, even as the technologies remain partially nascent, to ensure that future digital reorganization aligns with both scholarly requirements and public access imperatives.

2. CONTEXT AND METHODOLOGY

Research automation operates along a continuum. At one extreme lie fully autonomous systems capable of independently defining research objectives, selecting methodologies, executing research pipelines, and

disseminating results (Tang et al., 2025). At the other extreme are directed automation models, in which human researchers define research questions or identify specific gaps in the existing literature, delegating execution to artificial research agents (Beel et al., 2025). Both models are operational, and both hold significant implications for humanities and social science research.

The empirical foundation for this analysis is based on the author's role as principal investigator of an international research consortium comprising four European universities. This consortium investigates research automation applications across multiple disciplinary domains, including those pertinent to Digital Humanities and archival sciences. The project provides an institutional context for examining how automated research systems may integrate with humanities research infrastructures.

This analysis also draws on exploratory discussions conducted in 2025 with senior archival staff at the Historical Archives of the European Union (HAEU) in Fiesole and the Historical Archives of the Bank of Italy in Rome. These conversations were informal in nature and not designed as a systematic empirical inquiry; they served rather as consultations aimed at gauging institutional perspectives on digital transformation and the prospect of automated research systems. Both institutions are currently transitioning toward predominantly digital archival collections, a shift that necessitates new computational approaches to access, management, and research support (Colavizza et al., 2022; Shinde et al., 2024).

The perspective developed in this paper should be situated transparently. The author does not write from within archival science or as a specialist in archival theory, but as a researcher whose competence lies in the study of research automation and artificial researchers across heterogeneous scientific domains. The contribution offered here is therefore deliberately external to the archival profession: it consists in importing into the archival debate a conceptual framework—centred on artificial researchers and their integration models—that has taken shape in adjacent scientific fields where research automation has progressed more rapidly. This positioning is consistent with the interdisciplinary character of Digital Humanities, which has historically advanced through the cross-pollination of methods and frameworks between domains. The aim is not to displace archival expertise, which remains indispensable for any meaningful integration of automated systems into archival practice, but to bring into the archival conversation a set of developments and analytical categories that, if left unaddressed, risk shaping archival infrastructures from the outside without institutional participation.

3. CURRENT STATE AND TECHNICAL CHALLENGES

Despite recent advances, significant technical, methodological, and institutional challenges constrain the deployment of artificial researchers in humanities and social science contexts. The platforms emerging in 2024-2025, while demonstrative of feasibility, remain concentrated in domains where automation is either technically simpler or economically prioritized (Tang et al., 2025; Lu et al., 2024).

Disciplinary differences profoundly shape the efficacy and efficiency of research automation. In experimental sciences with robotic laboratory integration, artificial researchers can execute standardized protocols with high reliability (Merchant et al., 2023). By contrast, humanities and social science research engages heterogeneous source materials, interpretive ambiguities, multilingual corpora, and complex historical contexts—conditions that substantially complicate automation (Colavizza et al., 2022).

Furthermore, the structure of services and infrastructures with which artificial researchers must interface critically determines system feasibility and reliability. Institutional repositories, publisher-controlled bibliographic databases, library systems, archives, and proprietary AI-based services each present different access protocols, metadata standards, query capabilities, and usage restrictions (Groppe et al., 2025; Nguyen et al., 2025).

These factors mean that artificial researchers capable of sophisticated operations in controlled computational domains may perform inadequately when confronted with archival complexity. Current AI systems, including large language models, demonstrate persistent limitations in sustained reasoning, factual accuracy, and handling of specialized domain knowledge—limitations that directly impact their utility in humanities research contexts (Beel et al., 2025; Marcus, 2024).

Consequently, although the trajectory toward automated research systems in the humanities is evident, substantial refinement of techniques, standardization of interfaces, and development of specialized capabilities remain necessary. The systems discussed herein should be regarded as emergent rather than mature, with significant development required before achieving the reliability and sophistication observed in more advanced automation domains.

4. INTERNAL AND EXTERNAL INTEGRATION OF ARTIFICIAL RESEARCHERS

Two primary integration models can be identified for artificial researchers within archival ecosystems, each with distinct technical architectures and institutional implications.

External Access Model: In this configuration, archives provide programmatic access—typically via Application Programming Interfaces (APIs)—enabling external automated research systems to query collections, retrieve metadata, and access digitized materials. Researchers interact with their own artificial research platforms, which in turn interface with archival infrastructures. This model requires archives to establish governance frameworks that regulate machine access, define query limits, specify permissible operations, and monitor usage patterns.

Internal Deployment Model: Here, archives develop and maintain internal artificial archivist systems—intelligent agents capable of receiving inputs from external users (whether human researchers or external automated systems) and executing complex operations including semantic search, contextual guidance, metadata enrichment, and research path recommendation (Rim et al., 2025; Groppe et al., 2025; Nguyen et al., 2025). These internal systems function as sophisticated intermediaries, translating user queries into archival-specific operations.

Both models shift archival access from expert-dependent to system-mediated processes. Traditional archival research demands considerable tacit knowledge, including understanding institutional organization, mastering finding aids, interpreting archival classification systems, and navigating physical repository structures. Artificial researchers and digital archivists can encode and operationalize this expertise, thereby making it accessible to a broader range of users. However, both models also raise important questions regarding trust, interpretive authority, and the transformation of archival expertise from human skill to algorithmic operation (Jaillant & Rees, 2023).

5. BENEFITS AND PROSPECTS FOR PUBLIC ENGAGEMENT

The integration of artificial researchers into archival systems carries profound implications for public engagement—a dimension increasingly central to both Digital Humanities practice and archival mission (Ridge, 2014; Owens, 2018).

Traditional archival research entails substantial financial barriers. A doctoral researcher conducting multi-site archival work across European institutions can easily incur substantial costs in travel, accommodation, and subsistence—often several thousand euros for a single multi-site research trip. Early-career scholars, independent researchers, and interested citizens often lack access to institutional funding mechanisms that enable such expenditures. These economic barriers directly constrain who can conduct archival research and what research questions can be feasibly pursued.

Artificial researcher-mediated access dramatically reduces these barriers. Remote interaction with digital archives via intelligent systems eliminates travel requirements, compresses research timelines, and reduces overall project costs to marginal expenses of connectivity and computational resources. This economic democratization expands potential researcher populations to include those previously excluded by resource constraints.

Beyond economic barriers, archival research requires specialized competencies: facility with finding aids, understanding of archival organization principles, knowledge of institutional histories, and often language skills for historical documents. These competencies constitute significant entry barriers for non-specialists.

Artificial archivist systems (in practice, an "internal artificial researcher") can mediate these requirements, guiding users through complex archival structures and translating general research interests into specific document-retrieval strategies. Academic and non-academic researchers can interact with archival systems that interpret their queries, propose feasible research directions, and facilitate document access—without requiring years of archival training.

Crucially, these systems can perform feasibility assessment: rather than requiring users to independently determine whether adequate sources exist for a given research question, artificial archivists can evaluate available holdings and propose research topics known to be documentarily supported. This guidance function reduces wasted effort and frustration, particularly for inexperienced researchers who might otherwise invest substantial time pursuing questions for which sources prove inadequate. Understanding collection organization, navigating finding aids, and physically accessing materials consume substantial

time—time that constitutes another form of resource barrier. Professional researchers can dedicate this time; many potential users cannot.
Automated systems compress research timelines by performing rapid searches across large collections, identifying relevant materials through semantic analysis rather than manual review, and presenting synthesized results rather than requiring users to independently process extensive metadata. What might require many days of on-site work can potentially be accomplished in a few hours of remote interaction. These combined effects—economic, competential, and temporal accessibility—fundamentally expand potential archival user populations. This demographic expansion aligns with contemporary emphases on public engagement, citizen science, and participatory research (McClure et al., 2020). By lowering entry barriers, artificial researchers enable archives to function as genuinely public research infrastructures rather than specialist-dominated institutions. The democratization of AI tools more broadly reinforces these possibilities, as users become more familiar with conversational interfaces and AI-mediated information access (van der Sommen & de Groof, 2024).

6. INSTITUTIONAL PERSPECTIVES AND CRITICAL CONSIDERATIONS

From archival institutional perspectives, artificial researcher integration presents both opportunities and tensions. Understanding these dynamics is essential for a realistic assessment of likely trajectories.
The exploratory discussions with senior staff at HAEU and the Historical Archives of the Bank of Italy conducted in 2025 suggest that research automation, while recognized as part of a broader digital transformation agenda, is not yet a prominent concern in the strategic planning of these institutions. This limited prominence appears to reflect two interrelated factors: ongoing regulatory uncertainty surrounding the use of AI in scholarly and cultural heritage infrastructures, and the absence of archival standards adapted to the opportunities and demands of automated research workflows. The institutional dialogue remains largely framed by the legacy agenda of digitization and metadata enrichment, rather than by the prospective integration of artificial researchers. The resulting gap between the trajectory of research automation in adjacent scientific domains and its current institutional visibility in the archival sector reinforces the case for anticipatory planning developed below.
Archives face mounting pressures regarding relevance, demonstrating impact, and justifying resources. Expanding user populations through artificial researcher-mediated access provides measurable impact metrics: increased access requests, broader demographic engagement, and enhanced public visibility. Furthermore, AI-mediated access can reduce operational costs associated with physical user accommodation—staff time devoted to user orientation, physical retrieval of materials, and facility maintenance for reading rooms. For institutions managing growing digital collections with constrained staff resources, automation offers operational efficiency.
However, the integration of artificial researchers also challenges traditional archival functions and professional identities. When intelligent systems mediate access, guide research, and facilitate retrieval, the archivist’s expert mediating role is called into question. Professional archival work has traditionally centered on appraisal, arrangement, description, and reference services—functions that artificial systems may partially automate or reconfigure. This situation raises concerns not only about professional displacement but also about knowledge preservation, as archival expertise encompasses accumulated tacit knowledge regarding institutional histories, collection contexts, and interpretive frameworks. Encoding this knowledge algorithmically is a crucial step in the creation and training of artificial researchers and archivists.
Additionally, archival institutions must preserve their symbolic capital. For many archives, the physical facility serves as a material representation of institutional authority and prestige. Researchers visiting the archive experience not only document access but also an institutional encounter, reinforcing the significance of both the archival materials and the institutions that preserve them. Many institutions have invested substantial resources in archival facilities, and if access increasingly occurs remotely via artificial researchers, these investments may appear wasteful, with planned expansions or renovations potentially losing justification. At the same time, institutions recognize that expanded digital access enhances visibility and impact. However, algorithmic mediation introduces reputational risks, and traditional archival control over consultation becomes less effective in automated systems. Institutions must therefore develop governance mechanisms that preserve necessary oversight while enabling the expanded access promised by artificial researchers. They must also address significant technological investment requirements, as developing robust artificial archivist systems demands expertise in natural

language processing, semantic search, user interface design, and ongoing system maintenance. Smaller institutions may lack the resources for such development, potentially resulting in a two-tier system where major archives offer sophisticated automated access while others remain reliant on traditional methods. Finally, questions of authenticity, certification, and evidentiary value arise when access is algorithmically mediated. Mechanisms must be established to verify that retrieved materials accurately represent archival holdings and to ensure that artificial systems do not introduce interpretive biases or omissions. These concerns necessitate the development of new authentication and transparency protocols.

7. ANTICIPATORY PLANNING AND DIGITAL TRANSFORMATION

A central argument of this paper concerns timing: archival institutions and research communities must begin planning for the integration of artificial researchers *now*, even as technologies remain nascent. Several considerations justify this anticipatory approach.

First, the trajectory is clear even if the timeline remains uncertain. Artificial researchers have demonstrated feasibility in adjacent domains; extension to the humanities and social sciences is likely, not speculative.

Second, meaningful integration necessitates coordinated digital transformation across multiple dimensions. Artificial researchers require machine-readable materials, making it imperative for institutions to prioritize digitization efforts and establish standards that ensure materials are not only digitized but also structured for computational access. Additionally, rich and standardized metadata is essential for semantic search and intelligent retrieval, requiring investment in metadata schema development, controlled vocabularies, and entity linking. External access models also demand the creation of robust, well-documented APIs, while archival professionals must acquire new competencies through targeted training programs. Furthermore, institutions must establish policies governing automated access, including the definition of permissible operations, protection of sensitive materials, and assurance of ethical use. The development of these frameworks requires consultation among stakeholders and institutional deliberation.

Third, anticipatory planning enables collective action toward interoperability standards. One of the most significant current obstacles to the adoption of artificial researchers is the heterogeneity and incompatibility among systems. Institutional repositories employ divergent metadata schemas; publishers' platforms, even when universities hold legitimate subscription-based access rights, frequently deploy technical barriers that obstruct automated querying; archival systems lack standardised APIs; and proprietary AI services operate within closed ecosystems. These fragmented infrastructures compel artificial researchers to navigate multiple protocols, limiting both efficiency and reliability.

8. CONCLUSIONS

This paper has argued that artificial researchers and automated research platforms represent a structural transformation rather than marginal innovation within Digital Humanities and social science research. The archival domain offers a particularly instructive case. As collections become predominantly digital-native, traditional access modalities become inadequate. Artificial researchers—operating both as external systems interfacing with archival APIs and as internal digital archivists mediating user interactions—promise to fundamentally reconfigure access, expanding user demographics and reducing economic, competential, and temporal barriers to participation.

These developments carry profound implications for public engagement. By dramatically lowering entry barriers, artificial researcher-mediated archival access enables genuinely democratized primary source research. Archives can transition from specialist-dominated institutions to public research infrastructures serving diverse communities—from professional academics to citizen historians, educators, journalists, and community groups.

However, realizing these opportunities requires confronting significant technical challenges, navigating institutional tensions, and undertaking coordinated digital transformation. Artificial researcher systems remain partially nascent, particularly for humanities applications. Substantial refinement of techniques, standardization of interfaces, and development of domain-specific capabilities are necessary. Archival institutions face strategic decisions that balance democratizing access with preserving scholarly rigor and professional expertise.

Most critically, institutions must engage in anticipatory planning now. The trajectory toward automated research systems is sufficiently clear to justify proactive preparation, even as specific timelines remain uncertain.

ACKNOWLEDGEMENTS

The author wishes to thank the senior staff of the Historical Archives of the European Union (HAEU) in Fiesole and of the Historical Archives of the Bank of Italy in Rome for their availability and for the insights shared during the exploratory discussions that informed this analysis.
This contribution has developed in the context of the ATHENA SEED research project (University of Cagliari), which investigates the application of research automation and artificial researchers across epistemologically heterogeneous scientific domains.
AI assistance (Claude, Anthropic) was employed for initial literature synthesis, structural organization, and drafting support. All content has been reviewed, verified, and substantially revised by the author. Empirical material derives from direct research activities and institutional consultations conducted by the author.

REFERENCES

Beel, J., Kan, M.-Y., & Baumgart, M. (2025). Evaluating Sakana's AI Scientist for autonomous research. arXiv preprint. https://doi.org/10.48550/arXiv.2502.14297

Chapinal-Heras, D., & Díaz-Sánchez, C. (2023). A review of AI applications in human sciences research. *Digital Applications in Archaeology and Cultural Heritage*, 30, e00288. https://doi.org/10.1016/j.daach.2023.e00288

Colavizza, G., Blanke, T., Jeurgens, C., & Noordegraaf, J. (2022). Archives and AI: An overview of current debates and future perspectives. *Journal on Computing and Cultural Heritage*, 15(1), Article 13. https://doi.org/10.1145/3479010

Ghareeb, A. E., Chang, B., Mitchener, L., Yiu, A., Szostkiewicz, C. J., Laurent, J. M., Razzak, M. T., White, A. D., Hinks, M. M., & Rodriques, S. G. (2025). *ROBIN: A multi-agent system for automating scientific discovery*. arXiv. https://arxiv.org/abs/2505.13400

Gottweis, J., Weng, W.-H., Daryin, A., Tu, T., Palepu, A., Sirkovic, P., Myaskovsky, A., Weissenberger, F., Rong, K., Tanno, R., Saab, K., Popovici, D., Blum, J., Zhang, F., Chou, K., Hassidim, A., Gokturk, B., Vahdat, A., Kohli, P., Matias, Y., Carroll, A., Kulkarni, K., Tomasev, N., Dhillon, V., Vaishnav, E. D., Lee, B., Costa, T. R. D., Penadés, J. R., Peltz, G., Xu, Y., Pawlosky, A., Karthikesalingam, A., & Natarajan, V. (2025). *Towards an AI co-scientist*. arXiv. https://arxiv.org/abs/2502.18864

Groppe, J., Marquet, A., Walz, A., & Groppe, S. (2025). Automated archival descriptions with federated intelligence of LLMs. arXiv preprint. https://arxiv.org/abs/2501.08955

Jaillant, L., & Rees, A. (2023). Applying AI to digital archives: Trust, collaboration, and shared professional practices. *Digital Scholarship in the Humanities*, 38(2), 626–648. https://doi.org/10.1093/llc/fqac087

Lu, C., Lu, C., Lange, R. T., Foerster, J., Clune, J., & Ha, D. (2024). The AI Scientist: Towards fully automated open-ended scientific discovery. arXiv preprint. https://arxiv.org/abs/2408.06292

Marcus, G. (2024). What does it mean when an AI fails? *Communications of the ACM*, 67(1), 42–45. https://doi.org/10.1145/3611097

McClure, E. C., Sievers, M., Brown, C. J., Buelow, C. A., Ditria, E. M., Hayes, M. A., Pearson, R. M., Tulloch, V. J. D., Unsworth, R. K. F., & Connolly, R. M. (2020). Artificial intelligence meets citizen science to supercharge ecological monitoring. *Patterns*, 1(7), 100109. https://doi.org/10.1016/j.patter.2020.100109

Merchant, A., Batzner, S., Schoenholz, S. S., Aykol, M., Cheon, G., & Cubuk, E. D. (2023). Scaling deep learning for materials discovery. *Nature*, 624, 80–85. https://doi.org/10.1038/s41586-023-06735-9

Nguyen, H. D., Nguyen, T.-H. A., & Nguyen, T. B. (2025). LLM-based smart search for archive systems: A new approach to information retrieval. arXiv preprint. https://arxiv.org/abs/2501.09876

Owens, T. (2018). *The Theory and Craft of Digital Preservation*. Johns Hopkins University Press.

Ridge, M. (2014). *Crowdsourcing Our Cultural Heritage*. Ashgate Publishing.

Rim, K., King, O. C., Lynch, K., Verhagen, M., & Pustejovsky, J. (2025). AI-assisted archival metadata generation and enrichment. *Lecture Notes in Computer Science*, 13918, 234–248. https://doi.org/10.1007/978-3-031-40839-2_17

Shinde, G., Kirstein, T., Ghosh, S., & Franks, P. C. (2024). Artificial intelligence in archival science: A systematic literature review. arXiv preprint. https://arxiv.org/abs/2411.01140

Tang, J., Xia, L., Li, Z., & Huang, C. (2025). AI-Researcher: Autonomous end-to-end scientific research through large language models. arXiv preprint. https://arxiv.org/abs/2501.02471

van der Sommen, F., & de Groof, J. (2024). Risks and rewards of AI democratization in healthcare and beyond. *United European Gastroenterology Journal*, 12(4), 427–428. https://doi.org/10.1002/ueg2.12563

[6/C] AI e automazione per archivi e patrimonio

Dal problema storico alle IA: Il Database on the Slave Trade between the Mediterranean and the Atlantic (15th-16th Centuries)

SALVATORE SPINA[1], CARLO TAVIANI[2], JÖRG HÖRNSCHEMEYER[3]
[1] UNIVERSITÀ DEGLI STUDI DI CATANIA, SALVATORE.SPINA@UNICT.IT
[2] UNIVERSITÀ DEGLI STUDI DI TERAMO, CTAVIANI@UNITE.IT
[3] ISTITUTO STORICO GERMANICO DI ROMA, HOERNSCHEMEYER@DHI-ROMA.IT

ABSTRACT (ITALIANO)

Il contributo presenta il Database on the Slave Trade between the Mediterranean and the Atlantic (15th–16th centuries), progetto PNRR sviluppato dalle Università di Teramo e Catania con l'Istituto Storico Germanico di Roma. L'infrastruttura digitale mira a raccogliere e modellare dati archivistici per ricostruire reti, attori e spazi della tratta schiavistica tra tardo Medioevo e prima Età Moderna, restituendo centralità al ruolo del Mediterraneo nella formazione dei circuiti atlantici. Basato sulla piattaforma open-source Arches e progettato secondo i principi FAIR, il database integra fonti eterogenee e sperimenta l'uso dei Large Language Models per l'estrazione automatica delle informazioni. Il paper riflette criticamente sulle potenzialità e sui limiti dell'IA —dal rischio di allucinazioni alla difficoltà nel riconoscimento semantico— sottolineando la necessità di procedure *human-in-the-loop*. Il progetto si configura così come un laboratorio metodologico per la Digital History, volto a rendere la conoscenza storica *machine-readable* e computabile.
Parole chiave: Archivi inventati; ChatGPT; Metodologia digitale; Mercato atlantico; Tratta degli schiavi.

ABSTRACT (ENGLISH)

This paper presents the *Database on the Slave Trade between the Mediterranean and the Atlantic (15th–16th centuries)*, a PNRR-funded project developed by the Universities of Teramo and Catania in collaboration with the German Historical Institute in Rome. The digital infrastructure aims to collect and model archival data to reconstruct networks, actors, and spatial dynamics of the slave trade between the late Middle Ages and the Early Modern period, restoring centrality to the Mediterranean in the creation of Atlantic circuits. Built on the open-source platform *Arches* and designed according to FAIR principles, the database integrates heterogeneous sources while experimenting with the use of Large Language Models for automated information extraction. The paper critically reflects on both the potential and the limitations of AI —from the risk of hallucinations to challenges in semantic recognition— emphasising the need for human-in-the-loop procedures. The project thus emerges as a methodological laboratory for Digital History, aimed at making historical knowledge machine-readable and computable.
Keywords: Invented Archives; ChatGPT; Digital Methodology; Atlantic Trade; Slave Trade.

1. INTRODUZIONE

Il paper guarda alla presentazione del *Database on the Slave Trade between the Mediterranean and the Atlantic (15th–16th centuries)*, sviluppato grazie al finanziamento, su bando PNRR, che ha visto coinvolte l'Università di Teramo e Catania. Il progetto —ancora in corso— ambisce a creare un database che raccolga la documentazione utile a spiegare il fenomeno della tratta degli schiavi, tra Mediterraneo e Atlantico, nei secoli XV e XVI.
Nel quadro teorico del progetto si tengono in considerazione le dinamiche che portarono al consolidamento del "sistema-mondo" atlantico, secondo la nota definizione di Fernand Braudel (Braudel, 2003, 2015): uno spazio che, tra Medioevo e prima età moderna, si strutturò progressivamente in modo interconnesso e secondo complesse dinamiche economiche. In tale contesto, la circolazione di metalli preziosi, merci e capitali finanziari si accompagnò in modo sistematico alla mobilità forzata e volontaria di individui liberi e schiavizzati, configurando assetti economici e sociali di lunga durata.
Tali flussi misero in connessione gli arcipelaghi atlantici, le coste occidentali dell'Africa e il bacino mediterraneo, contribuendo all'attivazione di precoci circuiti del capitalismo moderno (Braudel, 1977, 2006) Se la storiografia ha ampiamente indagato le dimensioni economiche, politiche e sociali dei secoli XVII e XVIII, le fasi antecedenti e immediatamente successive al viaggio di Cristoforo Colombo risultano

ancora parzialmente influenzate proprio da una lettura retrospettiva, centrata sull'età della piena espansione coloniale. Questa prospettiva ha privilegiato l'analisi di un sistema ormai strutturato, segnato dall'intreccio tra piantagione e schiavitù e dallo sfruttamento intensivo delle risorse —dai metalli alle fibre tessili, dai prodotti agricoli alle specie animali— relegando in secondo piano le dinamiche formative del fenomeno (Bono, 2017; Delpiano, 2021; Fiume, 2012, 2012; Traselli, 1972).
Ciò che tende a rimanere in secondo piano —e che il progetto PRIN PNRR intende ricollocare al centro dell'analisi—, è il ruolo strutturante del Mediterraneo, in particolare delle sue infrastrutture finanziarie e commerciali, nella genesi dei *network* atlantici.
Inteso come "continente liquido" —secondo la definizione di Fernand Braudel (Braudel, 2011, 2017b, 2017a) —, il Mediterraneo si innesta, infatti, nei processi di integrazione economica di *lunga durata* ben prima dell'apertura atlantica —ossia prima della piena affermazione dell'Età Moderna—, fornendo modelli, pratiche e capitali, che verranno, successivamente, proiettati oltre lo Stretto di Gibilterra.
In questa prospettiva, la figura dell'"esploratore" europeo attivo nell'Africa subsahariana e negli arcipelaghi atlantici tra la metà del XV e l'inizio del XVI secolo viene reinterpretata non come *agente* isolato del viaggio, ma come attore inserito in reti mercantili e finanziarie strutturate e preesistenti. A partire dagli anni Cinquanta del Quattrocento, si osserva, infatti, un mutamento significativo nelle direttrici del capitale europeo: dal tradizionale asse orientale, imperniato sul Levante e sul Mar Nero, si passa progressivamente a un orientamento occidentale che privilegia la penisola iberica, il Maghreb e gli arcipelaghi atlantici del continente africano.
In questo processo di ridefinizione dei circuiti economici, un ruolo centrale fu svolto dai mercanti-banchieri genovesi (Taviani, 2024), capaci di articolare una rete fitta di insediamenti, partnership e investimenti finanziari diffusi nello spazio iberico, negli arcipelaghi atlantici e nel Mezzogiorno italiano (Armenteros-Martinez, 2022; Bonazza, 2023; Gioffrè, 1962, 1962, 1971; Rau, 1957; Schwartz, 2004; Sicking & Wijffels, 2020; Trasselli, 1969; Verlinden, 1970). Arcipelaghi come Madera e le Canarie diventano aree di produzione specializzata, in particolare nella coltivazione dello zucchero, fondata sull'impiego intensivo di manodopera schiavile, mentre Capo Verde emerse precocemente come snodo strategico della tratta atlantica delle persone ridotte in schiavitù. Ciò che emerge dall'analisi di tali dinamiche è la formazione di un sistema economico integrato (il capitalismo atlantico della prima età moderna), fondato sulla circolazione coordinata di capitali, merci, persone ridotte in schiavitù e risorse naturali (piante e animali) tra il Mediterraneo, l'Africa e l'Atlantico. Tuttavia, la ricostruzione di tali processi risulta complessa a causa della natura frammentaria delle fonti disponibili —in particolare gli atti notarili— che di solito hanno registrato singole transazioni e sono dispersi in diversi archivi e in conseguenza del fatto che il primo periodo di tale processo, il tardo Quattrocento e il primo Cinquecento, non è stato ancora studiato approfonditamente.
Sebbene il processo di apertura dell'Atlantico sia stato oggetto di ampie indagini storiografiche (Benton, 2009; Curtin, 1975; Elliott, s.d.; Fuson, 1995; Games, 2008; Morelli, 2013; Thornton, s.d.), solo in tempi recenti la ricerca ha iniziato a ricostruire, in modo sistematico le connessioni tra mercati differenti e contesti sovranazionali, guardando alle origini più remote della tratta delle persone ridotte in schiavitù—in particolare quelle avvenute tra Medioevo e prima Età Moderna—, dimostrando come tali prospettive siano state scarsamente esplorate, soprattutto per quanto concerne le loro premesse finanziarie, logistiche e infrastrutturali. In tale prospettiva si inserisce il database, che consente di integrare e mettere in relazione dati dispersi, rendendo ricostruibili le connessioni tra attori, luoghi e operazioni economiche, e permettendo così di studiare tali dinamiche non più come fenomeni isolati, ma come un sistema coerente.

2. IL DIGITAL TURN: DAGLI ARCHIVI AL WEB

Chi sono i cosiddetti "esploratori"? A quali reti di mercanti sono riconducibili? Quale network si costruisce intorno alla tratta delle persone ridotte in schiavitù? Quali gli spazi e le direzioni?
La risposta a queste domande esige una attenta valutazione della documentazione archivistica e l'individuazione di informazioni dettagliate; un lavoro che necessariamente richiede un impegno non limitato nel tempo, e la possibilità di costruire un vero e proprio strumento di controllo. È in questa direzione che si colloca il progetto di costruzione del *Database on the Slave Trade between the Mediterranean and the Atlantic (15th–16th centuries).*
Il database viene sviluppato da unità di ricerca presso l'Università di Teramo (prof. Carlo Taviani), l'Università di Catania (prof.ssa Lina Scalisi) e l'Istituto Storico Germanico di Roma (dr. Jörg

Hörnschemeyer), ed è stato finanziato, nel 2022, su bando PNRR (Next Generation EU – missione 4, componente 2, investimento 1.1).
Il lavoro, della durata di due anni, ha guardato a due direttive fondamentali, ossia, da un lato, integrare in un'unica infrastruttura digitale una quantità significativa di dati provenienti da diversi fondi archivistici (Genova, Siracusa, etc.) e, da un altro lato, sperimentare l'impiego di tecnologie innovative per l'acquisizione, l'estrazione e la traduzione delle informazioni desunte dalle fonti storiche in record strutturati del database. Il vincolo temporale, imposto dalla realizzazione del progetto, ha reso necessario il ricorso a strumenti informatici in grado di operare in modo efficiente su ampi corpora documentari, orientando il gruppo di ricerca verso l'utilizzo dei Large Language Models, oggi accessibili attraverso diverse piattaforme di intelligenza artificiale generativa.
All'interno della 'cassetta degli strumenti', caratterizzante il "mestiere di storico" —per riprendere una celebre definizione di Marc Bloch (Bloch, 1963) —, i tools di Named Entity Recognition (NER) erano già da tempo entrati, mostrandosi come prospettiva metodologica digitale della ricerca —con soluzioni quali spaCy, Flair, i modelli Transformers, Stanza o OpenNLP—, contribuendo significativamente al rinnovamento degli approcci quantitativi allo studio delle fonti, pur richiedendo competenze informatiche specifiche per consentire un'interazione efficace tra algoritmi e documentazione storica.
Da alcuni anni, tuttavia, la svolta digitale ha intrapreso un percorso di "maturazione" con la diffusione di sistemi di IA in grado di elaborare il linguaggio naturale su larga scala. Queste tecnologie, che rappresentano una prima concretizzazione delle ambizioni originarie della Computer Science legate alla simulazione del ragionamento umano (McCarthy et al., 1955), , hanno aperto nuove prospettive operative e teoriche in ambito archivistico e storico, incidendo profondamente sia sul fronte "interno", quindi sulle modalità di costruzione e organizzazione dei dati dentro gli archivi digitali, che su quello della loro analisi e interrogazione. E la creazione di questo database ha dato l'opportunità, al gruppo di ricerca, di testare le potenzialità di queste piattaforme, comprenderne i meccanismi e gettare le basi per la creazione di un sistema di prompt che possa tradursi in momenti di Machine Learning per l'addestramento di una IA storica, rispondendo, in tal senso, alla necessità evidenziata da Richard Ennals di 'spiegare' «alla [macchina] cosa noi [storici] facciamo» (Ennals, 1985) .

3. IL MODELLO SEMANTICO

L'obiettivo della realizzazione di questo database è quello di adottare la modellazione semantica dei dati[1], la quale rivela i suoi punti di forza soprattutto nelle aree in cui l'intelligenza artificiale incontra limitazioni, come nella rappresentazione esplicita di significati, relazioni e contesti. Sotto forma di 'grafici della conoscenza', i dati modellati semanticamente consentono «an interpretable representation of knowledge, the derivation of logical inferences, and a long-term stable structuring and storage that is not dependent on statistical training» (Pan et al., 2023); quindi, nel nostro caso, esso offre la possibilità di ricostruire, dettagliatamente, il network degli attori storici che giocarono un ruolo fondamentale nel fenomeno della schiavitù.
Da un altro lato, questo sistema consente di creare una base di conoscenza 'strutturata' da cui gli stessi sistemi di IA possono trarre vantaggio, ad esempio: (1) nell'addestramento attraverso categorie e relazioni chiaramente esplicitate, (2) nella contestualizzazione attraverso l'analisi di informazioni rilevanti, (3) nell'interpretazione dei risultati, attraverso la classificazione e la convalida di affermazioni generate dall'IA, basate su relazioni modellate esplicitamente.
Il modello di dati, sviluppato nell'ambito del progetto —sulla base del CIDOC CRM—, deve essere inteso sia come un approccio documentale, ovverosia rispondere alla necessità di rendere le copie digitali *accessibili*, *descrivibili* e *verificabili* (adeguate alla struttura FAIR), e quanto più fedeli alle fonti originali (Documenti Legali); sia come un quadro metodologico, per modellare il loro contenuto (Atti Legali) in una forma astratta, comparabile e analizzabile (fig. 1).

[1] https://gemene-st.dhi-roma.it/documentation/

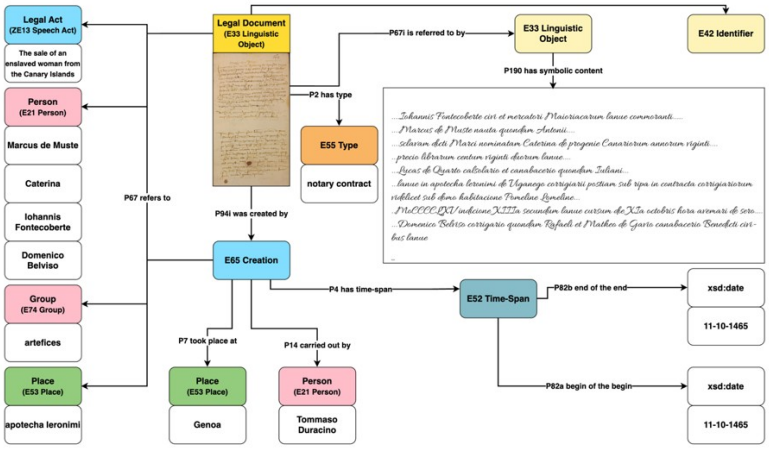


Figura 1. Esempio di atto notarile che descrive la vendita di una donna in schiavitù proveniente dalle Isole Canarie e le persone coinvolte in tale transazione.

Organizzando e memorizzando i dati sotto forma di "grafico della conoscenza", diventa possibile una interrogazione delle reti relazionali tra gli individui e i vari gruppi, come quelle create attraverso gli atti giuridici, tra i quali troviamo disposizioni testamentarie e atti di acquisto, vendita, donazione di beni e persone ridotte in schiavitù.
Nell'ambito del progetto, l'Istituto storico tedesco di Roma (DHI Rom) è responsabile, in qualità di partner infrastrutturale, dell'hosting sostenibile della piattaforma e di garantire l'accessibilità a lungo termine dei dati raccolti e memorizzati nel database.
Oltre all'uso di metodi basati sull'IA, per l'analisi delle fonti di ricerca, e del focus stesso della ricerca storica, il progetto dovrebbe essere inteso anche come "best practice case", che avvalora l'idea metodologica di come tali infrastrutture possano essere create con risorse finanziarie e umane relativamente limitate.
Il database si basa sulla piattaforma open source di gestione dei dati Arches[2] —piattaforma open-source sviluppata dal Getty Conservation Institute e dal World Monuments Fund[3]—, ed è stato sviluppato con il supporto di *knowledge engineering experts* dell'azienda Takin.solutions[4].
L'interfaccia web offre ampie funzionalità di ricerca, navigazione e visualizzazione, e consente il riutilizzo dei dati attraverso formati Linked Open Data standardizzati come JSON-LD. Tutti i vocabolari, le tassonomie e i thesauri utilizzati nel progetto sono pubblicati anche come file SKOS.
Il tutto viene concepito come 'progetto scalabile', che consente, da un lato, l'integrazione di set di dati aggiuntivi, e, da un altro lato, l'estensione e implementazione del modello semantico.
In futuro, per sfruttare al massimo i punti di forza sia delle tecnologie di IA, che dei *semantically modeled knowledge graphs*, miriamo a integrare i dati già acquisiti, attraverso un processo di «retrieval-augmented generation (RAG) workflow» (Oche et al., 2025), con l'ausilio di intelligenze artificiali, allo scopo di consentire ai ricercatori del progetto, nella futura raccolta dati, un lavoro di disambiguazione delle *entities* e dei termini, nonché la contestualizzazione spaziale, temporale e tematica delle informazioni.

4. MODELLI IA NELLA COSTRUZIONE DEL DATASET

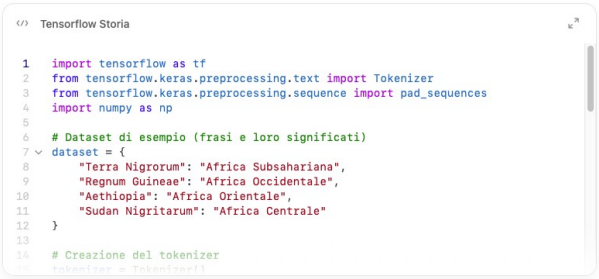
Se vero che, ancora, vi è una profonda "difficoltà", da parte delle IA, sul ragionamento storico, è ancor più vero che il lavoro di estrapolazione dei dati, con tali strumenti, è un processo dai risultati non sempre positivi. Il progetto, infatti, ha consentito al team di ricerca di valutare tutte le possibilità offerte dalle tecnologie IA ed individuare la piattaforma che, ad oggi, mostra di rispondere con più efficacia alla

[2] https://www.archesproject.org/
[3] L'istituto riunisce professionisti del patrimonio e sviluppatori di software provenienti da tutto il mondo, allo scopo di sviluppare un sistema che consenta di creare database in cui riunire dati che possano "spiegare" patrimoni culturali, inventarî digitali, e altre tipologie di prodotti, attraverso diversi sistemi integrati di visualizzazione dei dati, finanche la georeferenziazione delle informazioni.
[4] https://www.takinsolutions.com/

necessità di estrapolare dati storici dai documenti d'archivio, e la loro successiva organizzazione in un file CSV, per rispondere alle necessità dell'ontologia che sottostà al database.
Il problema iniziale sta nella necessità di evitare *outcomes* "allucinati" (Maleki et al., 2024; Salvagno et al., 2023). Quello delle risposte 'prive di senso', in ambito storico, raggiunge un livello di criticità sostanzialmente diverso rispetto a quello degli altri ambiti disciplinari. Quando portiamo l'IA a quel livello d'analisi che caratterizza la ricerca storica, il risultato che, spesso, si ottiene, è privo di validità. Ciò, sicuramente, è da addebitarsi alla inadeguata quantità di conoscenza storica su cui sono state allenate queste piattaforme. Dentro i grandi modelli, l'informazione storica "complessa" è pressoché nulla. I dati su cui vengono allenate le IA sono ricavati da enciclopedie online, e da qualche sito ad accesso aperto (senza la creazione di profili utente). Tutto quello che riguarda la grande letteratura e la Storiografia non è stato datificato ed utilizzato per il training —certamente, a causa di violazione di copyright sulle pubblicazioni, accesso negato ai file delle pubblicazioni digitali, e altri aspetti che impediscono all'IA di "leggere" testi e documenti storiografici. Le fonti archivistiche (su cui si fonda il lavoro di ricerca dello storico), per il loro 'logico' *status* cartaceo, non sono consultabili dalle IA; e anche i grandi progetti di digitalizzazione di materiale d'archivio non guardano ad una adeguata "traduzione" dei testi in formato *machine-readable*, e quindi non utilizzabile dalle IA per rispondere adeguatamente ai prompt che le vengono dati.
Tale fenomeno, durante il procedimento, si è palesato in più fasi, ma, in particolar modo, quando si chiedeva di effettuare NER su testi in cui le entità non avevano un nome o un toponimo specifico e (storicamente) riconoscibile. Ad esempio, su un testo latino, tratto da Carlo Taviani The Genoese Merchant Network and Gold (ea. 1450-1530) (2024, 206), di cui si riporta, qui, solamente la parte interessata («[…] in dicto loco Orani, in quo loco eum vendidit dictus Dominicus cuidam mauro speciario qui multum frequentat in *terra nigrorum* mercatorio modo; quam venditionem fecit dictus Dominicus mediante ipso teste, ex quo causam habuit videndi dictum papirru, ex quo etiam scit que testificatur. Interrogatus quotenis est, respondit etatis annorum triginta quinque vel circa»), ChatGPT 3.5 riesce nell'individuazione delle entità specifiche, ma palesa una "difficoltà" nell'individuazione di una entità nominata, la cui formulazione nel documento non è specificatamente toponomastica, ossia l'espressione «*terra nigrorum*», la quale si riferisce all'area tra il Sahara e l'Africa occidentale, probabile traduzione della nota formula araba *Bilād al-Sūdān* (بلاد السودان).
La mancata individuazione di questa essenziale informazione, sia per la valenza storica, che per la correttezza del network che deve restituire il dataset, ha obbligato il team a ragionare su un possibile training in grado di potenziare l'IA con le informazioni corrette.
Per operare efficacemente, sono stati scelti vari modelli, i quali sono stati scaricati in locale, attraverso la piattaforma Ollama, allo scopo di creare una IA "esperta" nel fenomeno storico della schiavitù. I test vengono effettuati su macchina MAC STUDIO, con chip Apple M2 Max, con sistema operativo macOS Sequoia 15.4.1, 12 Core, 64 GB di Ram, GPU integrata con supporto Metal 3.
Configurata la piattaforma Ollama[5], e la Chatbox AI[6], per l'interazione con gli LLMs, sono stati scaricati i modelli deepseek-v2, openai/hello, openchat, gemma, mistral, olmo2, deepseek-r1, llama3.1.
I modelli più performanti, dal punto di vista dei risultati, sono quelli più grandi (671b, 405b, etc.). Ma più grande è il modello, maggiori sono le prestazioni hardware richieste. Ad esempio, un prompt di NER, nel caso di un modello di 236bv (deepseek-v2), espleta il processo in circa 40 minuti. I tempi aumentano quando ci spostiamo su modelli ancora più grandi. Quelli più piccoli vengono gestiti senza difficoltà dall'hardware, ma i risultati che offrono sono meno dettagliati e i processi di estrazione delle *entities* non portano i risultati sperati. I modelli compongono liste di nomi non complete, e la maggior parte di loro non riesce a ricostruire correttamente il quadro delle relazioni famigliari, partendo dalle descrizioni dei legami presenti nel testo del documento.
L'unico modello in grado di soddisfare il prompt è stato il "deepseek-r1:671b", il quale, però, ha impiegato 57 minuti per comporre la lista e le relazioni famigliari.
Tutti i modelli, inoltre, in remoto, non sono in grado di organizzare i dati in un file EXCEL e consentire il download di quest'ultimo, per la costruzione dei record del database.
Questo quadro, così come descritto, si traduce in limiti reali nel lavoro dello studioso. Tempi lunghi di elaborazione e computazione non corretta creano un divario tra il mondo della Storia e quello delle

[5] https://ollama.com
[6] https://chatboxai.app/it-IT

tecnologie di IA, le quali, seppur promettenti nelle loro prospettive di impiego, non presentano ancora un addestramento adeguato nei due settori di ricerca.
Un ulteriore test è stato effettuato mediante costruzione di un codice Python, con piattaforma Jupyter Notebook[7], attraverso la quale è stato costruito un codice per l'estrazione delle entità nominate, che ha restituito, nel dettaglio, le varie entità, mostrando, però, un "bug" che necessitava un ulteriore lavoro (in termini di tempi di programmazione) di costruzione del codice: non veniva, infatti, rispettata la suddivisione dei vari documenti, ma il processo estrapolava tutte le *entities* presenti nella pagina fornita, creando una lista in cui queste erano elencate, senza divisione di sorta . A tutto questo, si aggiunga la necessità di configurazione costante delle API Key dei singoli modelli, per cui, pur valutando positivamente il risultato, non è stato implementato alcun codice per continuare il lavoro in questa direzione.
Il processo, invece, è totalmente diverso e realmente promettente, quando si passa alle versioni online delle IA, soprattutto nei modelli a pagamento (in abbonamento). Nel nostro caso, è stato scelto di utilizzare la piattaforma ChatGPT di OpenAI. Anche in questo caso, però, dopo diverse fasi di training e di controllo (*human-in-the-loop*), i prompt venivano seguiti da *outcomes* corretti, in cui le singole entità sono state tutte individuate e raggruppate.
Nel caso delle entità non esplicite («terra nigrorum»), sfruttando la funzionalità stessa di ChatGPT di creare codici per le varie piattaforme (come, ad esempio, per Python), è stato sviluppato un codice TensorFlow, al fine di potenziare il modello con i dati certi relativi ai significati delle espressioni toponomastiche non chiare, e su altre che sono state individuate in ulteriori documenti (fig. 2).


Tensorflow Storia

```
import tensorflow as tf
from tensorflow.keras.preprocessing.text import Tokenizer
from tensorflow.keras.preprocessing.sequence import pad_sequences
import numpy as np

# Dataset di esempio (frasi e loro significati)
dataset = {
    "Terra Nigrorum": "Africa Subsahariana",
    "Regnum Guineae": "Africa Occidentale",
    "Aethiopia": "Africa Orientale",
    "Sudan Nigritarum": "Africa Centrale"
}

# Creazione del tokenizer
```

Figura 2. Codice Tensorflow per addestramento

Successivamente, i dati sono stati organizzati in file EXCEL e CSV specifici —grazie ad un nuovo dettagliato prompt— per essere utilizzati per la strutturazione del sito internet, il *browsing* e il "riuso", secondo i principi FAIR.

5. PER UNA CONCLUSIONE

There's a lot of talk about AI…; attribuiamo grande valore metodologico all'IA, ma la Storia non ha, ancora, la possibilità di diventare una GPT specifica e "personalizzata" per il mestiere di storico, almeno fino a quando non si riuscirà a costruire una *meta*ontologia, o un dataset storicamente 'completo', su cui allenarla. "Pensare storicamente" significa *trasferire*, come affermava Richard Ennals, le competenze reali su cui si fonda il pensiero di ogni singolo storico e la sua chiave metodologica. Significa datificare, da un lato, il mestiere —pensiamo ad un digital twin (Banfi et al., 2023; Batty, 2018; Correia et al., 2023; El Saddik, 2018; Eramo et al., 2022; Fuller et al., 2020; Gabellone, 2022; Gräßler et al., 2023) — e, da un altro lato, la conoscenza storica, ossia rendere *machine-readable-computable* quello che la Storiografia ha realizzato sino ad oggi.
Quello della Digital History (DHy), infatti, è uno scenario metodologico che richiede una presa di posizione, la quale spinge al coinvolgimento della comunità degli studiosi. Il discorso *sul* digitale implica un rinnovamento del "discorso *sulla* Storia", che deve abbracciare una metodologia che fa delle IA degli strumenti realmente diversi rispetto a quelli che, fino ad ora, hanno caratterizzato il lavoro dello storico in ambiente informatico.
Questa prospettiva, se da un lato guarda all'IA quale *ally* capace di *ragionare* di Storia, da un altro lato, probabilmente, riuscirebbe a colmare quel divario "umano" che ha caratterizzato il fenomeno della schiavitù, nel tentativo di portare la Storia là dove le è stato impedito di arrivare.

[7] https://jupyter.org

BIBLIOGRAFIA

Armenteros-Martinez, I. (2022). Enslaved Africans in Medieval and Early Modern Iberia. In Oxford Research Encyclopedia of African History. https://doi.org/10.1093/acrefore/9780190277734.013.864

Banfi, F., Dellù, E., Stanga, C., Mandelli, A., Roncoroni, F., Sivilli, S., Pepe, G., & Cacudi, G. (2023). Representing intangible cultural heritage of humanity: From the deep abyss of the past to digital twin and XR of the Neanderthal man and Lamalunga cave (Altamura, Apulia). The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLVIII-M-2–2023, 171–181. 29th CIPA Symposium "Documenting, Understanding, Preserving Cultural Heritage. Humanities and Digital Technologies for Shaping the Future" - 25–30 June 2023, Florence, Italy. https://doi.org/10.5194/isprs-archives-XLVIII-M-2-2023-171-2023

Batty, M. (2018). Digital twins. Environment and Planning B: Urban Analytics and City Science, 45(5), 817–820. https://doi.org/10.1177/2399808318796416

Benton, L. (2009). A Search for Sovereignty: Law and Geography in European Empires, 1400–1900. Cambridge University Press.

Bloch, M. (con un contributo di Strayer, J. R.). (1963). The Historian's Craft: Ou metier d'historien (P. Putnam, Trad.). Knopf.

Bonazza, G. (2023). Slavery in the Mediterranean. In D. A. Pargas & J. Schiel (A c. Di), The Palgrave Handbook of Global Slavery throughout History (pp. 227–242). Springer International Publishing. https://doi.org/10.1007/978-3-031-13260-5_13

Bono, S. (2017). Schiavitù mediterranea. Nuova informazione bibliografica, 56(4), 675–698. https://doi.org/10.1448/87956

Braudel, F. (1977). La dinamica del capitalismo (G. Gemelli, Trad.). Il Mulino.

Braudel, F. (2003). Scritti sulla storia. Bompiani. http://www.bompiani.it/catalogo/scritti-sulla-storia-9788845254123

Braudel, F. (2006). Civiltà materiale, economia e capitalismo. Le strutture del quotidiano (C. Vivanti, Trad.). Einaudi.

Braudel, F. (2011). Civiltà e imperi del Mediterraneo nell'età di Filippo II. Mondadori.

Braudel, F. (2015). Storia, misura del mondo. Il Mulino.

Braudel, F. (2017a). Il Mediterraneo, lo spazio, la storia, gli uomini, le tradizioni. Bompiani.

Braudel, F. (2017b). La Méditerranée. FLAMMARION.

Correia, L. F. R., dos Santos Bartholo, R., Brufato, A. W., & Thomé Sanchez, E. C. (2023). Toward a Digital Twin for Cultural Heritage. In J. V. Carvalho, A. Abreu, P. Liberato, & A. Peña (A c. Di), Advances in Tourism, Technology and Systems: Selected Papers from ICOTTS 2022, Volume 1 (pp. 419–430). Springer Nature. https://doi.org/10.1007/978-981-99-0337-5_35

Curtin, P. D. (1975). Economic Change in Precolonial Africa: Senegambia in the Era of the Slave Trade/Supplementary Evidence. Univ of Wisconsin Pr.

Delpiano, P. (2021). La schiavitù in età moderna. Gius.Laterza & Figli Spa.

El Saddik, A. (2018). Digital Twins: The Convergence of Multimedia Technologies. IEEE MultiMedia, 25(2), 87–92. IEEE MultiMedia. https://doi.org/10.1109/MMUL.2018.023121167

Elliott, J. H. (s.d.). Imperi dell'Atlantico. America britannica e America spagnola, 1492-1830 (M. Magnani, Trad.). Einaudi.

Ennals, R. (1985). Artificial Intelligence: Applications to Logical Reasoning and Historical Research. Ellis Horwood.

Eramo, R., Bordeleau, F., Combemale, B., Brand, M. van den, Wimmer, M., & Wortmann, A. (2022). Conceptualizing Digital Twins. IEEE Software, 39(2), 39–46. IEEE Software. https://doi.org/10.1109/MS.2021.3130755

Fiume, G. (2012). Schiavitù mediterranee. Corsari, rinnegati e santi di età moderna. Bruno Mondadori.

Fuller, A., Fan, Z., Day, C., & Barlow, C. (2020). Digital Twin: Enabling Technologies, Challenges and Open Research. IEEE Access, 8, 108952–108971. IEEE Access. https://doi.org/10.1109/ACCESS.2020.2998358

Fuson, R. H. (1995). Legendary Islands of the Ocean Sea. Pineapple Pr Inc.

Gabellone, F. (2022). Digital Twin: A new perspective for cultural heritage management and fruition. Acta IMEKO, 11(1), Articolo 1. https://doi.org/10.21014/acta_imeko.v11i1.1085

Games, A. (2008). Atlantic History: Definitions, Challenges, and Opportunities. In J. C. Guarnieri & J. Davis, Teaching American History in a Global Context. Routledge.

Gioffrè, D. (1962). Il commercio d'importazione genovese alla luce dei registri del dazio, 1495-1537. In Studi in onore di Amintore Fanfani (Vol. 5, pp. 113–242). Giuffrè.

Gioffrè, D. (1971). Il mercato degli schiavi a Genova nel secolo XV. Bozzi.

Gräßler, I., Maier, G. W., Steffen, E., & Roesmann, D. (2023). The Digital Twin of Humans: An Interdisciplinary Concept of Digital Working Environments in Industry 4.0. Springer Nature.

Maleki, N., Padmanabhan, B., & Dutta, K. (2024). AI Hallucinations: A Misnomer Worth Clarifying. 2024 IEEE Conference on Artificial Intelligence (CAI), 133–138. https://doi.org/10.1109/CAI59869.2024.00033

McCarthy, J., Rochester, N., Minsky, M. L., & Shannon, C. E. (1955). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955.

Morelli, F. (2013). Il mondo atlantico. Una storia senza confini. Carocci.

Oche, Agada Joseph; Folashade, Ademola Glory; Ghosal, Tirthankar; Biswas, Arpan (2025): A Systematic Review of Key Retrieval-Augmented Generation (RAG) Systems: Progress, Gaps, and Future Directions. arXiv: https://doi.org/10.48550/arXiv.2507.18910

Pan, Jeff Z., Simon Razniewski, Jan-Christoph Kalo, Sneha Singhania, Jiaoyan Chen, Stefan Dietze, Hajira Jabeen, Janna Omeliyanenko, Wen Zhang, Matteo Lissandrini, Russa Biswas, Gerard de Melo, Angela Bonifati, Edlira Vakaj, Mauro Dragoni, and Damien Graux (2023): Large Language Models and Knowledge Graphs: Opportunities and Challenges. arXiv: https://doi.org/10.48550/arXiv.2308.06374

Rau, V. (1957). A family of Italian merchants in Portugal in the XVth century: The Lomellini. In Studi Armando Sapori (Vol. 1, pp. 715–726).

Salvagno, M., Taccone, F. S., & Gerli, A. G. (2023). Artificial intelligence hallucinations. Critical Care, 27(1), 180. https://doi.org/10.1186/s13054-023-04473-y

Schwartz, S. B. (2004). Tropical Babylons: Sugar and the Making of the Atlantic World, 1450-1680. Univ of North Carolina Press.

Sicking, L., & Wijffels, A. (2020). Introduction Flotsam and Jetsam in the Historiography of Maritime Trade and Conflicts. In Conflict Management in the Mediterranean and the Atlantic, 1000-1800: Actors, Institutions and Strategies of Dispute Settlement. Brill Nijhoff. https://brill.com/display/title/39048

Taviani, C. (2024). The Genoese Merchant Network and Gold (ea.1450-1530). In T. B. F. Cummins (A c. Di), Global Gold: Aesthetics, Material Desires, Economies in the Late Medieval and Early Modern World. Villa I Tatti.

Thornton, J. (s.d.). L'Africa e gli africani nella formazione del mondo atlantico. 1400-1800 (L. Cobbe, Trad.). Il Mulino.

Traselli, C. (1972). Considerazioni sulla schiavitù in Sicilia alla fine del Medioevo. Clio, 8, 67–90.

Trasselli, C. (1969). Genovesi in Sicilia. In Atti della Società Ligure di Storia Patria, nuova serie (No. 9/2; Fascicolo 9/2, pp. 153–178). Società Ligure di Storia Patria. https://www.storiapatriagenova.it/Scheda_vs_info.aspx?Id_Scheda_Bibliografica=711

Verlinden, C. (1970). The Beginnings of Modern Colonization;: Eleven Essays. Cornell Univ Pr.

Peirce Interprets Peirce: Digitization, Automation, and Interpretation in Charles Peirce's Manuscripts

ALESSANDRO ADAMOU[1], SEBASTIAN FEIL[2], DAVIDE PICCA[3], CARLO TEO PEDRETTI[4], DARIO RODIGHIERO[5],

LORENZO ZANGARI[6]

[1] BIBLIOTHECA HERTZIANA – MAX PLANCK INSTITUTE FOR ART HISTORY, ITALY – ALESSANDRO.ADAMOU@BIBLHERTZ.IT
[2] RWTH AACHEN UNIVERSITY, GERMANY – SEBASTIAN.FEIL@IFAAR.RWTH-AACHEN.DE
[3] UNIVERSITY OF LAUSANNE, SWITZERLAND – DAVIDE.PICCA@UNIL.CH
[4] BIBLIOTHECA HERTZIANA – MAX PLANCK INSTITUTE FOR ART HISTORY, ITALY – CARLO.PEDRETTI@BIBLHERTZ.IT
[5] UNIVERSITY OF GRONINGEN, NETHERLANDS – D.RODIGHIERO@RUG.NL
[6] UNIVERSITY OF LAUSANNE, SWITZERLAND – LORENZO.ZANGARI@UNIL.CH

ABSTRACT (ENGLISH)

This paper presents a workflow-oriented approach to interpreting Charles S. Peirce's manuscripts through computational methods. Building on recent high-resolution digitization of the Peirce archive, it combines probabilistic transcription, text analysis, semantic modeling, and data visualization to examine how automation reshapes archival access and interpretive practice. Rather than treating computation as a means of producing static objects, the workflow frames digital archives as mediating research infrastructures that expand the interpretive space while preserving scholarly judgment. By approaching Peirce's manuscripts as multimodal objects—integrating text, diagrams, and revisions—the paper reflects on philosophical patterns, continuities, and tensions that become visible only at corpus scale. The contribution situates itself within current Digital Humanities debates on openness, mediation, and interpretive accountability, and argues that digitization and automation function as enabling conditions for shared, reflexive interpretation.
Keywords: Charles S. Peirce; Manuscript Studies; Digital Hermeneutics; Semiotics; Semantic Web

ABSTRACT (ITALIANO)

Questo articolo propone un approccio basato su workflow per interpretare i manoscritti di Charles S. Peirce con metodi computazionali. Partendo dalla recente digitalizzazione ad alta risoluzione dell'archivio, il lavoro integra trascrizione automatica, linguistica computazionale, modellazione semantica e visualizzazione dei dati per indagare come l'automazione stia trasformando l'accesso agli archivi e le pratiche interpretative. Invece di trattare i prodotti di processi computazionali come entità statiche, l'approccio qui presentato concepisce gli archivi digitali come infrastrutture di mediazione che ampliano lo spazio interpretativo senza sostituire il giudizio dell'esperto di dominio. Considerando i manoscritti di Peirce come oggetti multimodali, dove testo, diagrammi e revisioni convivono, l'articolo propone un flusso che palesi motivi filosofici, continuità e tensioni visibili solo su larga scala. Il contributo dialoga con il dibattito attuale nelle Digital Humanities su apertura, mediazione e responsabilità interpretativa, argomentando che digitalizzazione e automazione rappresentano condizioni abilitanti per una pratica ermeneutica condivisa e riflessiva.
Parole chiave: Charles S. Peirce; studi sui manoscritti; Ermeneutica Digitale; semiotica; Web Semantico

1. INTRODUCTION

Charles S. Peirce's philosophy has exerted a lasting influence across a wide range of disciplines, underscoring his foundational role in American pragmatism and scientific methodology (Atkin 2013; Sheriff 1989; Liszka 2021; Mazur and Sticksel 2021; Cristalli 2022; Strand 2022; Burch 2021; Zhang 2019; Jariah et al. 2022). While Peirce's published writings have been extensively studied, a large body of his unpublished manuscripts, vital for a complete picture of his thought, remains comparatively underexplored.

The comparatively limited engagement with Peirce's manuscripts is not due to a lack of scholarly interest but rather to long-standing challenges related to organization and access. Foundational efforts such as Richard Robin's catalog and the work of the Peirce Edition Project and the Institute for Studies in Pragmaticism provided essential tools for navigating such an extensive corpus (Robin 1967; Peirce

Edition Project 2021). These initiatives—together with later work, among others (e.g., Bella et al. 2016; Keeler 2020b; Keeler and Kloesel 1997)—established a scholarly infrastructure that remains indispensable. Given the sheer volume of Peirce's manuscripts, however, these approaches were somewhat constrained by publishing technologies and formats that were not designed to support large-scale comparison or exploratory analysis.

An important step toward broader access was John Deely's 1994 digitization of the eight-volume Collected Papers of Charles S. Peirce, which enabled full-text search and digital consultation of Peirce's published corpus (Peirce 1994). Despite its significance, this effort inherited the editorial and structural limitations of the Collected Papers, including the non-chronological and selective thematic organization of Peirce's writing, and media constraints resulting from the fact that in its original edition as a CD-ROM for the MS-DOS operating system, the edition omitted aspects of the visual content (including the use of color) found in the manuscripts. As a result, the content of manuscripts—many of which contain diagrams, revisions, marginalia, and color-coded reasoning—remains only partially accessible.

Recent developments in technologies open up new possibilities for engaging with Peirce's manuscripts. High-resolution re-digitizations of original documents address the material issues related to deteriorating paper and insufficient earlier scans, particularly with respect to diagrammatic and chromatic features (Everett 2021; Harvard University, n.d.; Keeler 2020a). At the same time, advances in Natural Language Processing (NLP) allow Peirce's corpus to be studied at a scale and level of complexity previously unattainable through automated transcription, annotation, and Large Language Models (LLMs).

Building on the aforementioned developments, this paper proposes a workflow-oriented approach to Peirce's manuscript corpus. By combining machine-assisted transcription, semantic structuring, data visualization, and the ability of LLMs to generalize across heterogeneous tasks, the text reflects on how current technologies reshape archival workflows and open new interpretive perspectives on Peirce's work within the framework of a four-year funded research project. In doing so, the paper situates itself within current Digital Humanities debates on openness, mediation, and interpretive accountability, and argues that digitization and automation should be understood not as endpoints, but as conditions for the enhancement of thorough philosophical inquiry.

2. BACKGROUND

Scholarship on Charles S. Peirce has long emphasized the centrality of his semiotic theory, particularly his Universal Categories and triadic sign relations—icon, index, and symbol—as foundational to his philosophical and scientific thought (Peirce 1955; Peirce 1960, 1977). Extensive work in semiotics and philosophy has explored these concepts as key to understanding perception, reasoning, and meaning-making processes (Pape 1990, 2015; Liszka 1996; Nöth 2016; Sonesson 2013; Stjernfelt 2007; Farias and Queiroz 2017; Feil 2024). Within this literature, Peirce's emphasis on diagrammatic reasoning—including diagrams and existential graphs—has been recognized as a distinctive feature that foregrounds the material and visual dimensions of thought.

In addition, several recent studies have addressed the multimodal character of Peirce's work, highlighting the analytical relevance of diagrams, graphical structures, and non-linear reasoning (Roberts 2009; Engel et al. 2012; Bateman 2018; Pflaeging et al. 2021). Research in semiotics and Digital Humanities has further shown how diagrams function as complex semiotic resources, enabling forms of reasoning that cannot be fully captured through linear text alone (Hiippala and Bateman 2022; Bellucci and Pietarinen 2024). Despite these advances, most analyses remain grounded in close reading or theoretical interpretation, with comparatively limited application of computational methods to large manuscript corpora.

In parallel, Digital Humanities has developed a wide range of computational approaches for the analysis of historical corpora, demonstrating their capacity to identify stylistic patterns and conceptual recurrences across large collections (Jockers 2013; Underwood 2019). Yet approaches that combine such methods with an explicit semiotic framework remain uncommon, particularly when applied to heterogeneous manuscript corpora that integrate textual, graphical, and material features (Ciula et al. 2023). In the case of Peirce, large-scale computational studies that engage directly with both his semiotic theory and his manuscript corpus remain scarce, as existing work tends to privilege either theoretical interpretation without computational mediation or computational analysis without sustained engagement with Peirce's conceptual apparatus.

Taken together, the state of the art points to a structural gap. While Peirce's semiotic theory and diagrammatic reasoning have been extensively studied, and while computational methods in Digital Humanities have matured, few approaches integrate these strands within a coherent methodological workflow. Addressing this gap calls for frameworks that treat digitization, modeling, and visualization not as endpoints, but as enabling conditions for interpretive openness, capable of supporting relational inquiry across complex and multimodal manuscript corpora.

3. METHODS

This section presents the project workflow as a methodological framework for engaging with large, heterogeneous manuscript corpora. Rather than a sequence of technical steps, the workflow is understood as a set of research decisions that shape how Peirce's manuscripts become legible, comparable, and interpretable. Digitization, automation, and visualization are treated as mediating practices that condition inquiry (see Fig. 1). The workflow builds on the recent large-scale digitization of Peirce's manuscripts at Harvard's Houghton Library, accessible via HOLLIS and IIIF services, providing a high-resolution corpus anchored in archival sources.

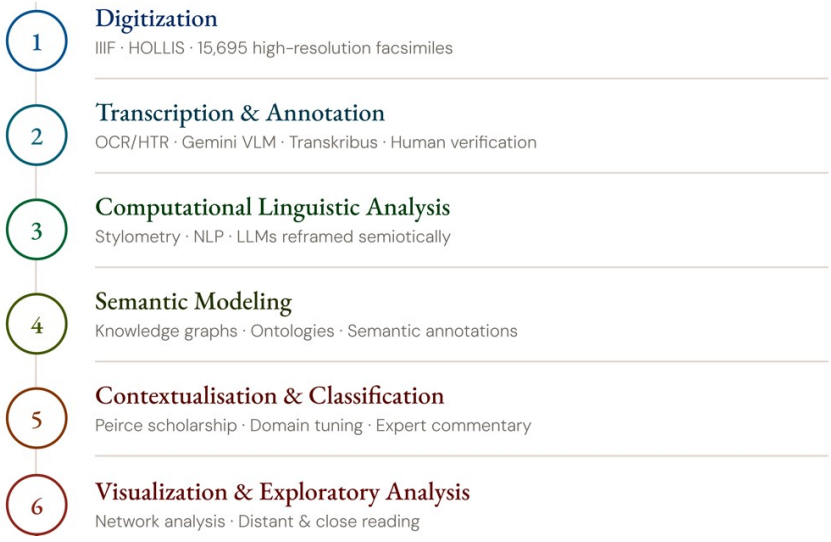


Figure 2. The project workflow in six stages, moving from digitization to visualization and exploratory analysis.

Automatic transcription and annotation. Transcription and annotation are approached as interpretive processes rather than neutral conversions. The recognition of characters (OCR) and of handwritten text (HTR) are treated as probabilistic techniques producing provisional texts that require human verification, particularly given the presence of revisions, marginalia, diagrams, and heterogeneous writing practices. Annotation and metadata creation actively structure the corpus, shaping how manuscripts can be queried and compared. Exploratory analyses at this stage focus on semantic and thematic regularities, while acknowledging the evolving nature of the data.

Computational linguistic analysis. Computational linguistic methods are employed as heuristic instruments for exploring Peirce's writings at scale. Building on earlier stylometric work, the workflow combines frequency analysis, syntactic parsing, and machine-learning–based representations to surface recurring concepts. These methods are not used for exhaustive classification, but to generate hypotheses and points of attention. Computational abstractions are explicitly aligned with Peirce's semiotic framework, ensuring continuity with concepts such as representamina and sign relations. Furthermore, we aim to reframe LLMs from a semiotic perspective, shifting the focus from performance to underlying sign processes. This enables a two-way analysis: examining how such models engage with Peirce's philosophical vocabulary, and exploring how Peirce's triadic framework can inform their generative capabilities.

Semantic modeling and knowledge representation. Semantic modeling structures emerging insights through relational representations rather than fixed taxonomies. Knowledge graphs are used to articulate connections among concepts, manuscripts, and semiotic categories. Ontologies will be adapted to

accommodate Peirce's distinctions while remaining open to revision. Semantic annotations link textual and graphical elements across the corpus, supporting traversal and comparison among them.
Contextualization and classification. Machine-generated output from all stages of the project is contextualized and assessed relative to the existing body of published material and the current state and requirements of Peirce scholarship. This entails the tuning of machine-generated output according to established knowledge about major Peircean themes, as well as the preliminary commentary on established knowledge within the Peirce community according to material made available by and findings generated in the project.
Visualization and exploratory analysis. Visualization is treated as an exploratory interpretive practice rather than an explanatory endpoint. Techniques from cultural analytics and network analysis provide orientation within the corpus, making patterns and anomalies perceptible while preserving access to manuscript sources. Visualizations support movement between distant and close reading, with design choices emphasizing transparency, legibility, and traceability. In this sense, visualization complements textual and semantic analysis as a mediating scholarly tool, without claiming interpretive authority.

4. PRELIMINARY RESULTS

Six months into the project, the exploratory work conducted demonstrates the feasibility and scholarly value of integrating semiotic theory with computational analysis for the study of Peirce's manuscripts. An earlier pilot stylometric study based on a selected subset of manuscripts transcribed using Transkribus showed that automated methods can reliably surface recurring conceptual and rhetorical patterns across heterogeneous documents (Kahle et al. 2017). In particular, the analysis highlighted Peirce's recurrent use of concepts such as "habit" and characteristic configurations of abductive reasoning, while maintaining a low transcription error rate (Picca et al. 2023). These results indicate that probabilistic transcription and text-based modeling can support large-scale exploration of Peirce's corpus while remaining sensitive to his philosophical vocabulary and argumentative structures.
Building on this pilot, the full set of Peirce manuscripts digitized by Harvard's Houghton Library has since been processed through the project transcription pipeline. From the IIIF Manifests of 233 manuscript items made available via the Harvard HOLLIS system, 15,695 high-resolution facsimiles were retrieved, after automatically excluding blank pages.
A targeted evaluation was conducted on approximately 200 manuscript pages sampled across five Peircean manuscripts of different periods and content types. On a shared 20-page test set, Gemini was compared with a Transkribus PyLaia model fine-tuned on 80 manually transcribed pages of the same hand. Gemini reached 2.36% CER and 4.41% WER (strict), while the fine-tuned PyLaia model reached 4.98% CER and 13.95% WER. The gap between character-level and word-level accuracy for PyLaia indicates that even a dedicated fine-tuned HTR system produces locally plausible but lexically unstable output on Peirce's handwriting, whereas the VLM maintains consistency at word level. These results position zero-shot VLMs as a reasonable production route for large-scale transcription of the archive, while preserving the need for philological verification on contested passages.
The manuscript pages were then transcribed zero-shot with Google Gemini Flash 3. In a post-processing phase, an inline diplomatic markup scheme for TEI P5 has been employed, with the goal of preserving authorial revisions (deletions, insertions, substitutions). The resulting machine-generated corpus now provides a homogeneous basis for corpus-scale analyses across Peirce's archive.
Subsequent studies extended this line of inquiry to the visual and multimodal dimensions of the manuscripts. Computational pipelines developed within the project enabled the extraction and evaluation of visual features from Peirce's diagrams, with a particular focus on Existential Graphs (Pedretti et al. 2025, n.d.). Using Vision–Language Models, this work assessed the capacity of contemporary systems to recognize diagrammatic forms and visual regularities.
Peirce claimed that "all necessary reasoning is diagrammatic" (CP 5.162) and described his visual logic system, the Existential Graphs, as "moving pictures of thought" (CP 4.8). His manuscripts embody this conviction in pages where text and visual elements intertwine, employing as many as four colors to make critical logical distinctions (Keeler 2020a). The manuscript page is a single semiotic surface that conventional printed editions have rendered only partially (Keeler 2020b). On the full digitized corpus, a CLIP-based classifier trained on 1,317 manually annotated pages identified 1,519 pages (9.7%) containing visual elements, concentrated in the Logic and Mathematics categories; the density peaks in

the 1895–1899 period, with 50.6% of Logic pages and 92.3% of Mathematics pages containing diagrams, in correspondence with Peirce's intensive development of the Existential Graphs (Pedretti et al. 2025). Machine-generated outputs (transcriptions, diagram detections, classifications) are anchored on the IIIF canvas of each manuscript page through the Web Annotation Data Model. Each bounding box produced by YOLOv8m layout segmentation becomes an annotation linking a canvas region to a semantic class; 1,139 diagram and 874 text-block boxes were annotated across 443 pages for training (see Fig. 2). The IIIF Manifest thus becomes an editorial surface where diagrams, logical notation, sketches, and other visual elements sit alongside the transcription, without altering the underlying facsimile.

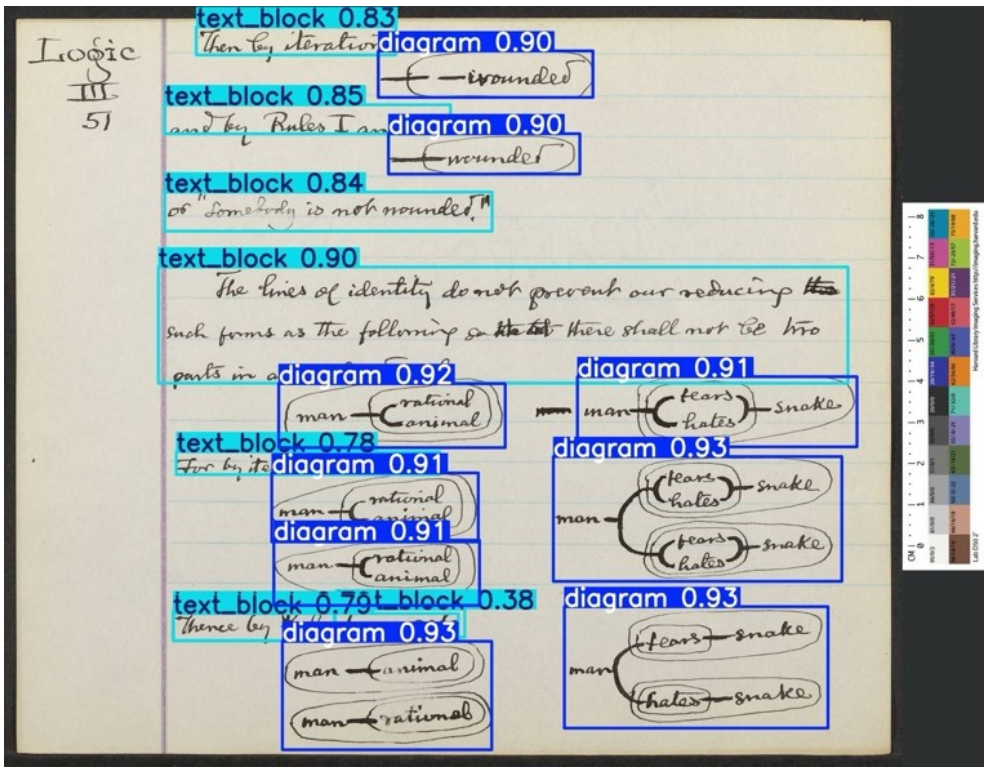


Figure 3. Fine-tuned YOLOv8m output showing diagram (blue) and text block (light blue) detection on MS 431.

Moreover, testing five Vision–Language Models on 27 manually annotated Existential Graphs covering Alpha and Beta variants, the results show strong performance at the morphological level (counting cuts, lines, spots) and at the relational level (containment, connection), but systematic failure at the representational level, where the diagram must be read as a logical formula (Pedretti et al., n.d.). Even when prompts stated Peirce's endoporeutic (from the outside to the inside) reading procedure, the models produced errors of scope and negation, reading "Some man is not wounded" as "No man is wounded" or collapsing ∃y∀x into ∀x∃y.

This empirical gap is further clarified by theoretical work examining the relationship between semiotics and large language models (Picca and Zangari, n.d.). That analysis situates contemporary AI systems within a semiotic perspective, showing how their pattern-based operations intersect with, but remain distinct from, Peirce’s account of sign processes and reasoning. Read together, the results suggest that computational systems can effectively support the organization, comparison, and exploration of Peirce’s manuscripts, while interpretive reasoning remains anchored in semiotic analysis and philosophical judgment.

A first public prototype of the resulting digital edition[1] has been developed using TEI Publisher v10 as a platform. Each manuscript page is rendered alongside its IIIF facsimile, with authorial revisions resolved interactively from the underlying TEI markup (see Fig. 3). The prototype serves a double purpose: as a scholarly access point to the transcribed corpus, and as a testbed for further layers—e.g. critical apparatus and ontology-driven annotations linking textual and diagrammatic elements—that will be integrated in the subsequent phases of the project. Moreover, by making the edition available at this

[1] The digital edition is available at https://peirce.humanitiesconnect.net.

intermediate stage the transcribed material is exposed as an object of scholarly verification rather than as a black-box output, allowing domain experts to inspect, annotate, and eventually correct this baseline.

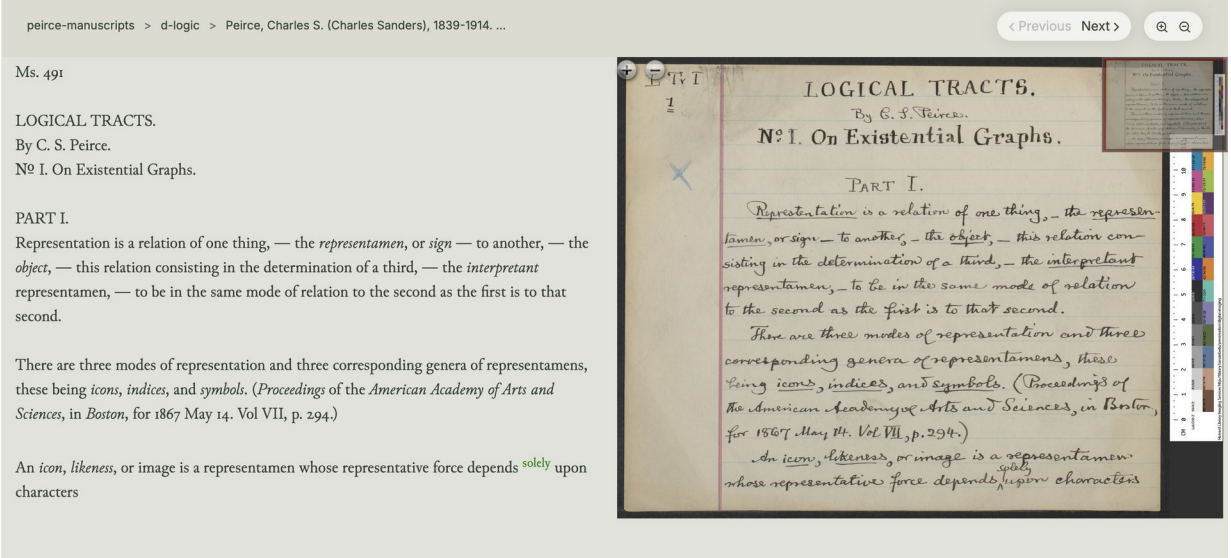


Figure 4. Diplomatic TEI transcription of MS 491, aligned with the IIIF facsimile

5. DISCUSSION

This project addresses a scholarly public composed of researchers, students, editors, and advanced readers engaged with Peirce's work. Public engagement is understood here as the creation of shared research infrastructures that enable exploration, comparison, and critical dialogue. By emphasizing access to manuscript materials, traceability of analytical transformations, and openness of methodological choices, the workflow supports forms of engagement grounded in reuse and collective interpretation.
The workflow also brings into view a set of methodological conditions that shape interpretation. Automated transcription produces variable results across manuscript types, handwriting styles, and material conditions, leading to differentiated data quality within the corpus. Computational models highlight recurring and statistically prominent patterns, while less frequent or ambiguous elements require careful interpretive attention. Framing LLMs as semiotic machines (Picca 2025) shifts us away from an intentional-stance view that treats them as quasi-psychological agents (Shanahan 2024). Instead, it allows us to describe them as systems that model sign relations, rather than as devices that merely recombine surface linguistic forms through statistical regularities. In a Peircean context, this shift motivates the development of semiotic-informed LLMs (explicitly aligned with triadic sign relations), and it frames their outputs as operational support for archival inquiry (e.g., chronological dating, while leaving interpretive judgment anchored in the manuscripts). Semantic modeling and visualization organize relations through formal structures and visual abstractions, which guide reading by foregrounding certain connections. These characteristics situate interpretation within an iterative process, where analytical results remain connected to manuscript sources and open to revision.
Within this framework, machine learning models contribute to archival inquiry by supporting classification and dating practices. Model-assisted clustering and comparison offer new perspectives on existing manuscript classifications, including those established by Robin, by revealing affinities and structural regularities across the corpus. Stylometric, semantic, and contextual signals provide additional evidence for situating manuscript pages in time, enriching historical and philological analysis. In these applications, AI serves as a tool for orienting scholarly investigation, extending the range of questions that can be asked while remaining embedded in interpretive practice.

6. CONCLUSION

This paper has proposed a workflow for engaging with large, heterogeneous manuscript corpora that integrates digitization, modeling, and visualization as mediating research practices. Designed around interpretive openness and traceability, the workflow is intended to be transferable to similar archival contexts in which textual, visual, and material features intersect. In the second phase of the project,

training language models on manuscript-level evidence to support the interpretation of Peirce's published writings offers a promising direction for exploring how computational methods can extend, rather than replace, scholarly reading.

REFERENCES

Atkin, A. (2013). Peirce's Theory of Signs (Edward N. Zalta, Ed.). Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/entries/peirce-semiotics/

Bateman, J. (2018). Peircean semiotics and multimodality: Towards a new synthesis. Multimodal Communication, 7.

Bella, M., Maddalena, G., & Tienne, A. D. (2016). What is Happening to the Peirce Project? European Journal of Pragmatism and American Philosophy, 8(2). https://doi.org/10.4000/ejpap.664

Bellucci, F., & Pietarinen, A. (2024). Existential graphs: History and interpretation. In C. d. Waal (Ed.), The Oxford Handbook of Charles S. Peirce (online). Oxford Academic. https://doi.org/10.1093/oxfordhb/9780197548561.013.16

Burch, R. (2021). Charles Sanders Peirce. In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Summer 2022). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2022/entries/peirce/

Ciula, A., Eide, Ø., Marras, C., & Sahle, P. (2023). Modelling Between Digital and Humanities: Thinking in Practice. Open Book Publishers. https://doi.org/10.11647/OBP.0369

Cristalli, C. (2022). Unconscious inferences in perception in early experimental psychology: From Wundt to Peirce. Journal of the History of the Behavioral Sciences, 58(4), 432–448. https://doi.org/https://doi.org/10.1002/jhbs.22211

Engel, F., Queisner, M., & Viola, T. (Eds.). (2012). Das bildnerische Denken: Charles S. Peirce [OCLC: ocn793401061]. Akademie Verlag.

Everett, D. (2021, September). Charles Sanders Peirce was America's greatest thinker [Publication Title: Attention to the Unseen]. Retrieved December 9, 2022, from https://attentiontotheunseen.com/2021/09/26/charles-sanders-peirce-was-americasgreatest-thinker

Farias, P., & Queiroz, J. (2017). Visualizing triadic relations: Diagrams for Charles S. Peirce's classifications of signs. Information Design Journal, 23(2), 127–147.

Feil, S. (2024, October). Pragmatismus als Theorie der Vorbegrifflichkeit. Brill | Fink. https://doi.org/10.30965/9783846768945

Harvard University. (n.d.). Harvard Library [Publication Title: Hollis.harvard.edu]. https://hollis.harvard.edu/

Hiippala, T., & Bateman, J. (2022). Semiotically-grounded distant view of diagrams: Insights from two multimodal corpora. Digital Scholarship in the Humanities, 37, 405–425. https://doi.org/10.1093/llc/fqab063

Jariah, R., Fathu, R., & Amir, M., P. (2022). Social Problems In Drama 13 Reasons Why: Peirce Semiotics Approach. TEKSTUAL, 20, 48. https://doi.org/10.33387/tekstual.v20i1.4508

Jockers, M. L. (2013). Macroanalysis: Digital Methods and Literary History. University of Illinois Press. https://doi.org/10.5406/illinois/9780252037528.001.0001

Kahle, P., Colutto, S., Hackl, G., & Muhlberger, G. (2017). Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents. 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), 19–24. https://doi.org/10.1109/ICDAR.2017.307

Keeler, M. (2020a). Pragmatically Improving Access to Peirce's Archive. Chinese Semiotic Studies, 16(1), 167–187. https://doi.org/10.1515/css-2020-0009

Keeler, M. (2020b). The Hidden Treasure of C. S. Peirce's Manuscripts. Chinese Semiotic Studies, 16(1), 155–166. https://doi.org/10.1515/css-2020-0008

Keeler, M., & Kloesel, C. (1997). Communication, Semiotic Continuity, and the Margins of the Peircean Text (D. Greetham, Ed.). Margins of the Text - University of Michigan Press. http://conceptualgraphs.org/revelator/web/papers/keelermargins1997.pdf

Liszka, J. (2021). Charles Peirce on Ethics, Esthetics and the Normative Sciences. Routledge. https://doi.org/10.4324/9781003160892

Liszka, J. J. (1996). A general introduction to the semiotic of Charles Sanders Peirce. Indiana University Press.

Mazur, L., & Sticksel, I. (2021). An empirical study of psychology and logic. Abduction and belief as normalizing habits of positive expectation. New Ideas in Psychology, 63, 100874. https://doi.org/https://doi.org/10.1016/j.newideapsych.2021.100874

Nöth, W. (2016). Habits, habit of change, and the habit of habit change according to Peirce. In D. E. West & M. Anderson (Eds.), Consensus on Peirce's concept of habit (pp. 35–63). Springer.

Pape, H. (1990). Charles S. Peirce on objects of thought and representation. Noûs, 24(3), 375–395.

Pape, H. (2015). Charles S. Peirce zur Einführung (2nd ed.). Junius.

Pedretti, C. T., Picca, D., & Rodighiero, D. (2025). Moving pictures of thought: Extracting visual knowledge in Charles S. Peirce's manuscripts with vision-language models. In T. Arnold, M. Fantoli, & R. Ros (Eds.), Anthology of Computers and the Humanities (pp. 1454–1467, Vol. 3). https://doi.org/10.63744/fkFGJ6wSzDPV

Pedretti, C. T., Picca, D., Rodighiero, D., Feil, S., & Adamou, A. (n.d.). Moving Pictures of Thought: Evaluating Vision-Language Models on Peirce's Existential Graphs [status: forthcoming]. Journal of Digital History.

Peirce, C. S. (1960). Collected Papers of Charles Sanders Peirce, Vol. I Principles of Philosophy and Vol. II Elements of Logic (C. Hartshorne & P. Weiss, Eds.). The Belknap of Harvard University Press.

Peirce, C. S. (1977). Semiotic and significs. The correspondence between Charles S. Peirce and Victoria Lady Welby (C. S. Hardwick, Ed.). Indiana University Press.

Peirce, C. S. (1955). Logic as semiotic: The theory of signs. In Justus Buchler (Ed.), Philosophical writings of Peirce (pp. 98–105). Dover Publications.

Peirce, C. S. (1994). The Collected Papers of Charles Sanders Peirce (J. Deely, Ed.). Harvard University Press.

Peirce Edition Project. (2021). Peirce Edition Project. Retrieved February 9, 2023, from https://peirce.iupui.edu/

Pflaeging, J., Wildfeuer, J., & Bateman, J. (Eds.). (2021). Empirical multimodality research: Methods, Applications, Implications. De Gruyter. https://doi.org/10.1515/9783110725001-001

Picca, D. (2025). Not minds, but signs: Reframing llms through semiotics. https://arxiv.org/abs/2505.17080

Picca, D., Schnyder, A., Kostina, E., Adamou, A., Rodighiero, D., & Schnapp, J. (2023). Orchestrating cultural heritage: Exploring the automated analysis and organization of Charles S. Peirce's PAP manuscript. Proceedings of the 34th ACM Conference on Hypertext and Social Media. https://doi.org/10.1145/3603163.3609066

Picca, D., & Zangari, L. (n.d.). Framing large language models through semiotics: A review [status: submitted].

Roberts, D. D. (2009). The existential graphs of Charles S. Peirce. Mouton de Gruyter.

Robin, R., S. (1967). Robin Catalog. Retrieved February 9, 2023, from https://peirce.sitehost.iu.edu/robin/rcatalog.htm

Shanahan, M. (2024). Talking about large language models. Communications of the ACM, 67(2), 68–79.

Sheriff, J. (1989). The Fate of Meaning: Charles Peirce, Structuralism, and Literature. Princeton Legacy Library. https://doi.org/10.1515/9781400859979

Sonesson, G. (2013). The natural history of branching: Approaches to the Phenomenology of Firstness, Secondness, and Thirdness. Signs and Society, 1(2), 297–325.

Stjernfelt, F. (2007). Diagrammatology: An investigation on the borderlines of phenomenology, ontology, and semiotics [OCLC: 255632682]. Springer.

Strand, T. (2022). A semiotic model of learning. Chinese Semiotic Studies, 17, 153–162.

Underwood, T. (2019). Distant Horizons: Digital Evidence and Literary Change. University of Chicago Press. https://doi.org/10.7208/chicago/9780226612973.001.0001

Zhang, Y. (2019). A Semiotic Study on Print Advertisements of Luxury Perfume Brands for Women. Asian Journal of Social Sciences Studies, 4(1). https://doi.org/https://doi.org/10.20849/ajsss.v4i1.540

A New Semantic Model for Mythologiae: from Data to Interpretation

BIANCA LA MANNA
UNIVERSITY OF BOLOGNA, ITALY – BIANCA.LAMANNA2@UNIBO.IT

ABSTRACT (ENGLISH)

Starting from the MythLOD project, which aimed at re-evaluating Mythologiae, an image repository of classical mythology representation, through Semantic Web technologies, this research aims at revisiting and updating the existing semantic model in Linked Open Data (LOD) according to the latest ontologies and descriptive standards. This project addresses the need to replace obsolete ontologies' versions or those that are not maintained, and enrich the schema with new ontologies, domain-specific terminology, and standard classification for representing relationships between iconographical works and textual sources. Furthermore, it will revise the reference model and methodology used to represent assertions, interpretations, and the provenance of the statements, to provide a more comprehensive and interoperable semantic representation of the recognition of mythological themes in visual arts, facilitating the analysis and reuse of these data in scholarly research.
Keywords: Linked Open Data; Semantic Web; Ontologies; RDFstar.

ABSTRACT (ITALIANO)

Un nuovo modello semantico per Mythologiae: dai dati alle interpretazioni. Partendo dal progetto MythLOD, nato allo scopo di rivalorizzare Mythologiae, un repository di immagini relative alla rappresentazione della mitologia classica, attraverso le tecnologie del web semantico, questa ricerca ha lo scopo di rivisitare e aggiornare il modello semantico in Linked Open Data (LOD) esistente con le versioni più recenti delle ontologie e degli standard descrittivi. Questo progetto si dedica alla sostituzione necessaria delle versioni delle ontologie ormai obsolete o non più mantenute, e all'arricchimento dello schema con nuove ontologie, terminologie di settore e classificazioni standard per descrivere il rapporto tra opera iconografica e citazione testuale. Si dedica, inoltre, alla rivisitazione del modello di riferimento e della metodologia utilizzata per la rappresentazione delle asserzioni, interpretazioni e della provenienza delle affermazioni, per offrire una rappresentazione semantica dell'atto di riconoscimento dei temi mitologici nelle arti visuali più completa e maggiormente interoperabile, facilitando l'analisi e il riutilizzo di questi dati nella ricerca scientifica.
Parole chiave: Linked Open Data; Semantic Web; Ontologies; RDFstar.

1. INTRODUCTION

Mythologiae[1] is a digital repository, born as a didactic project at the Department of Classical Heritage (DBC) of the University of Bologna, for collecting images of visual artworks representing narrative and mythographic structures. In 2022, the MythLOD project (Pasqual & Tomasi, 2022) converted this collection into Linked Open Data (LOD), providing a layered semantic framework for the description of data that goes from factual data to interpretation, with a focus on the relation between the subjects of the artworks and the textual sources.

This paper will discuss the model published by MythLOD and update it according to the recent developments in the Semantic Web technologies. In the first part, we will outline the original schema, with an analysis of the ontologies and RDF technologies employed. In the second part, the criteria and the results of the revision process will be presented. Finally, the implementation of the relationships between the different layers of information will be revised, proposing RDF-star as a solution for representing statements on factual data.

This work constitutes the first step for creating a public and accessible LOD triplestore for the digital collection of Mythologiae, enhancing interoperability, exchange, and crossing of data related to classical mythology.

[1] https://mythologiae.unibo.it/ (cons. 06/02/2026).

2. MYTHLOD SCHEMA OVERVIEW

The Mythologiae repository is centered on visual artworks, which are the primary object of description. As with any work of art, these objects are subject to assertions concerning the recognition of iconographical elements, cultural phenomena, references to other works, etc. These assertions are the result of an act of interpretation, made by someone at some time, and which is based on specific elements or contextual information. To manage this kind of complex representation, the general structure proposed by MythLOD (Pasqual & Tomasi, 2022) relies on the MIMA data model, which handles these different types of data through a 4-layered approach (Daquino et al, 2020):

- Layer 0. Factual data that are part of scholars' background knowledge.
- Layer 1. The scope of scholars' questionable statements.
- Layer 2. Context information for hypothesis assessment.
- Layer 3. Provenance information of the mining processes.

In the current implementation of the MythLOD knowledge graph (Pasqual & Tomasi, 2021), the various layers have been managed through the Nanopublication model (Greg et al., 2026), which provides "a structured approach to making statements about statements" (Menotti et al., 2025). Each item is represented as a nanopublication, with data from each layer contained in separate named graph, organized as follows: layer 0, which represents the factual data, has been built under the namespace of the project (myth:factual-data), while each one of the other layers corresponds to a specific named graph of the Nanopublication, i.e., layer 1 is a named graph of type np:Assertion, layer 2 of type np:Provenance, and layer 3 of type np:PublicationInfo.
Layer 0 of the MythLOD schema was intended to describe the physical object of the digital collection with its metadata. They were based on the following ontologies:

- Erlangen CRM (ECRM)[2]: an OWL implementation of the CIDOC Conceptual Reference Model (CRM) (Görz et al., 2008). Nowadays, it has been overtaken by the module CRM OWL[3].
- Erlangen Functional Requirements for Bibliographic Records object-oriented (EFRBRoo)[4], an implementation on OWL of FRBRoo, which has now been succeeded by LRMooo (IFLA Working Group & CIDOC CRM SIG, 2025).

Layer 1 modelled the relationships between the themes or characters identifiable in the image and textual sources that were related to the same subjects, and was based mainly on the Humanities Citation Ontology (HuCit)[5]. This ontology at present is not available since it was built on the aforementioned, ECRM and EFRBROO[6], which are nowadays outdated.
The layer 2 provides contextual information about the assertions and is based on five ontologies, which are still available and still in use with the same versions:

- FRBR-aligned Bibliographic Ontology (FaBiO), which manages published and publishable entities and bibliographic references (Peroni & Shotton, 2012; Shotton et al., 2019).
- The Historical Context Ontology (HiCO), for representing the context of a claim (Daquino et al., 2020; Daquino & Tomasi, 2015).
- The Citation Typing Ontology (CiTO) (Peroni & Shotton, 2012; Shotton et al., 2018).
- The PROV Ontology (PROV-O) to model provenance information (Lebo et al., 2013).
- The Canadian Writing Research Collaboratory (CWRC) ontology (Brown et al., 2023), for the representation of certainty.

Layer 3 is entirely based on PROV-O to store information about the publication of the data.

[2] https://erlangen-crm.org/ (cons. 06/02/2026).
[3] https://cidoc-crm.org/Issue/ID-678-towards-crm-owl (cons. 05/02/2026).
[4] https://erlangen-crm.org/efrbroo (cons. 06/02/2026).
[5] https://hucitlib.readthedocs.io/en/latest/ (cons. 06/02/2026).
[6] https://cidoc-crm.org/Resources/humanities-citation-ontology-hucit (cons. 06/02/2026).

3. SCHEMA UPDATING: FACTUAL DATA AND ASSERTIONS

The structure of the knowledge graph has been maintained in alignment with the MIMA model, updating one layer at a time, according to the specific cases.
In the original model layer 0, as it emerged, relied on ontologies that are no longer in use. In particular, the migration from FRBRoo to LRMoo did not involve only a change of namespace, but included a revision of the conceptual model behind the schema, although the ontology is still based on the FRBR model (IFLA Study Group on the FRBR, 1998). We updated the schema to reflect these changes:

- The physical object, which in MythLOD was represented by the class efrbroo:F4_Manifestation_Singleton, instance of efrbroo:F2_Expression, is now an instance of the class the lmroo:F5_Item, which is an exemplification of a lmroo:F3_Manifestation.
- The classes efrbroo:F27_Work_Conception and efrbroo:F28_Expression_Conception have been substituted by lmroo:E27_Work_Creation and lmroo:F28_Expression_Creation, with a shift from the focus on the act of conception to that of creation.

We also enriched the previous schema with additional classes that opens to a more comprehensive representation of the physical object, in particular, aligning it to the concept of artwork as defined in the ICON ontology (Sartini et al., 2023): the class lmroo:F5_Item is now a representation of a visual item (crm:E36_Visual_Item) which is a broad match of the concept of icon:Artwork (see Fig. 1)[7].

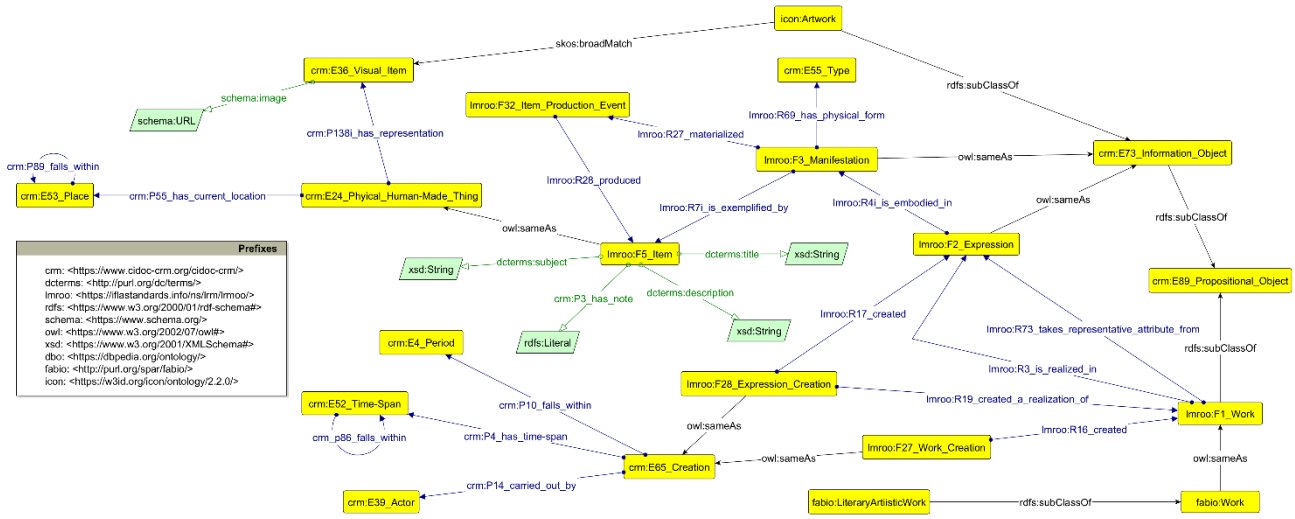


Figure 1. The conceptual model of the updated layer 0 – factual data.

Through ICON ontology, it will be possible to include in our schema statements or assertions about the recognition of an iconographical element with different levels of abstraction.
Moving to layer 1, which contains the statements resulting from interpretation, we want to describe the relationships that connect visual artworks, in whole or in their components, to the textual sources. The original model used primarily the HuCiT ontology, which, regardless of the technical problems, provided a direct relation between an artwork's subject and the canonical citation of its textual source. However, we would like to investigate and represent additional types of references between iconographical and textual objects, for example, adaptation of classical myths in modern literature, or generic references to a common mythological subject.
To incorporate in the future this kind of representation, we decided to add to the model a class that represents a narrative unit of a myth, which we will call "mytheme". It will act as an intermediate and common reference between the iconographical element and the textual one, enabling comparisons of different works and their interpretations of the same mytheme. In the conceptual model (Fig. 2) is possible to see how the bridge between the visual work as represented above and the literary works has been formalized.

[7] All the conceptual models have been graphically represented by means of Graffoo (Peroni, 2013).

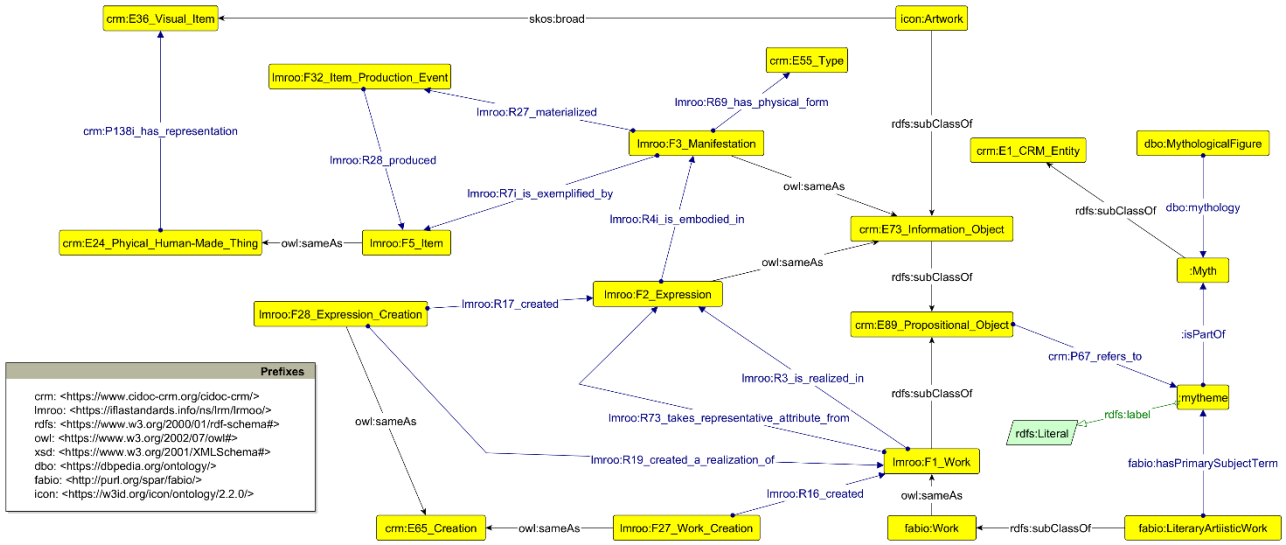


Figure 2. The conceptual model of the updated layer 1 – assertions.

Layers 2 and 3 do not require a structural update, since the ontologies implemented are still maintained and in use. Future work will be aimed at expanding and enriching these two layers with new metadata[8].

4. FROM NANOPUBLICATION TO RDFSTAR

Taking as an example the painting "Dedalo e Icaro" by Anton Van Dyck[9], from the digital collection of Mythologiae, we can see how the factual data about the artwork are actually modelled and presented (see Fig. 3).



Figure 3. "Dedalo e Icaro" by Anton Van Dyck, as described in Mythologiae.

According to the updated model, we can now serialize them in RDF (Fig. 4).

[8] In layer 2, for example, the concept of certainty of a statement is currently modelled through the generic class "Certainty" of the CWRC ontology, but domain-specific ontologies, such as the Argument Modelling Ontology (AMO), can provide a more comprehensive description (Vitali & Peroni 2011).
[9] https://mythologiae.unibo.it/index.php/2018/05/15/dedalo-e-icaro-3/ (cons. 05/02/2026).

```
prefix mythologiae: <https://mythologiae.unibo.it/>
prefix lmroo: <https://iflastandards.info/ns/lrm/lrmoo/>
prefix dcterms: <http://purl.org/dc/terms/>
prefix rdfs: <https://www.w3.org/2000/01/rdf-schema#>
prefix xsd: <https://www.w3.org/2001/XMLSchema#>
prefix crm: <https://www.cidoc-crm.org/cidoc-crm/>
prefix schema: <https://www.schema.org/>
prefix viaf: <http://viaf.org/viaf/>

mythologiae:dedalo-icaro-item a lmroo:F5_Item;
        dcterms:title "Dedalus and Icarus"^^rdfs:Literal;
        dcterms:description "Dedalo mentre con la mano sinistra tocca un'ala del
                figlio Icaro solleva il dito indice della mano destra. Il giovane
                ha già indossato le ali e guarda di fronte a sé."^^xsd:String;
        crm:P55_has_current_location viaf:151879312;
        crm:P3_note "Dipinto datato circa 1620"^^rdfs:Literal;
        schema:url <https://ago.ca/collection/object/2556>.
mythologiae:dedalo-icaro-expression a lmroo:F28_Expression_Creation;
        crm:P14_carried_out_by viaf:17231738;
        crm:P10_falls_within mythologiae:Arte-moderna;
        crm:P4_has_time-span mythologiae:1620;
        crm:P67_refers_to mythologiae:addestramento-icaro.
mythologiae:dedalo-icaro-manifestation a lmroo:F3_Manifestation;
        lmroo:R69_has_physical_form mythologiae:pittura.
```

```
viaf:17231738 a crm:E39_Actor;
        rdfs:label "Anton Van Dyck".

mythologiae:Arte-moderna a crm:E4_Period.

mythologiae:1620 a crm:E52_Time-Span;
        crm:P86_falls_within mythologiae:XVII-secolo;
        rdfs:label "1620"^^xsd:gYear.

mythologiae:XVII-secolo a crm:E52_Time-Span;
        rdfs:label "XVII secolo"^^rdfs:Literal.

mythologiae:pittura a crm:E55_Type.

mythologiae:addestramento-icaro a :mytheme.

viaf:151879312 a crm:E53_Place;
        rdfs:label "Art Gallery of Ontario, Toronto"^^rdfs:Literal.

<https://ago.ca/collection/object/2556> a schema:URL.
```

Figure 4. "Dedalo e Icaro" by Anton Van Dyck, RDF serialization.

In the MythLOD model, this item would be represented as a named graph, containing factual data; statements about its references to the mytheme (mythologiae:addestramento-icaro) would be part of a named graph of type np:Assertion, and the provenance of this assertion would be contained in a third graph of type np:Provenance. Expressing contextual information about the factual data through nanopublications is particularly helpful for provenance tracking; however, since we want to expand the schema with more nuanced contextual information and combine or exchange data with third-party knowledge graphs, a more flexible solution would be preferable. This function of expressing statements on statements is known as "Reification"[10], and is already implemented inside RDF, through various methods (Uzeir, 2023). For our scope, we will propose the use of RDF-star (Hartig, 2017), which is based on triple terms[11], i.e., RDF triples used as a term of another triple to express statements on it, and allows for keeping all the information on the same knowledge graph. Currently, it has been implemented in the new model to express the provenance of the statement relative to the recognition of the reference of an image to some mytheme (see Fig. 5).

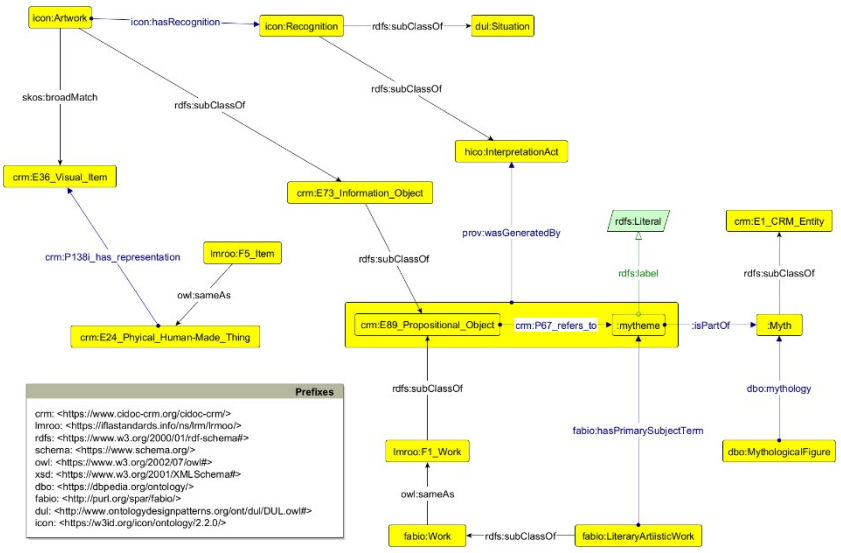


Figure 5. Layer 2 – provenance: example of a statement on a statement.

In this way, the provenance of the statement can be added to the RDF serialization of an item by adding a single line (see Fig. 5).

[10] https://www.w3.org/wiki/RdfReification (cons. 07/02/2026).
[11] https://www.w3.org/TR/rdf12-concepts/#dfn-triple-term (cons. 07/02/2026).

```
<<mythologiae:dedalo-icaro-expression crm:P67_refers_to mythologiae:addestramento-icaro>> prov:wasGeneratedBy mythologiae:icaro-interpretazione-1.

mythologiae:icaro-interpretazione-1 a hico:InterpretationAct.
```

Figure 5. "Dedalo e Icaro" by Anton Van Dyck, RDF serialization.

Being RDF native, it has the advantage of being supported by the main triplestore. Furthermore, it has its own query language, SPARQL-star (Arndt et al., 2021), thus allowing to easily retrieve the information both about the statement and about its context.

5. CONCLUSIONS

This project revisits and builds on the semantic model proposed by MythLOD, which already anticipated several key challenges of the Semantic Web, such as representation of interpretation, attribution, and the semantic linking of images and data through shared themes or subjects. As for many projects built under the principle of reuse (Wilkinson et al., 2016), however, it is exposed to the risk of technological obsolescence and lack of maintenance of the ontologies on which it depends. The present work aimed at refreshing the conceptual model by updating the ontologies with the latest versions and aligning it with the new practices that have emerged in recent years, opening the model to new forms of integration and reuse. Future work will focus on increasing the expressive granularity of the model through the analysis of taxonomies and categorical or domain-based classifications and exploring the representation of additional nuances, such as the degree of certainty of the statements, aimed at increasing trustability and expressiveness of the knowledge graph to support research on cultural data on the semantic web.

ACKNOWLEDGEMENTS

This project is part of the author's PhD research and is co-funded by the Department of Classical Philology
and Italianistics (FICLIT) and the Department of Cultural Heritage (DBC) of the University of Bologna.

REFERENCES

Arndt, D., Broekstra, J., DuCharme, B., Lassila, O., Patel-Schneider, P. F., Prud'hommeaux, E., Thibodeau, T. Jr., & Thompson, B. (2021, Decembre 17). RDF-star and SPARQL-star. Final Community Group Report 17 December 2021. https://w3c.github.io/rdf-star/cg-spec/2021-12-17.html.

Brown, S., Cummings, J., Drudge-Willson, J., Faulkner, C., Lemak, A., Long, J., Martin, K., Mo, A., Mustafa, L., Penancier, J., Simpson, J., Singh, G., Stacey, D., Stewart, H., Warren, R., Black, C., Crompton, C., Flynn, A., Holland, K., … Members of the Orlando Project. (2023, October 16). The CWRC Ontology Specification 0.99.88 Documentation. https://sparql.cwrc.ca/ontologies/cwrc-2023-10-16.html.

CIDOC CRM Special Interest Group (SIG). (2024, February). Volume A: Definition of the CIDOC Conceptual Reference Model version 7.1.3. https://cidoc-crm.org/sites/default/files/cidoc_crm_version_7.1.3.pdf.

Daquino, M., Pasqual, V., & Tomasi, F. (2020). Knowledge representation of digital hermeneutics of archival and literary sources. JLIS.IT, 11, 59-76. ISSN: 2038-5366. https://doi.org/10.4403/jlis.it-12642.

Daquino, M., Peroni, S., & Tomasi, F. (2020, March 31). HiCO, Historical Context Ontology Documentation. http://purl.org/emmedi/hico.

Daquino, M., & Tomasi, F. (2015). Historical Context Ontology (HiCO): A Conceptual Model for Describing Context Information of Cultural Heritage Objects. https://doi.org/10.1007/978-3-319-24129-6_37.

Gray, A. J. G., Burger, K., Kotoulas, S., Loizou, A., Tkachenko, V., Waagmeester, A., Askjaer, S., Pettifer, S., Harland, L., Haupt, C., Batchelor, Vazquez, M., Fernández, J. M., Saito, J., Gibson, A., Wich, L., van Dam, J., & Wilkinson, M. (2026, February 05). Nanopublication Guidelines. https://nanopub.net/guidelines/working_draft/.

Görz, G., Schiemann, B., & Oischinger, M. (2008). An implementation of the CIDOC Conceptual Reference Model (4.2.4) in OWL-DL. Conference paper for CIDOC 2008.

Hartig, O. (2017, June). Foundations of RDF* and SPARQL* - An Alternative Approach to Statement-Level Metadata in RDF. In Proceedings of the 11th Alberto Mendelzon International Workshop on Foundations of Data Management (AMW), Montevideo, Uruguay. http://ceur-ws.org/Vol-1912/paper12.pdf.

IFLA LRMoo Working Group & CIDOC CRM Special Interest Group (SIG). (2025, October). LRMoo: object-oriented definition and mapping from the IFLA Library Reference Model version 1.1.1. https://cidoc-crm.org/sites/default/files/LRMoo_V1.1.1.pdf.

IFLA Study Group on the Functional Requirements for Bibliographic Records (FRBR). (1998). Functional requirements for bibliographic records: final report. München: K. G. Saur.

Lebo, T., Sahoo, S., McGuinness, D., Belhajjame, K., Cheney, J., Corsar, D., Garijo, D., Soiland-Reyes, S., Zednik, S., & Zhao, J. (2013, April 30). PROV-O: The PROV Ontology. https://www.w3.org/TR/prov-o/.

Menotti, L., Marchesin, S., Giachelle, F., & Silvello, G. (2025). Provenance-driven nanopublications: representing source lineage and trust networks for multi-source assertions. Int J Digit Libr 26, 24. https://doi.org/10.1007/s00799-025-00431-x.

Pasqual, V. & Tomasi, F.. (2021). mythlod/1.0 (Version 1.0) [Data set]. Zenodo. http://doi.org/10.5281/zenodo.4639821.

Pasqual, V. & Tomasi, F. (2022). Linked Open Data Per La Valorizzazione Di Collezioni Culturali: Il Dataset MythLOD. AIB Studi 62(1), 149-68. https://doi.org/10.2426/aibstudi-13301.

Peroni, Silvio. (2013, October 28). Graffoo. https://essepuntato.it/graffoo/specification/.

Peroni, Silvio & Shotton, David. (2012). FaBiO and CiTO: ontologies for describing bibliographic resources and citations. Journal of Web Semantics, 17, 33-43. https://doi.org/10.1016/j.websem.2012.08.001.

Sartini, B., Baroncini, S., van Erp, M., Tomasi, F., & Gangemi, A. (2023). ICON: An Ontology for Comprehensive Artistic Interpretations, J. Comput. Cult. Herit. 16, 3, Article 59 (September 2023). https://doi.org/10.1145/3594724.

Shotton, D., Peroni, S., Ciccarese, P., & Clark, T. (2018, February 16). CiTO, the Citation Typing Ontology Documentation. http://purl.org/spar/cito.

Shotton, D., Peroni, S., Ciccarese, P., & Clark, T. (2019, February 19). FRBR-aligned Bibliographic Ontology (FaBiO) Documentation. http://purl.org/spar/fabio.

Vitali, F., & Peroni, S. (2011, May). The Argument Model Ontology (AMO). https://sparontologies.github.io/amo/current/amo.html.

Uzeir, E. (2023). Benchmarking “Expressing without Asserting” Approaches on a Large Wikidata Dump [Unpublished master’s thesis], 15-23. University of Bologna.

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Sci Data volume, 3(1). https://doi.org/10.1038/sdata.2016.18.

[7/A] Ontologie, FAIR data e modellazione della conoscenza

The ATLAS Guidelines for producing FAIR research products in the Digital Humanities

CHIARA MARTIGNANO,[1] GIORGIA RUBIN,[2] SEBASTIANO GIACOMINI,[3] ALESSIA BARDI,[4] MARINA BUZZONI,[5] MARILENA DAQUINO,[6] RICCARDO DEL GRATTA,[7] ANGELO MARIO DEL GROSSO,[8] FRANZ FISCHER,[9] ROBERTO ROSSELLI DEL TURCO,[10] FRANCESCA TOMASI[11]

[1] UNIVERSITÀ DI PADOVA, ITALIA – CHIARA.MARTIGNANO@UNIPD.IT
[2] CNR-ILC, ITALIA – GIORGIA.RUBIN@ILC.CNR.IT
[3] UNIVERSITÀ DI BOLOGNA, ITALIA – SEBASTIANO.GIACOMIN2@UNIBO.IT
[4] CNR-ISTI, ITALIA – ALESSIA.BARDI@ISTI.CNR.IT
[5] UNIVERSITÀ CA' FOSCARI VENEZIA, ITALIA – MBUZZONI@UNIVE.IT
[6] UNIVERSITÀ DI BOLOGNA, ITALIA – MARILENA.DAQUINO2@UNIBO.IT
[7] CNR-ILC, ITALIA – RICCARDO.DELGRATTA@ILC.CNR.IT
[8] CNR-ILC, ITALIA - ANGELO.DELGROSSO@ILC.CNR.IT
[9] UNIVERSITÀ CA' FOSCARI VENEZIA, ITALIA - FRANZ.FISCHER@UNIVE.IT
[10] UNIVERSITÀ DI TORINO, ITALIA - ROBERTO.ROSSELLIDELTURCO@UNITO.IT
[11] UNIVERSITÀ DI BOLOGNA, ITALIA – FRANCESCA.TOMASI@UNIBO.IT

ABSTRACT (ENGLISH)

The ATLAS guidelines were developed as part of the PRIN 2022 ATLAS project to address a gap in the Digital Humanities (DH) landscape: the lack of practical guidance for making Humanities digital research products compliant with the FAIR principles (Findability, Accessibility, Interoperability, and Reusability). Building upon the ALLEA (All European Academies) recommendations, the document offers specific strategies for five key categories: digital scholarly editions, text collections, software tools, ontologies, and linked open datasets. The guidelines complement existing disciplinary standards by providing direct references to tools and best practices. Tested with undergraduate and PhD students, they demonstrated a high degree of usefulness and completeness, confirming their role as an essential resource for improving the quality, citability, and long-term relevance of DH research.
Keywords: FAIR principles; guidelines; research outputs.

ABSTRACT (ITALIANO)

Le linee guida ATLAS per realizzare prodotti di ricerca FAIR nelle Digital Humanities
Abstract tradotto in italiano.
Le linee guida ATLAS sono state sviluppate nell'ambito del progetto PRIN 2022 ATLAS per colmare una lacuna nel panorama delle Digital Humanities (DH): la mancanza di indicazioni pratiche per rendere i prodotti della ricerca umanistica digitale conformi ai principi FAIR (Findability, Accessibility, Interoperability, and Reusability). Basandosi sulle raccomandazioni dell'ALLEA (All European Academies), il documento offre strategie specifiche per cinque categorie chiave: edizioni accademiche digitali, raccolte di testi, strumenti software, ontologie e set di dati aperti collegati. Le linee guida integrano gli standard disciplinari esistenti fornendo riferimenti diretti a strumenti e best practice. Testate con studenti universitari e dottorandi, hanno dimostrato un alto grado di utilità e completezza, confermando il loro ruolo di risorsa essenziale per migliorare la qualità, la citabilità e la rilevanza a lungo termine della ricerca DH.
Parole chiave: principi FAIR; linee guida; output della ricerca.

1. INTRODUCTION

Data FAIRness is now a key requirement for public research funding, including Horizon 2020. But how exactly is data FAIRness achieved? The FAIR principles (Wilkinson et al., 2016) are a well-known set of guidelines for making scholarly research data findable, accessible, interoperable, and reusable by leveraging web technologies, metadata standards, and open formats. Officially presented in an article published in 2016, the FAIR principles have since gained widespread recognition within the global scientific community.

In recent years, many resources have been produced by universities, data stewards and other research institutions to promote the use of the FAIR principles and educate scholars about them. However, despite various resources being available, the DH practitioner community lacks comprehensive and practical guidelines for implementing FAIR principles in the research products that characterise the field of DH, such as digital scholarly editions, linked open datasets, and text collections.
The guidelines presented in this paper aim at filling this gap and are one of the key outcomes of the PRIN 2022 ATLAS project,[1] which was a collaborative effort involving the University of Bologna, Ca' Foscari University of Venice and the CNR Unit of Pisa (comprising the Institute of Information Science and Technologies "Alessandro Faedo" and the Institute for Computational Linguistics "A. Zampolli"). The project's primary objective was to create a catalogue of DH research on Italian cultural heritage using semantic web technologies. Focusing on a selection of pilot projects (Martignano et al., 2025a), we investigated the broad field of DH research outputs in order to collect data for the catalogue and study the most effective ways to model and semantically represent them. We observed that most of the pilot projects were not fully compliant with the FAIR principles (Martignano et al., 2025a). Even when they were compliant, they lacked important aspects to foster reuse in the DH, such as user documentation and an explicit reference to the models and standards used. We developed a semantic framework in the form of an OWL-based ontology (Giacomini et al., 2025) that captures the diversity of DH research outputs (Giacomini et al., 2025) and their contents, and enables the representation of key methodological aspects for achieving full FAIR compliance and high-quality research results. The work conducted to develop the ontology and the knowledge graph underlying our catalogue formed the basis for creating the guidelines, as outlined in section 3.

2. STATE OF THE ART AND OBJECTIVES

Scholars currently have a wide range of resources available to educate themselves on the FAIR principles, including video courses, guides, testing tools and informative materials. Alongside the many resources produced within research infrastructures such as DARIAH, CLARIN, and others, other examples of these resources include the following web portals: How to FAIR,[2] which is aimed mostly at individual scholars who want to learn the FAIR principles from scratch; GO-FAIR,[3] which provides a quick guide on how to implement FAIRness with a focus on machine actionability; and FAIRsFAIR (Fostering FAIR Data Practices in Europe),[4] which is a comprehensive portal offering resources to help data stewards implement FAIR data repositories, as well as supporting educators in training personnel and students on the FAIR principles. The main limitation of these resources is that they are intended for a broad audience and tend to offer general recommendations applicable across academic fields. Notable resources specifically designed for the humanities and social sciences are PARTHENOS's 'Guidelines to FAIRify Data Management and Make Data Reusable' (PARTHENOS et al., 2018) and the recommendations of the ALLEA E-Humanities Working Group, 'Sustainable and FAIR Data Sharing in the Humanities' (Harrower et al., 2020).
These two resources have their strengths and limitations. PARTHENOS's guidelines are concise and provide clear tips on how to achieve FAIR compliance. However, they do not include references to existing resources for a more in-depth study of the FAIR principles, nor to community-specific standards and best practices. The ALLEA recommendations are also very clear and well organised, and are intended for all scholars in the humanities. However, their main limitations are that: 1) they do not provide specific strategies for implementing the various types of output produced by scholars in the humanities, and 2) only a few existing tools are mentioned as examples within the recommendations.
Given that FAIR compliance requires adherence to standards, it is important to consider the many different format and metadata standards, as well as the guidelines specific to different academic fields, that have been developed over the years. Scholars who are relatively new to digital humanities (DH) and its best practices may have difficulty identifying the standards that best suit their research objectives and the strategies for properly implementing them.
Our guidelines aim to provide best practices and implementation strategies, as well as useful references to community-specific standards and guidelines for creating FAIR-by-design research products. Our

[1] https://dh-atlas.github.io/.
[2] https://www.howtofair.dk/.
[3] https://www.go-fair.org/.
[4] https://www.fairsfair.eu/.

guidelines are intended for DH scholars who want to learn about or deepen their understanding of the FAIR principles, as well as for those who are new to DH and want to explore the field's distinctive research products.

3. METHODOLOGY

The development of the guidelines was based on a review of existing FAIR data implementation guidelines, as described in the previous section, and an analysis of pilot projects falling under the following five categories: digital scholarly editions (e.g., Codice Pelavicino Digitale, National Edition of Aldo Moro's Works), text collections (e.g., Musisque Deoque, ALIM), software tools (e.g., EVT, Voyant Tools), ontologies (e.g., SPAR Ontologies, HiCO), and linked open datasets (e.g., Zeri & LODE, DanteSources). The full list of the selected pilot projects is presented in the project's whitebook (Martignano et al., 2025a), one of the main outcomes of our research. This volume is aimed at the data curators of our catalogue, as it presents the semantic framework and the tools we developed for data access and curation. Due to the expertise of our project team and the time limitations, we decided to focus on these five categories. While these categories do not cover the full range of DH research products, they represent a significant part of them. The pilot analysis enabled us to identify existing best practices, guidelines and standards particular to each type of research product. It also led to a critical reflection on the tension between FAIR compliance and actual reusability in the Humanities. We observed that, despite the publication of relevant and FAIR-compliant metadata, most DH research outputs, fail to be reusable due to the following common issues:

- **Incomplete or unclear metadata on the landing pages of research products.** Key metadata for FAIR compliance, such as licences, access rights and the names of all creators and contributors, is often hidden on a secondary webpage or in the deposited versions of the products, or is omitted entirely. This makes it more difficult for users to properly credit the authors and assess how the products can be used.
- **Lack of effective and comprehensive user documentation.** Describing the methodologies applied and including references to the data models, standards and software tools used is crucial to encouraging other researchers to reuse the research product or its production workflow to create a similar one.
- **Unclear production status.** In DH, it is common practice to publish research products before completion in order to test them or gather feedback during development. However, more often than not, the status of the product is not explicitly declared in the documentation or on the landing page.
- **Undocumented changes and improper version management.** This issue is related to the previous one, as published research products often undergo many changes that are neither uniquely identifiable nor documented.

We found that existing guidelines for data FAIRness did not address these common issues. However, many of these resources were useful for learning about the FAIR principles from a more general point of view. For this reason, we decided to build upon existing work, particularly the ALLEA recommendations, as these focus on the Humanities. Our guidelines are designed to complement existing resources and help DH practitioners and scholars produce digital research outputs that are FAIR-compliant, reusable and high-quality from the outset. The added value of the ATLAS guidelines lies in the translation of high-level principles into specific workflows for the creation of complex DH objects, with a level of granularity and completeness that is missing in current literature.
The ALLEA recommendations guide users through every stage of a digital object's lifecycle, from planning to dissemination and preservation. Each set of recommendations is dedicated to a specific phase and begins with a detailed introduction that explains the rationale and benefits. This is followed by a curated list of web resources and references for further learning.
As our focus was on specific aspects and problems to consider when producing FAIR digital scholarly editions, text collections, software tools, ontologies, and linked-open datasets, we structured our guidelines into five main sections, one for each category.
We decided to frame all recommendations within the ALLEA digital object life cycle framework, simplifying it slightly and creating a tag for each phase:

- Identify: recommendations focused on identifying your own research data, community standards and best practices, as well as existing resources, such as tools and datasets, that could be useful for creating the research product.
- Plan: recommendations about planning your research product lifecycle and what to consider from the outset.
- Produce: recommendations for implementing your research product, paying particular attention to formats, metadata and data models.
- Deposit: recommendations for properly depositing and preserving your research product, includingconsiderations for licences.
- Disseminate: recommendations on how to effectively reach your audience and make your research product reusable.

We then created colourful tags bearing the names of the listed phases. Each recommendation was assigned one or more tags depending on which of the phases it pertained to. This way, users are guided to consult the more general ALLEA recommendations before delving into the strategies and tips provided by our guidelines.
Compared to the ALLEA recommendations, the organisation of our guidelines aims to simplify and summarise salient aspects that may result complex to novices (as illustrated in Figure 1). Firstly, we decided to number our recommendations for easier reference and access. Secondly, we aimed to make our guidelines very practical to help users quickly identify useful resources such as community standards, tools, examples and guides. For this reason, each section begins with an overview of the standards and existing guidelines relating to the research product category, including external links. Within each set, each recommendation is structured as follows:

- A number and a short title for quick and easy identification;
- The recommendation itself;
- The phase tags;
- One or two paragraphs providing the rationale and benefits of applying the recommendation, along with tips;
- A list of links to useful resources.



Figure 1. Comparison of the structure differences between the ALLEA recommendations and the ATLAS guidelines.

The ATLAS guidelines begin with a detailed review of existing guidelines and a brief description of the project's objectives and methodology. We have also decided to present research infrastructures and organisations active in the field of the humanities and social sciences, as they play a key role in implementing and educating people about FAIR principles. These include ARIADNE, CLARIN, DARIAH, E-RIHS, CESSDA and OPERAS. Our aim was to provide users with not only best practices, but also references to existing communities of scholars and practitioners from a range of academic fields, including archaeology, linguistics, history and political science.
Finally, the guidelines conclude with a set of general recommendations and strategies that address the common reusability issues listed at the beginning of this section.

4. RESULTS

We have implemented the ATLAS guidelines in two formats: as a print-ready publication and as an online website.[5] This allows users to explore more easily our guidelines together with the many online resources we provide.
The first version of the guidelines is available on Zenodo as an integral part of the first version of the project's whitebook (Martignano et al., 2025a). The second version of the guidelines is soon to be released as part of an open access volume that also comprises the whitebook.
As you can see in the example below (Figure 2), there are slight editorial differences in the organisation of the contents between the volume and the website.

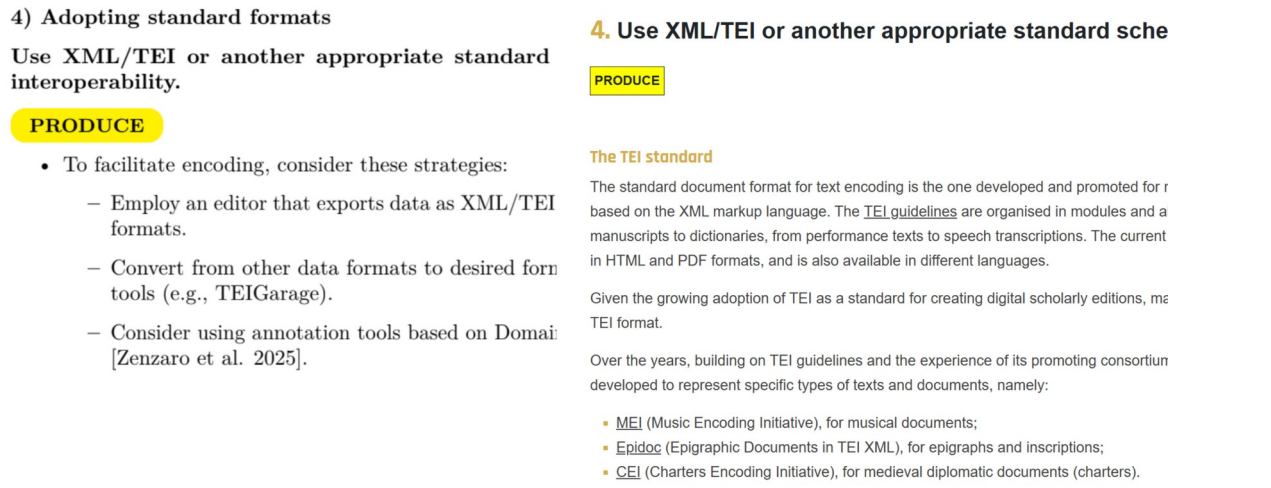


Figure 2. Comparison of the print-oriented and web version of a recommendation in the ATLAS guidelines.

We had the opportunity to assess the perceived value of the guidelines via surveys (Giacomini, Martignano, et al., 2025; Martignano & Rubin, 2025) during the final workshop and a seminar presentation of the project. At the first event, the survey population was more varied, comprising five Master's students, a PhD candidate, a professor and three researchers. At the second event, eleven people responded to our survey. Nine of these were Master's students in Digital Humanities, while one was a Bachelor's student. We asked them to rate the clarity of the terminology, the completeness of the guidelines, and their usefulness on a Likert scale from 1 to 5. We also asked them to indicate how much of the information was new to them (see Table 1 for the results). The more expert population at the first event perceived the terminology as clearer, while the undergraduates found more new information. The students at the second event may not have found the guidelines easy to read because they needed more detailed, tailored descriptions (which they can find in the Whitebook); hence the more neutral ratings. PhD students and researchers, on the other hand, are already familiar with the subject matter and therefore appreciate the 'operational' summary of the guidelines.

[5] https://dh-atlas.github.io/guidelines/.

	Average rating at the first event	Average rating at the second event
Clarity	4,5/5	3,33/5
Completeness	4,4/5	3,78/5
Usefulness	4,5/5	3,78/5
New information	2,9/5	4,22/5

Table 1. Average ratings of PhD and undergraduate students collected when testing the effectiveness of the ATLAS guidelines

It is important to note that these evaluations represent a preliminary assessment phase. Despite the limited sample size, the feedback collected was instrumental in refining the first complete version of the guidelines, particularly in optimising their structure and visual formatting for better readability and usability. To provide more robust evidence of their effectiveness, we plan to broaden the evaluation by disseminating the guidelines through international scientific venues (conferences, blogs, mailing lists and community forums), accompanied by a more extensive validation survey and a call for contributions.

5. CONCLUSIONS

The ATLAS guidelines aim at becoming a vital resource for disseminating the FAIR principles to the wider DH practitioner community. Building upon ALLEA directives, the guidelines focus on five common output types: digital scholarly editions, text collections, software, ontologies and linked open datasets. They provide standards and advice for every stage of the data lifecycle, including planning, modelling and publishing. The aim of this resource is to complement, rather than replace, existing disciplinary standards, thereby enhancing the overall quality of research products. Furthermore, by facilitating the creation of FAIR-compliant digital resources, we ensure that cultural heritage data is truly accessible and reusable for the general public, educators, and local communities. By following the ATLAS guidelines, scholars can create resources that are technically compliant, easier to cite and evaluate, and more likely to be reused, thus ensuring their long-term relevance in the DH landscape.

ACKNOWLEDGEMENTS

Funded by the European Union- Next Generation EU, Mission 4 Component 1 CUP J53D23013000006.

REFERENCES

Daquino, M., Bardi, A., Buzzoni, M., Del Gratta, R., Del Grosso, A. M., Fischer, F., Tomasi, F., & Rosselli Del Turco, R. (2024). *The ATLAS: A knowledge graph of digital scholarly research on Italian Cultural Heritage*. In *Atti Del XIII Convegno Annuale AIUCD 2024*, edited by Antonio Di Silvestro and Daria Spampinato. https://doi.org/10.6092/unibo/amsacta/7927.

Daquino, M., Bardi, A., Buzzoni, M., Del Gratta, R., Del Grosso, A. M., Fischer, F., Tomasi, F., & Rosselli Del Turco, R. (2024). *The ATLAS: A knowledge graph of digital scholarly research on Italian Cultural Heritage*. https://doi.org/10.6092/unibo/amsacta/7927.

Giacomini, S., Bardi, A., Buzzoni, M., Daquino, M., Del Gratta, R., Del Grosso, A. M., Fischer, F., Martignano, C., Del Turco, R. R., Rubin, G., & Tomasi, F. (2025). ATLAS: Towards a Knowledge Graph of International Scholarly Research on the Italian Digital Cultural Heritage. *CEUR Workshop Proc.*, *3937*. https://ceur-ws.org/Vol-3937/paper4.pdf.

Giacomini, S., Daquino, M., Tomasi, F., Bardi, A., Buzzoni, M., Del Gratta, R., Del Grosso, A. M., Fischer, F., Martignano, C., Rosselli Del Turco, R., & Rubin, G. (2025). *DH ATLAS: Ontology v2.0*. Zenodo. https://doi.org/10.5281/ZENODO.16780282.

Giacomini, S., Martignano, C., & Rubin, G. (2025). *DH ATLAS: Closing Workshop – Focus Group Survey Results v1.0* [Data set]. Zenodo. https://doi.org/10.5281/ZENODO.17332837.

Harrower, N., Maryl, M., Biro, T., Immenhauser, B., & ALLEA Working Group E-Humanities. (2020, February 18). *Sustainable and FAIR Data Sharing in the Humanities: Recommendations of the*

ALLEA Working Group E-Humanities - Digital Repository of Ireland. ALLEA - All European Academies. https://doi.org/10.7486/DRI.tq582c863.

Martignano, C., Bardi, A., Buzzoni, M., Daquino, M., Del Gratta, R., Del Grosso, A. M., Fischer, F., Giacomini, S., Rosselli Del Turco, R., Rubin, G., & Tomasi, F. (2025a). *DH ATLAS: White book v1.2*. Zenodo. https://doi.org/10.5281/ZENODO.14925266.

Martignano, C., Bardi, A., Buzzoni, M., Daquino, M., Del Gratta, R., Del Grosso, A. M., Fischer, F., Giacomini, S., Rosselli Del Turco, R., Rubin, G., & Tomasi, F. (2025b). *The ATLAS Guidelines v1.0*. https://doi.org/10.5281/ZENODO.17167253.

Martignano, C., & Rubin, G. (2025). *DH ATLAS: Digital Culture Seminar – Feedback Survey v1.0* [Data set]. Zenodo. https://doi.org/10.5281/ZENODO.17332765.

PARTHENOS, Hollander, H., Morselli, F., Uiterwaal, F., Admiraal, F., Trippel, T., & Di Giorgio, S. (2018). *PARTHENOS Guidelines to FAIRify data management and make data reusable*. https://doi.org/10.5281/zenodo.2668479.

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18.

RATIO: strumenti per la creazione e la gestione di risorse e collezioni FAIR nell'ambito delle scienze umane

FEDERICO SILVERSTRI[1], LORENZO MANCINI[2], LAURA BAGGIANI[3], MARCO BAGIACCHI[4]

[1] ISTITUTO PER IL LESSICO INTELLETTUALE EUROPEO E STORIA DELLE IDEE (CNR-ILIESI), ITALIA, FEDERICO.SILVESTRI@CNR.IT
[2] ISTITUTO PER IL LESSICO INTELLETTUALE EUROPEO E STORIA DELLE IDEE (CNR-ILIESI), ITALIA, LORENZO.MANCINI@CNR.IT
[3] ISTITUTO PER IL LESSICO INTELLETTUALE EUROPEO E STORIA DELLE IDEE (CNR-ILIESI), ITALIA, LAURA.BAGGIANI@CNR.IT
[4] ISTITUTO PER IL LESSICO INTELLETTUALE EUROPEO E STORIA DELLE IDEE (CNR-ILIESI), ITALIA, MARCO.BAGIACCHI@CNR.IT

ABSTRACT (ITALIANO)

Il paper presenta la suite RATIO sviluppata dall'unità operativa CNR-ILIESI all'interno del progetto infrastrutturale H2IOSC per conto di OPERAS-IT. La suite è costituita da tre servizi principali che mettono a disposizione altrettanti strumenti: FAST (Fairness ASsesment Tool), CONNECT (CollectiONs maiNtanencE and produCTion) e CONNECT-TED (Transcription and Edition Dashboard). I tre strumenti possono essere utilizzati singolarmente o all'interno di un flusso di lavoro che ha nell'adesione ai principi dell'Open Science e del paradigma FAIR uno dei suoi punti di forza. Oltre a descriverne i principi di funzionamento e le principali caratteristiche, il paper presenta anche il modello dei dati implementato in RATIO che è alla base dell'integrazione tra i servizi.
Parole chiave: FAIRificazione; Linked Open Data; Collezioni digitali; Edizioni digitali.

ABSTRACT (ENGLISH)

Ratio: tools for creating and managing FAIR resources and collections in Humanities.
The paper presents the RATIO suite developed by the CNR-ILIESI operational unit within the H2IOSC infrastructure project on behalf of OPERAS-IT. The suite consists of three main services that provide three tools: FAST (Fairness ASsesment Tool), CONNECT (CollectiONs maiNtanencE and produCTion) and CONNECT-TED (Transcription and Edition Dashboard). The three tools can be used individually or within a workflow that has adherence to the principles of Open Science and the FAIR paradigm as one of its strengths. In addition to describing its operating principles and main features, the paper also presents the data model implemented in RATIO, which is the basis for the integration among the services.
Keywords: FAIRification; Linked Open Data; Digital Collections; Digital editions.

1. INTRODUZIONE

Giunto alla sua conclusione, il progetto Humanities and Cultural Heritage Italian Open Science Cloud (H2IOSC) ha costituito un importante momento di riflessione e sviluppo per una molteplicità di aspetti e strumenti legati alle Digital Humanities.[1] Nato attraverso la federazione dei nodi italiani di quattro infrastrutture di ricerca europee (CLARIN, E-RIHS, DARIAH, OPERAS), H2IOSC ha avuto come obiettivo primario quello di creare un cluster collaborativo tra di esse, in modo da fornire un punto di accesso unitario, integrato e scalabile a servizi, dataset e varie altre tipologie di risorse (Sichera, Marras, Pasini, 2024) con un focus specifico sulle scienze umane e del patrimonio culturale. Un'attenzione particolare è stata rivolta anche al tema della sostenibilità, con l'impegno a mantenere attivo e aggiornato il cluster per almeno 10 anni.
Tra le iniziative messe in campo, particolare rilevanza hanno avuto i progetti pilota del Work Package 7, attraverso cui è stato possibile sviluppare nuovi strumenti di lavoro per le comunità di riferimento – o aggiornarne/perfezionarne di esistenti. In questo contesto è maturato anche il lavoro svolto da OPERAS-IT e dalle unità operative a esso collegate. Il CNR-ILIESI, in particolare, ha sviluppato una piccola suite di strumenti denominata RATIO, attraverso cui si intende mettere in condizione ricercatori e progetti di ricerca nell'ambito delle scienze umane di creare, importare, gestire, manutenere e archiviare risorse

[1] https://www.h2iosc.cnr.it/ (cons. 31/01/2026)

digitali e set di dati, ponendosi così come un punto di riferimento per la diffusione di pratiche di produzione e gestione di dati e risorse secondo il paradigma FAIR. L'adozione di questi principi - che prevedono la pubblicazione di dati reperibili, accessibili, interoperabili e riutilizzabili – avviene attraverso l'uso di standard aperti e condivisi, metadati ricchi e identificatori persistenti.
RATIO è costituito da tre strumenti che nascono con finalità diverse, descritti nelle pagine seguenti: FAST (Fairness ASsesment Tool), per guidare l'utente nella creazione di record di risorse in accordo con i principi FAIR; CONNECT (CollectiONs maiNtanencE and produCTion), per importare, gestire e arricchire collezioni eterogenee di dati (principalmente testuali, bibliografici e lessicali); e CONNECT-TED (Transcription and Edition Dashboard), per creare trascrizioni ed edizioni di testi a stampa. Tutta la suite è sviluppata con tecnologie open source ed è utilizzabile integralmente in ambiente web attraverso un sistema di autenticazione federato. Seppur fruibili singolarmente, è nella loro integrazione che va individuato il valore aggiunto di RATIO rispetto a progetti di ispirazione analoga.

2. STATO DELL'ARTE

Negli ultimi anni la gestione FAIR dei dati nell'ambito delle scienze umane si è sviluppata attraverso una serie di strumenti, piattaforme e infrastrutture che supportano la valutazione, la modellazione e la progettazione di risorse digitali interoperabili. Le iniziative internazionali qui presentate condividono l'obiettivo di rendere i dati delle Humanities più strutturati, riusabili e integrabili nei contesti di ricerca.
Un primo contributo rilevante proviene da FUJI - FAIR Assessment Tool[2], un servizio automatizzato per la valutazione della FAIRness dei dataset, che si presuppongono già dotati di identificativi propri. Lo strumento applica indicatori standardizzati per misurare Findability, Accessibility, Interoperability e Reusability, restituendo report dettagliati e suggerimenti operativi. La sua adozione è crescente anche nel contesto umanistico, dove la verifica della qualità dei dati digitali è un requisito sempre più centrale, sebbene fatichi nel rappresentare aspetti contestuali, istituzionali o relazionali tipici dei dati umanistici.
Accanto agli strumenti di valutazione, FactGrid – Wikibase for Historical Knowledge Graphs[3] rappresenta una piattaforma collaborativa che integra pratiche FAIR nella modellazione dei dati. Basato su Wikibase, consente la creazione di knowledge graph aperti e interoperabili, particolarmente adatti alla rappresentazione di entità, eventi e relazioni storiche. L'uso di identificatori persistenti e modelli semantici condivisi favorisce la riusabilità dei dati e la loro integrazione nel web semantico. È essenzialmente in virtù di queste caratteristiche e per la possibilità di armonizzare diversi modelli dati, mantenendo ampia flessibilità nella scelta delle informazioni rilevanti che differenti progetti nell'ambito delle SSH hanno optato per l'utilizzo di Wikibase (Koho, 2023, Zapala, 2023 e Lozana Rossenova, 2023, che segnala anche alcune difficoltà connesse con l'integrazione di dati da Wikidata).
A queste piattaforme si affiancano metodologie che mirano a incorporare la FAIRness fin dalla progettazione. La metodologia FAIR-by-Design di Skills4EOSC[4] propone un approccio sistematico (di non sempre facile implementazione) per integrare i principi FAIR nel ciclo di vita dei dati, dai modelli concettuali ai workflow di produzione. Questo orientamento è particolarmente rilevante per le discipline umanistiche, caratterizzate da una grande varietà di formati e pratiche.

3. RATIO: FUNZIONI, COMPONENTI E ARCHITETTURA

La piattaforma RATIO consente la gestione di collezioni di risorse digitali attraverso la descrizione, l'archiviazione e il supporto a diverse tipologie di oggetti e dei loro raggruppamenti ed è strutturato attorno a tre servizi chiave, parzialmente indipendenti e ciascuno dotato di front-end e back-end, che ne implementano le macro-funzioni e la cui interconnessione determina l'architettura della piattaforma nel suo complesso. I tre componenti di RATIO sono:
1. Strumento di FAIRificazione (FAST) – Mette a disposizione una serie di funzionalità progettate per guidare gli utenti nella descrizione dei prodotti delle loro attività di ricerca aiutandoli ad allineare i propri dati ai principi FAIR. Questo strumento supporta l'arricchimento dei metadati, gli identificatori persistenti, le licenze aperte e l'esportazione in formati standardizzati per promuovere l'apertura, la trasparenza e la riutilizzabilità a lungo termine. Consente l'importazione e la creazione ex novo di set di metadati e propone una valutazione della loro FAIRness, con suggerimenti integrati su come aumentarne il livello.

[2] Immagine generata con IA
[3] https://www.skills4eosc.eu/resources/fair-by-design-methodology (cons. 31/01/2026)
[4] https://factgrid.de/ (cons. 31/01/2026)

Lo strumento è stato progettato con schemi TEI-XML personalizzati su misura per ciascuno dei tipi di risorse che gestisce, garantendo coerenza e aderenza col suddetto standard. Le strutture di tutti gli schemi sono state sottoposte a controllo di validazione durante la fase di test.
La valutazione della FAIRness viene effettuata in base a una serie di criteri parzialmente dipendenti dall'oggetto, che tuttavia condividono una logica comune. Le metriche utilizzate sono un sottoinsieme di quelle stabilite dal FAIR Data Maturity Model Working Group della Research Data Alliance (FAIR Data Maturity Model Working Group, 2020). Ove possibile, viene indicata l'equivalenza con le metriche FAIRisFAIR (Devaraju, 2020 e versioni successive fino a Hubert, 2025). Le metriche sono collegate a un percorso specifico negli schemi TEI, fornendo una misurazione FAIR univoca per ogni tipo di risorsa, nonché un quadro generale comune che viene dettagliato in base alle peculiarità di ciascun oggetto. Il livello di FAIRness viene quantificato in percentuale e rappresentato graficamente con una visualizzazione che distingue chiaramente il contributo dei singoli principi FAIR, permettendo di cogliere immediatamente punti di forza e aree da migliorare (Fig. 1).

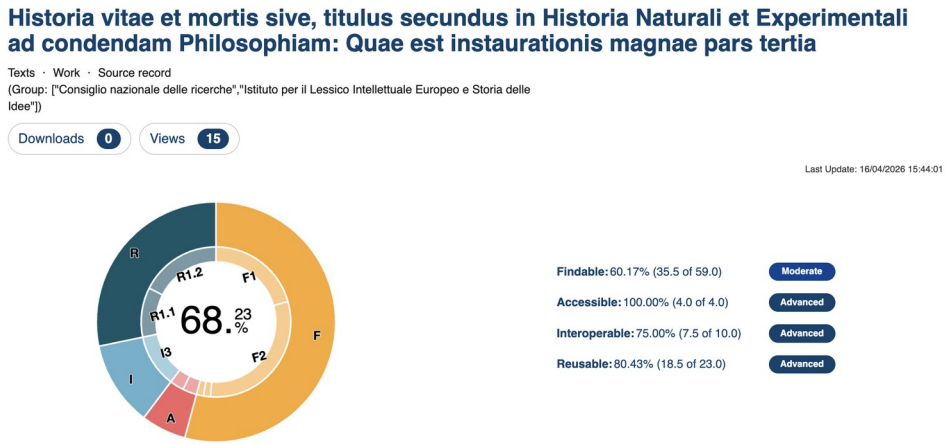


Figura 1. Rappresentazione grafica del livello di FAIRness di un record in FAST.

2. Ambiente Wikibase (CONNECT): Un ambiente semantico flessibile per la gestione di metadati strutturati, dati di autorità e relazioni complesse tra entità, particolarmente adatto per set di dati biografici e prosopografici. Supporta etichette multilingue, principi di dati aperti collegati e integrazione con repository e ontologie esterni. L'ambiente Wikibase è destinato alla creazione, alla gestione e all'hosting di set di dati per vari progetti di ricerca che utilizzano la piattaforma. L'istanza Wikibase è pensata per consentire la descrizione di tutte le classi di oggetti (entità), sia gestiti e archiviati direttamente sulla piattaforma (ad esempio trascrizioni digitali), sia semplicemente citati (testi stampati, persone, istituzioni, progetti, collezioni, ecc.). La scelta della tecnologia Wikibase è stata determinata principalmente dalla sua flessibilità e dal formato altamente generalizzato di descrizione delle informazioni che fornisce, e in secondo luogo dalla sua natura parzialmente autodescrittiva (cfr. infra) e dalla possibilità nativa di implementare facilmente i principi FAIR.

3. Ambiente di trascrizione (CONNECT-TED): Uno spazio di lavoro per la trascrizione, l'annotazione e la codifica di testi sia preesistenti su supporti fisici, sia nati in formato digitale. Supportando flussi di lavoro collaborativi, consente il tagging semantico, il controllo delle versioni e l'allineamento con gli standard esistenti (come TEI e IIIF), garantendo risorse testuali di alta qualità e interoperabili.
Sebbene ciascuna di queste parti abbia un proprio livello di autonomia, l'obiettivo dell'integrazione è quello di incorporare tutte le funzionalità necessarie per la creazione e la gestione di collezioni digitali aperte. Ogni componente aggiunge una caratteristica fondamentale alla suite, costituendo un flusso di lavoro ideale per la creazione di collezioni, pur servendo al contempo al proprio scopo specifico e autonomo. In generale, dunque, la funzione specifica di ciascun componente può essere realizzata in modo autonomo; tuttavia, la gestione completa della collezione richiede l'integrazione di tutti i componenti (Fig.2).

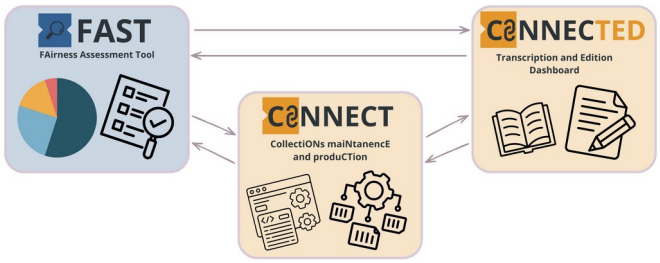


Figura 2. Interazione tra le componenti di RATIO.

Lo strumento di FAIRificazione è il principale punto di accesso ed è responsabile dell'assegnazione dell'ID univoco RATIO all'oggetto creato. I metadati pubblicati in FAST possono essere collegati automaticamente alla piattaforma CONNECT, dove vengono archiviati. Qui i metadati vengono gestiti e possono essere arricchiti in base all'obiettivo specifico del progetto, al di là dei requisiti minimi richiesti dallo strumento FAST. I metadati possono essere eventualmente aggiornati tramite CONNECT dal momento che l'allineamento tra i due strumenti è bidirezionale.
FAST consente anche il caricamento di file di dati (testi, immagini, etc.) accompagnati da un set di metadati che vengono inviati all'ambiente di trascrizione e lì archiviati. Se i file sono in uno dei formati supportati possono inoltre essere gestiti e visualizzati direttamente. Durante il processo di importazione, i dati archiviati o gestiti nell'ambiente di trascrizione vengono automaticamente collegati agli elementi (Items) corrispondenti in CONNECT. A ogni elemento corrisponde una pagina dove vengono raccolti tutti i riferimenti presenti agli oggetti correlati. Ad esempio, una trascrizione viene collegata alla sua descrizione, che a sua volta contiene riferimenti alla fonte a stampa, all'autore del testo e al progetto di ricerca di riferimento. Gli utenti possono decidere se un oggetto di riferimento debba essere trattato come un elemento, e quindi avere i propri metadati, oppure come una semplice etichetta. Entrambe le tipologie vengono indicizzate e dunque sono ricercabili e utilizzabili per la navigazione tra le risorse. In questo modo, l'ambiente di trascrizione integra nella propria visualizzazione gli elementi rilevanti di CONNECT, offrendo una visualizzazione integrata di dati e metadati. Allo stesso modo, ogni elemento di CONNECT contiene un riferimento ai file di dati (testi, immagini, etc.) pertinenti presenti nell'ambiente di trascrizione, se ve ne sono. L'architettura è stata progettata a partire dall'analisi dei comportamenti richiesti in due diversi scenari d'uso principali che definiscono le funzioni chiave della piattaforma in relazione alla FAIRness e, più in generale, alla coerenza con i principi dell'Open Science:

1) FAIRificazione dei dati pre-esistenti. Il flusso di lavoro per la definizione dell'architettura del sistema si è basato sulle risorse esistenti nell'archivio digitale del CNR-ILIESI. Queste si sono rivelate un buon caso di studio perché, nonostante siano quasi esclusivamente collezioni di oggetti testuali, presentano sfide diverse per quanto riguarda la FAIRificazione. Essendosi accumulate nel corso degli anni, sono disponibili in diversi formati e standard di codifica, talvolta interni all'Istituto, nonché in diversi sistemi, livelli di completezza e descrizioni dei metadati.
2) Generazione di dati FAIR da zero. Il flusso di lavoro per questa seconda opzione è stato costruito sulla base dei progetti che saranno ospitati sulla piattaforma[5].
Attraverso l'analisi di questi due casi, il sistema è stato progettato in modo da avere un flusso di immissione dati unico, al fine di garantire l'unicità degli ID interni. Allo stesso tempo, è sempre possibile aggiornare i dati e i metadati in qualsiasi componente del sistema, e questi aggiornamenti vengono riconosciuti e registrati da tutti i componenti attraverso flussi bidirezionali.

4. STRUTTURA DEI DATI

RATIO ambisce a servire un'ampia gamma di utenti accademici, in particolare nelle scienze umane e a consentire, oltre alla conservazione e gestione di specifiche risorse, di integrare diversi tipi di dati al fine di supportare lo studio delle reti intellettuali, la ricostruzione di contesti storici e l'analisi della trasmissione delle idee.

[5] Segnatamente: l'Archivio Biografico della Cultura Scientifca Italiana (www.iliesi.cnr.it/abcsi.php); Leibniz correspondents and aquaintances (www.leibnitiana.eu) ed il Lessico digitale dello slavo ecclesiastico.

La piattaforma è progettata per gestire una varietà di oggetti digitali, fornendo schemi di metadati prestabiliti che descrivono sia gli oggetti digitali stessi, sia gli oggetti fisici a cui si riferiscono. Questi schemi sono stati progettati per garantire la conformità ai principi FAIR e per essere il più generici possibile, in modo da adattarsi a un'ampia gamma di progetti di ricerca. In quest'ottica, si è ritenuto di fornire strumenti per arricchire le descrizioni in base agli scopi specifici dei vari progetti predisponendo la serie di informazioni che, per ogni tipo di oggetto, ne consentano l'identificazione univoca, cercando un equilibrio tra approccio open world e pre-definizione dei dati, tenendo in considerazione, per quanto possibile a un singolo progetto, alcune delle osservazioni sviluppate in Tóth-Czifra (2020). La mediazione tra l'esigenza di personalizzazione (parziale) dei modelli di dati con descrizioni goal-oriented e l'inserimento dei dati in un quadro comune coerente con il quadro semantico del progetto H2IOSC, che a sua volta contribuisce a rendere possibile la condivisione di dati tra progetti differenti, è stata ottenuta grazie alla struttura altamente generalizzata del meta-modello di dati in Wikibase, in cui forme o gerarchie non sono fisse, ma possono essere derivate solo dalla relazione tra le proprietà e tra gli elementi. A giustificare la scelta hanno inoltre contribuito: la possibilità di presentare sia i dati che alcune delle loro proprietà in forma strutturata e univoca attraverso qualificatori o riferimenti e la possibilità di specificare e ordinare più valori per una proprietà. Quest'ultima caratteristica in particolare facilita il trattamento dei dati contrastanti nelle fonti, cosa piuttosto comune nei progetti a carattere storico, dando la possibilità di identificare un valore preferito senza perdere informazioni, permettendo anche a progetti diversi di compiere scelte differenti.
Per consentire tanto la condivisione quanto la personalizzazione, è stato costruito un modello di dati a più livelli, suddividendo secondo regole prestabilite le informazioni presenti negli schemi TEI generati da FAST su entità differenti della piattaforma CONNECT. Il modello include un insieme di entità di primo livello ("persona" o "opera"), che sono collegate tramite proprietà personalizzabili ad altre entità di primo livello ("progetto"), con cui condividono una relazione specifica ("è responsabile di"), o a entità di livello superiore che le descrivono ("scheda biografica", "testo a stampa"). Il motivo di questa struttura è che progetti diversi, come ha dimostrato l'analisi dei casi d'uso, possono voler descrivere gli stessi oggetti, evidenziando aspetti diversi e utilizzando quindi insiemi di proprietà diversi o fare scelte diverse nei casi in cui nelle fonti siano attestati dati contrastanti (Fig. 3).

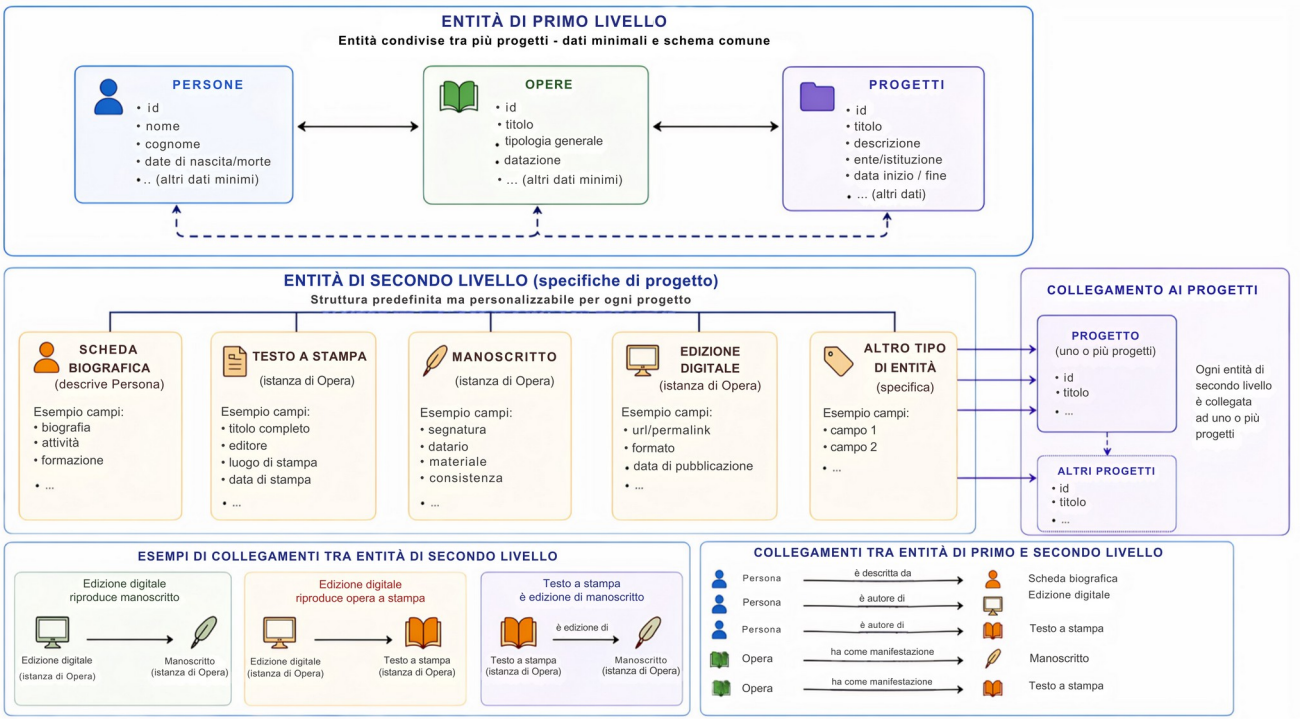


Figura 3. Rappresentazione delle interazioni tra entità di primo e secondo livello in CONNECT[6].

[6] https://www.f-uji.net/ (cons. 31/01/2026)

Inoltre, il modello a livelli consente di differenziare i permessi degli utenti riferendoli a elementi descritti come appartenenti a un progetto specifico, garantendo così che le modifiche dei dati non abbiano un'eredità immediata tra i progetti.
La categoria principale di oggetti digitali sulla piattaforma è costituita da edizioni digitali di oggetti testuali, comprese trascrizioni di fonti e testi di nuova generazione, come biografie in raccolte biografiche o descrizioni di progetti di ricerca. Gli schemi di metadati proposti mirano a descrivere gli elementi salienti delle edizioni digitali, tra cui la data di pubblicazione, gli identificatori (ad esempio DOI) e la codifica. Essi definiscono anche i possibili ruoli relativi al testo, come autore, editore o traduttore. Gli schemi consentono di descrivere la fonte o le fonti dell'edizione digitale, con una differenziazione basata sul tipo di supporto testuale (edizione stampata o manoscritto). Ciò consente di identificare le informazioni rilevanti (ad esempio, editore, luogo e anno di stampa per un'edizione stampata; luogo di conservazione, firma e composizione del codice per un manoscritto). Gli schemi sono specificati in modo più dettagliato per consentire la gestione di casi specifici, come documenti d'archivio o articoli di riviste. In generale, i tipi sono classificati in base ai metadati necessari per una descrizione sufficientemente completa. Questo approccio è stato preferito al tentativo di applicare una tassonomia predeterminata basata su criteri di classificazione predefiniti. Poiché l'obiettivo è quello di fornire descrizioni di oggetti fisici o digitali che costituiscono fonti, i tipi di testo non sono generalmente inclusi nella parte predefinita degli schemi di metadati. Tuttavia, sono previste due eccezioni: le lettere, poiché è stato ritenuto utile consentire l'inclusione di informazioni specifiche quali il mittente e il destinatario, la data di spedizione ed eventuali allegati; e dizionari ed enciclopedie, per consentire di utilizzarli come fonti per la costruzione di raccolte lessicali indicando informazioni aggiuntive quali il numero di voci e l'argomento. In ogni caso, gli utenti hanno la possibilità di descrivere il contenuto dei testi o specificare ulteriormente la natura del supporto materiale utilizzando proprietà specifiche nell'ambiente CONNECT.
L'altra categoria principale di oggetti non digitali che possono essere descritti utilizzando schemi di metadati è quella delle persone. Queste possono essere presenti come entità primarie in raccolte specifiche (biografiche o prosopografiche) o come oggetti con relazioni con altri oggetti primari (ad esempio, “essere l'autore di” un testo). A causa di questo doppio ruolo degli “oggetti-persona” si è deciso di dare comunque agli utenti la possibilità di generare una voce biografica quando una persona è collegata a un testo in un ruolo specifico (ad esempio come traduttore). Se l'utente sceglie di non farlo, il sistema gestirà il nome della persona come etichetta e garantirà che sia presente nelle funzioni di ricerca. Qualora scelga di farlo, viene generato uno schema contenente gli elementi necessari per identificare in modo univoco la persona in questione, pur consentendo specifiche personalizzazioni.
La piattaforma può supportare altri tipi di oggetti digitali, in particolare immagini. Queste possono essere gestite sia all'interno di CONNECT-TED — nel qual caso sono descritte come fonti dell'edizione — sia come elementi autonomi di una collezione. In quest'ultimo caso, hanno un proprio schema di metadati a fini identificativi.
Oltre alle risorse testuali, i metadati della piattaforma includono uno schema di descrizione del progetto che specifica le istituzioni coinvolte, i ruoli e le funzioni dei partecipanti e gli eventuali enti finanziatori. In RATIO, un progetto è dunque un oggetto unico tra gli altri, sia all'interno di FAST che di CONNECT, con una propria descrizione e un proprio insieme di metadati, comprensivo delle informazioni solitamente obbligatorie per i progetti finanziati da istituzioni. Al contempo, dal punto di vista del modello di dati, il progetto è lo “strumento” principale che integra le diverse parti della piattaforma e il mezzo principale per raggruppare gli oggetti in una collezione (o in un insieme più ampio) mettendo in relazione diverse entità con un unico progetto di ricerca. Ferma restando la possibilità di definire autonomamente sottoinsiemi di risorse, il riferimento a (almeno) un progetto è obbligatorio per ogni oggetto e per la gestione di ogni collezione perché consente di definire restrizioni di accesso e livelli di autorizzazione, evitando così interferenze tra i progetti.
L’insieme delle funzioni qui delineate mira, dunque, a garantire la gestione in tutte le sue fasi di collezioni digitali anche composte da tipologie di oggetti differenti, consentendo la coesistenza di progetti diversi sulla medesima piattaforma. Centrale per questa consistenza è l’equilibrio che si è cercato di trovare tra l’autonomia dei singoli progetti, con la possibilità di implementare modelli personalizzati e gestire in autonomia i propri dati e la condivisione ed il riutilizzo delle informazioni e dei dati esistenti garantita dalla strutturazione di un modello a più livelli e dalla definizione di un nucleo minimale di dati comuni.

RINGRAZIAMENTI

RATIO è stato sviluppato all'interno del progetto H2IOSC - Humanities and cultural Heritage Italian Open Science Cloud finanziato dall'Unione Europea – NextGenerationEU – NRRP M4C2 - Codice progetto IR0000029 - CUP B63C22000730005. Si ringraziano per il lavoro svolto i rappresentanti delle aziende coinvolte nello sviluppo di RATIO: per LINKS Management and Technology, Alessandra Caggiano, Flavio D'Asaro, Lorenza Di Lallo, Alfredo Serafini; per Net7, Nicola Baglini, Luca De Santis, Marco Zizi.

BIBLIOGRAFIA

Devaraju, A., Mokrane, M., Herterich, P., de Vries, J., Davidson, J., Huber, R., & Cepinskas, L. (2020). FAIRsFAIR Data Object Assessment Metrics (0.2). https://doi.org/10.5281/zenodo.3775794.

FAIR Data Maturity Model Working Group. (2020). FAIR Data Maturity Model. Specification and Guidelines (1.0). Zenodo. https://doi.org/10.15497/rda00050.

Huber, R. (2025). FAIRsFAIR Data Object Assessment Metrics (0.8). https://doi.org/10.5281/zenodo.15045911.

Koho, M., Coladangelo, L. P., Ransom, L., & Emery, D. (2023). Wikibase model for premodern manuscript metadata harmonization, linked data integration, and discovery. ACM Journal on Computing and Cultural Heritage, 16(3), 1-25. https://doi.org/10.1145/3594723.

Rossenova, L., Duchesne, P., & Blümel, I. (2023). Wikidata and Wikibase as complementary research data management services for cultural heritage data. Wikidata 2022 : Wikidata Workshop 2022, Proceedings of the 3rd Wikidata Workshop 2022 co-located with the 21st International Semantic Web Conference, Kaffee L. A., Razniewski S., Amaral G., Alghamdi, K.S.. https://doi.org/10.25968/opus-2573.

Sichera, P., Marras, C., & Pasini, E. (2024). Orchestrazione API per workflow applicativi nell'ambito delle Digital Humanities. https://doi.org/10.5281/zenodo.14187534.

Tóth-Czifra, E. (2020). The risk of losing the thick description: Data management challenges faced by the arts and humanities in the evolving FAIR data ecosystem. Digital technology and the practices of humanities research, Edmond J (eds) Cambridge: Open Book Publishers, 235-266. https://doi.org/10.11647/OBP.0192.10.

Zapała, A., & Królik, T. (2023). Wikibase as an environment for harmonisation of data about the past: The example of WikiHum. DARIAH Annual Event 2023: Cultural Heritage Data as Humanities Research Data?, Budapest, Hungary. https://doi.org/10.5281/zenodo.8009577.

Making Uncertainty Explicit: The Preservation of Interpretive Complexity across Digital Humanities Project Workflows

TOMMASO BATTISTI[1], VALENTINA PASQUAL[2], GIULIA RENDA[3], MARILENA DAQUINO[4]

[1] DEPARTMENT OF CLASSICAL PHILOLOGY AND ITALIAN STUDIES, UNIVERSITY OF BOLOGNA, ITALY - TOMMASO.BATTISTI5@UNIBO.IT
[2] DEPARTMENT OF CLASSICAL PHILOLOGY AND ITALIAN STUDIES, UNIVERSITY OF BOLOGNA, ITALY - VALENTINA.PASQUAL2@UNIBO.IT
[3] DEPARTMENT OF CLASSICAL PHILOLOGY AND ITALIAN STUDIES, UNIVERSITY OF BOLOGNA, ITALY – GIULIA.RENDA3@UNIBO.IT
[4] DEPARTMENT OF CLASSICAL PHILOLOGY AND ITALIAN STUDIES, UNIVERSITY OF BOLOGNA, ITALY - MARILENA.DAQUINO2@UNIBO.IT

ABSTRACT (ENGLISH)

This contribution examines how the interpretive nature of humanities knowledge manifests and propagates through workflows in Digital Humanities projects. Focusing on three critical phases—data acquisition, knowledge representation, and web-based dissemination—we analyse selected case studies to reveal how interpretive phenomena are preserved, negotiated, simplified, or omitted at each stage. Our analysis reveals a systematic gap between the rich knowledge preserved in data models and that communicated through dissemination interfaces, where the rhetorical appeal of an "aesthetics of certainty" often overrides scholarly transparency.
Keywords: uncertainty; semantic web; data acquisition; knowledge representation; web-based dissemination

ABSTRACT (ITALIANO)

Esplicitare l'incertezza: la preservazione della complessità interpretativa nei flussi di lavoro dei progetti di Digital Humanities. Questo contributo analizza come la natura interpretativa della conoscenza umanistica si manifesti e si propaghi nel corso dello sviluppo di progetti di Digital Humanities. Concentrandoci su tre fasi critiche, ovvero la creazione dei dati, la rappresentazione della conoscenza e la disseminazione sul web, esaminiamo come i fenomeni interpretativi vengano preservati, negoziati, semplificati od omessi in una selezione di casi di studio. L'analisi mette in luce un divario tra la ricchezza di conoscenza preservata durante la modellazione dei dati e quella comunicata attraverso le interfacce di disseminazione, dove l'attrattiva retorica di un'"estetica della certezza" tende spesso a prevalere sulla trasparenza scientifica.
Parole chiave: incertezza; web semantico; acquisizione dei dati; rappresentazione della conoscenza; disseminazione web.

1. INTRODUCTION

Humanistic knowledge is often partial, evolving, or shaped by subjective perspectives. These conditions give rise to epistemic uncertainties—defined as the lack of definitive scholarly knowledge—that stand in contrast to the purely stochastic or algorithmic uncertainty of computational processes (Therón Sánchez et al., 2019). Consequently, imprecision, disagreement, credibility issues, ignorance, and data incompleteness introduce uncertainty into Digital Humanities (DH) projects at multiple stages. Depending on design decisions, uncertainty may be reduced, ignored, propagated, or even amplified (Therón Sánchez et al., 2019; Windhager, Salisu, & Mayr, 2019; Windhager, Salisu, Schreder, et al., 2019; Panagiotidou et al., 2022; Conroy et al., 2024; Windhager et al., 2024).

From the moment data are created, each stage of a DH workflow involves acts of interpretation—such as processes of selection, division, and classification—that embed particular perspectives and negotiate ambiguities, conflicts, and contradictions (Drucker, 2011; Drucker, 2014; Gitelman, 2013; D'Ignazio and Klein, 2020; Piotrowski, 2023). Modelling and communicating uncertainty have been shown to support users' trust and perceived transparency among expert users (Fisher et al., 2012; Windhager, Salisu, Schreder, et al., 2019), since this ensures methodological reliability by explicitly acknowledging the adopted perspective (Berg et al., 2018).

This contribution starts from the assumption that uncertainty is fundamental to the humanities and should therefore be treated as an asset (Piotrowski, 2023), requiring the design of appropriate

methodologies to preserve it throughout the data acquisition, knowledge representation, and dissemination phases of a DH project workflow. Despite growing theoretical recognition, the preservation of epistemic uncertainty and interpretive phenomena is often insufficiently addressed in DH practice, both in modelling and in dissemination outcomes (Barabucci et al., 2021; Battisti and Daquino, 2025a).
In this contribution, we present a selection of research projects that explicitly acknowledge uncertainty as an integral component of their DH workflows, in order to extract design policies and methods that drive the phases of data acquisition, knowledge representation, and web-based dissemination. The projects exemplify different research domains on heterogeneous cultural objects and historical periods and are used to analyse how uncertainty emerges, is propagated, and is managed across the DH workflow.

2. BACKGROUND AND STATE OF THE ART

Barabucci et al. identify several recurring problems affecting the way cultural-heritage metadata are originally created, namely: **reticence**, the omission of well-known information due to haste, limited skills, or inadequate policies and tools; **flattening**, where complex, multi-layered entities are collapsed into a single, inconsistent record; **coercion**, the forcing of information into unsuitable fields when no appropriate schema elements exist; and **dumping**, where important details are relegated to free-text descriptions, making them inaccessible to computational processing (Barabucci et al., 2021). Moreover, varying reliability, inherent subjectivity, and varying degrees of confidence in scholars' claims are well-known limitations that need to be harmonised, normalised, or resolved when devising a structured dataset (e.g., a knowledge graph). Over the years, Cultural Heritage (CH) domain standards have addressed this multifaceted issue with different modelling approaches, such as the introduction of n-ary relations (e.g., the attribute assignment construct) in CIDOC CRM for museum heritage (Doerr et al., 2007), or through the definition of specific ontology extensions (e.g., CRM-inf is the CIDOC CRM Argumentation Model that formalises scholars' inferences and observations as first-class entities, distinguishing between the historical event or entity and the interpretive act that asserts its existence or properties (Doerr et al., 2023)), or by defining explicitly uncertain properties such as in RiC-O (e.g., rico:relationCertainty) (Clavaud & Wildi, 2021).
Alongside domain-specific ontologies, the Semantic Web community recognises the limitations of encoding statements as objective, universal truths, and proposes a range of ontology-independent solutions for representing assertions and their provenance in RDF. These approaches enrich RDF statements with meta-assertions—such as authorship or source—while avoiding the proliferation of specialised vocabularies. Proposed solutions include standard RDF reification, singleton properties, Named Graphs and models for organising them, alternative serialisations such as Notation3, and extensions to the graph model, including RDF-star and Conjectures (Daquino et al., 2022; Barabucci et al., 2021).
In practice, large-scale collaborative platforms such as Wikidata—where disagreement, uncertainty, vagueness, and subjectivity occur across many scenarios—adopt mixed approaches combining ontology-dependent and independent solutions. Despite such modelling efforts, empirical studies show that annotators use uncertainty-related representation patterns sparingly (less than 1% of annotations on CH data in Wikidata) (Di Pasquale et al., 2024), which reveals a tendency to oversimplify the interpretive nature of humanistic knowledge for the sake of dissemination.
While interfaces for experts tend to be more inquiry-oriented, interfaces for casual users usually provide more engagement-oriented experiences that do not retain complex, potentially misleading, scholarly argumentations (Windhager, Salisu, Schreder, et al., 2019). However, interfaces shape the ways CH is perceived and interpreted by different users and communities (Dörk et al., 2011; Whitelaw, 2015; Windhager, Federico, et al., 2019; Drucker, 2013) which too often prioritise task efficiency within a "user-as-customer" paradigm rather than fostering interpretive engagement, thereby giving the illusion of certainty, completeness, and singularity of meaning (Drucker, 2013). Generous interfaces were proposed to challenge this paradigm, foregrounding exploration, interpretation, and sense-making as central activities, aligning with the epistemic characteristics of humanistic enquiry (Whitelaw, 2015). Relatedly, making the evidential status of represented information explicit—distinguishing evidence, hypotheses, and degrees of probability—is key for CH visualisations, aligning them with research practices and supporting proper user interpretation and evaluation (Denard, 2016). Such an approach can be observed, for instance, in domain-specific implementations, as in the case of bookbinding and codicological modelling (Campagnolo, 2015; Porter & Campagnolo, 2021), where structured descriptions

are transformed into uncertainty-aware diagrammatic representations recognising the complexity of humanities scholarly discourse.
In this sense, these practices resonate with broader perspectives in information visualisation and storytelling, which are argued to enhance the communication of complex information by offering an additional interpretive layer and producing memorable and engaging experiences (Gershon & Page, 2001; Kosara & Mackinlay, 2013; Shan et al., 2022; Renda et al., 2023), while promoting slower and more reflective experiences (Hinrichs et al., 2016; Elli et al., 2021)—for example, showing complexities on-demand to interested users only (Windhager, Salisu, & Mayr, 2019) was proposed to overcome the risk of cognitive overload for non-expert users due to the introduction of interpretive complexity in visualisation (Mayr et al., 2016; Windhager et al., 2018; Windhager, Salisu, Schreder, et al., 2019).
It has been argued that the scarcity of solutions and standard ways to visually communicate uncertainty with existing visualisation tools (Khulusi et al., 2020; Panagiotidou et al., 2022), and the lack of visualisation expertise, could be the reasons hindering the adoption of more critical approaches to present the interpretive nature of humanities data (Ma & Xiao, 2021). Indeed, surveys on web-based DH visualisation projects revealed that visualisation and narrative strategies often rely on generic, rarely customised, off-the-shelf solutions that oversimplify information rather than revealing its inherent complexity (Battisti & Daquino, 2025b).
While individual stages of the DH workflow (acquisition, representation and dissemination) have been studied in isolation, there is a notable lack of research tracking the continuity and propagation of uncertainty across the entire pipeline. With this work, we address this gap by analysing how uncertainty is negotiated at each stage in selected case studies, examining if and how dissemination interfaces maintain the interpretive awareness established during the data creation and representation phases.

3. CASE STUDIES

Inspired by existing conceptual frameworks that examine uncertainty propagation in humanities contexts (Therón Sánchez et al., 2019; Panagiotidou et al., 2022; Haghighatkhah et al., 2022), we distinguish three key phases that characterise typical DH project structures, namely **data acquisition**, **knowledge representation**, and **web-based dissemination**. The selected cases are examples of a broader landscape rather than an attempt to create a complete survey. In fact, it is a collection of projects reflecting the evolution of the research line at hand within the Digital Humanities Advanced Research Centre (DH.arc) of the University of Bologna over the past decade. They are presented in chronological order and share epistemic uncertainty as their foundational concern across the three aforementioned phases, yet vary in their application domains (e.g., art history, digital philology) and dissemination formats (e.g., catalogues, digital editions, and data analysis projects). Selection criteria include the following aspects: 1) original data are extracted from authoritative although subjective sources (e.g. literary works, institutional catalogues), 2) data are transformed into RDF according to one or more strategies that try to preserve the interpretive nature of original data (whether this limits to preserve subjectivity or expands on other features like ambiguity, uncertainty, and disagreement), and 3) the final graph is disseminated via web interfaces for heterogeneous audiences, hence adopting diverse solutions to mitigate the complexity of data.
The "Paolo Bufalini's Notebook" digital edition[1] uses Linked Open Data (LOD) technologies to enhance a semantic scholarly edition of the notebook, in which the author collates quotations, translations and comments (his or others') which implicitly debate about the influence among literary authors (Daquino et al., 2019). Data acquisition included manual transcription, identification and disambiguation of over 250 literary quotations and their sources, and the textual encoding using XML/TEI. Knowledge representation uses the Nanopublication meta-model to explicitly represent the interpretive nature of Bufalini's assertions, specifying provenance (e.g. whether an assertion belongs to Bufalini or to a quoted author) and reusing the HiCo and CiTo ontologies (e.g., source texts, interpretive criteria). The web application shows the interpretive layers underlying the transcription and encoding phases in the form of facets, which allow users to retrieve all the occurrences of a person or work recognised in the text (even when implicit), and in the facsimile, where annotations appear on hover, revealing the (implicit) sources of a quotation. However, all information is presented as factual statements, while the subjectivity and the degree of certainty (which is not modelled) remain implicit.

[1] https://projects.dharc.unibo.it/bufalini-notebook/.

mythLOD[2] is a digital collection that semantically links museum objects depicting classical mythology subjects to the literary sources sharing common themes (Pasqual & Tomasi, 2022). Uncertainty arises from subjective interpretive acts and is explicitly modelled using the Digital Hermeneutics approach (Daquino et al., 2020), including FRBRoo as the backbone ontology, Nanopublications to preserve the provenance of individual scholarly assertions, and PROV-O and HiCo to contextualise the hermeneutic information. The platform provides a browsable catalogue with colour-coded tags and filters to visually distinguish (questionable) scholarly claims from factual data. A storytelling section with interactive visualisations conveys quantitative aspects. However, visual components are off-the-shelf data visualisations that, *per se*, cannot fully differentiate between asserted interpretations and established facts, therefore proposing a simplified view on the data.

The Iconology dataset[3] is a Linked dataset that contains a selection of iconological interpretations over modern and Middle Ages artworks, manually extracted from Erwin Panofsky's works. Iconographic and iconological attributions are formally represented using the ICON ontology (Sartini et al., 2023) and n-ary relations as a reification method to represent meaning, subjects, and interpretive acts. Notably, the ontology allows to represent claims according to Panofsky's three levels of interpretation (visual recognition, iconographic interpretation and iconological interpretation). The ICON dataset supports interdisciplinary research into how meaning and interpretation are constructed and analysed within art history. The dissemination platform is intended to support data analysis tasks by combining project documentation, exploratory data analysis, visualisations, and an artwork gallery. Visualisations enable quantitative analysis of uncertainty-related features—such as the frequency of author's interpretation types or levels—treating interpretation and subjectivity as first-class citizens. The media gallery works as a filter on depicted subjects across the model's three interpretive levels while preserving the distinction between conflicting scholarly readings.

Broast[4] leverages Heider's documentary index to collect the scholarly debate on the authenticity of damaged, tampered-with, or altered medieval documents from Upper Austria (Haider, 2022). Uncertainty arises from incomplete or ambiguous information, such as uncertain production dates, and from the coexistence of conflicting scholarly opinions about the documents' authenticity over time. To this end, the Conjectures framework (Barabucci et al. 2021), an extension of the Named Graphs syntax that expresses scholarly opinions without asserting their content, was applied to formally represent scholarly claims as undisputed, disputed, or settled, along with evidence, provenance, and bibliographic context. The web platform allows the exploration of the evolution of the scholarly debate. It presents quantitative reports where interpretive aspects are the subject (e.g., mapping the locations of authenticity-debated documents) rather than emerging as a "perspective" on data. The catalogue can be filtered by type, creator, date, and document coverage. Here, the interpretive layer is preserved through sub-filters that allow perspectives that distinguish currently disputed, debated, and settled information (Pasqual, 2025).

artresearch.net[5] is a discovery platform that brings together the collections of the International Association of Photo Archives (PHAROS), providing access to millions of photographs of artworks. Records from seven institutions are enriched and reconciled using different approaches (manual annotation, computer vision, LOD). Uncertainty and subjectivity are preserved in the aggregated dataset, thanks to ontology-dependent methods (including application profiles of CIDOC-CRM and non-committal properties instead of equivalence assertions) to record conflicting attributions and unknown or ambiguous identities, which may entail disagreement in the reconciliation with shared authorities (Daquino et al., 2026) and Named Graphs to preserve provenance information. The platform supports cross-collection research, highlighting contrasting and multi-source information and image comparison. It provides text- and image-based search similarity capabilities with adjustable sensitivity, enabling both targeted lookup of known items and exploratory discovery through broader semantic and visual matches. Faceted filters include cascading terms to visually simplify vagueness and imprecision using a hierarchical organisation (from broader, vague terms to specialised, precise ones). Data provenance is highlighted with institutional logos, supporting comparison. So far, no information visualisation technique has been employed to address further epistemic uncertainties.

[2] https://dharc-org.github.io/mythlod/.
[3] https://iconology-dataset.streamlit.app/.
[4] https://projects.dharc.unibo.it/broast/.
[5] https://artresearch.net.

4. DISCUSSION AND CONCLUSION

The case studies present some recurring patterns in the management of uncertainty.
In data acquisition, uncertainty manifests as 1) implicit knowledge nearly always manually interpreted and extracted, 2) disputed assertions, either automatically retrieved (e.g. artwork attributions) or defined via policies (e.g. for reconciliation purposes). Across cases, uncertainty is handled through explicit manual annotation of interpretive layers (e.g. Bufalini, ICON), enrichment of pre-existing sources with additional evidence and provenance (e.g. mythLOD), and harmonisation of heterogeneous and even conflicting claims within the same (e.g. Broast) or across different datasets (e.g., artresearch.net). This variety of approaches confirms the need to treat interpretive subjectivity as a first-class entity from the earliest stages of data creation in DH projects, regardless of their scale.
Knowledge representation strategies in place are still scattered. Reification approaches (n-ary in ICON and Nanopublications in mythLOD and Bufalini), hybrid strategies combining n-ary and Named Graphs (artresearch.net), or RDF vocabulary extensions (Conjectures in Broast) tend to build on top of community standards (e.g., CIDOC-CRM). However, these notable solutions significantly impact data complexity and performance, affecting their wide adoption, and implying a steeper learning curve for "re-users" who have to deal with different, non-standard solutions to perform integrated queries. We argue that the development of shared, lightweight, reusable patterns to balance high expressivity with usability is fundamental to the DH international agenda, in order to effectively lower technical barriers for humanities scholars.
The study confirms the systematic gap between the importance given to the interpretive nature of humanities data in the scholarly literature and its actual implementation in web-based environments for scholars. Epistemic uncertainties seem to be better addressed through customised interface solutions, explicitly designed to enable interaction with the content based on its interpretive aspects (e.g., artresearch.net, Broast, ICON). When available, data visualisations present interpretive phenomena as a variable of the representation (ICON) or do not represent them at all. We believe that a shift toward layered engagement modalities is needed to encode interpretation, if not as a first-class variable, as a feature on demand (Windhager, Salisu, Schreder, et al., 2019), which we see not only as a mitigation strategy but also as a way to reveal qualitative aspects emerging from the data, thereby supporting critical approaches.
Overall, the projects show a chronological, gradual shift towards a greater awareness. In early projects, uncertainty is explicitly addressed at the first two stages, but it is not explicitly reflected in the visualisation interface (e.g., only on hover annotation in Bufalini). In mythLOD, colour-coded filters in the catalogue visualisation allow this distinction, but it disappears in data visualisations. In ICON, subjectivity and interpretive layers are the subject of data visualisations. It is worth noting that earlier projects store individual, non-contradicting claims (e.g., Bufalini's perspective, Panofsky's perspective, or single assertions in mythLOD). In contrast, the most recent projects (Broast and artresearch.net) shift their acquisition and representation interests towards multi-perspective datasets, formalising contrasting opinions from different sources. artresearch.net shows increasing awareness across all DH workflow phases, attempting to leverage generous approaches to exploration that do not neglect transparency and interpretive plurality.
However, despite such few notable examples, we believe the "modelling-to-dissemination" bottleneck remains, since most projects share sophisticated knowledge representation patterns and minimal effort in visualising such aspects. Previous studies have confirmed this tendency (Battisti & Daquino, 2025a; Battisti & Daquino, 2025b) and have highlighted two major related challenges: the scarcity of visualisations used as critical reflective tools, and the inherent tension between the desire to adopt reusable (off-the-shelf) visual solutions and the intrinsic diversity of humanities data, which demands case-specific design. Highly reusable tools often impose rigid formats that require data to be processed to fit predefined structures, leaving little room for unanticipated interpretive specificities. It is therefore not accidental that most DH contributions propose the definition of design patterns to be used with existing software, showcase techniques to encode different uncertainty dimensions, or suggest the use of narrative strategies as a mitigating strategy (Panagiotidou et al., 2022; Windhager, Salisu, & Mayr, 2019; Battisti & Daquino, 2025a).
In conclusion, this contribution argues that within the hybrid set of technical and theoretical knowledge of a digital humanist, more attention should be paid to design processes and dissemination-oriented skills to face situations where collaboration with interface or visualisation experts is not feasible. At the same

time, we believe future work should move beyond high-level guidelines toward an operational research framework that investigates how reusability can be reconciled with interpretive complexity. While high-level approaches offer interdisciplinary application, the field must still define a threshold of standardisation: the point where a tool moves from a helpful framework to a rigid structure that flattens the very knowledge it seeks to represent.

REFERENCES

Barabucci, G., Tomasi, F., & Vitali, F. (2021). Supporting Complexity and Conjectures in Cultural Heritage Descriptions. COLCO 2020: Proceedings of the International Conference Collect and Connect: Archives and Collections in a Digital Age, CEUR Workshop Proceedings, 2810, 104–115. https://hdl.handle.net/11250/2736994

Battisti, T., & Daquino, M. (2025a). Exploring data-driven narratives in Digital Humanities web-based projects: Features and impact. Diversità, Equità e Inclusione: Sfide e Opportunità per l'Informatica Umanistica Nell'Era Dell'Intelligenza Artificiale, Proceedings Del XIV Convegno Annuale AIUCD2025, 18–23. https://doi.org/10.6092/unibo/amsacta/8380

Battisti, T., & Daquino, M. (2025b). Uncertainty, narrativity, and critical approaches in Digital Humanities information visualisation projects. Proceedings of the 21st Conference on Information and Research Science Connecting to Digital and Library Science, 1–10.

Berg, H. V. D., Betti, A., Castermans, T., Koopman, R., Speckmann, B., Verbeek, K. A. B., Werf, T. V. der, Wang, S., & Westenberg, M. A. (2018). A Philosophical Perspective on Visualization for Digital Humanities. In 3Rd Workshop on Visualization for the Digital Humanities.

Campagnolo, A. (2015). Transforming structured descriptions to visual representations. An automated visualization of historical bookbinding structures [PhD Thesis]. University of the Arts London.

Clavaud, F., & Wildi, T. (2021). ICA records in contexts-ontology (RiC-o): A semantic framework for describing archival resources. In Linked archives 2021: Proceedings of linked archives international workshop 2021 co-located with 25th international conference on theory and practice of digital libraries (TPDL 2021) (pp. 79–92). https://enc.hal.science/hal-03965776

Conroy, M., Gillmann, C., Harvey, F., Mchedlidze, T., Fabrikant, S. I., Windhager, F., Scheuermann, G., Tangherlini, T. R., Warren, C. N., Weingart, S. B., Rehbein, M., Börner, K., Elo, K., Jänicke, S., Kerren, A., Nöllenburg, M., Dwyer, T., Eide, Ø., Kobourov, S., & Betz, G. (2024). Uncertainty in humanities network visualization. Frontiers in Communication, 8, 1305137. https://doi.org/10.3389/fcomm.2023.1305137

Daquino, M., Giovannetti, F., & Tomasi, F. (2019). Linked Data per le edizioni scientifiche digitali. Il workflow di pubblicazione dell'edizione semantica del quaderno di appunti di Paolo Bufalini. Umanistica Digitale, 3(7). https://doi.org/10.6092/issn.2532-8816/9091

Daquino, M., Mambelli, F., & Kozlov, A. (2026). Many-to-many. Usability challenges of entity reconciliation in art history and photographic studies. Journal of Documentation, 1–19. https://doi.org/10.1108/JD-09-2025-0284

Daquino, M., Pasqual, V., & Tomasi, F. (2020). Knowledge representation of digital hermeneutics of archival and literary sources. JLIS: Italian Journal of Library, Archives and Information Science= Rivista Italiana Di Biblioteconomia, Archivistica e Scienza Dell'informazione: 11, 3, 2020, 59–76.

Daquino, M., Pasqual, V., Tomasi, F., & Vitali, F. (2022). Expressing without asserting in the arts. Proceedings of the Italian Research Conference on Digital Libraries. Padova, Italy, CEUR Workshop Proceedings. https://ceur-ws.org/Vol-3160/short14.pdf

Denard, H. (2016). A New Introduction to The London Charter. In Paradata and Transparency in Virtual Heritage (pp. 57–71). Routledge. https://doi.org/10.4324/9781315599366-7

Di Pasquale, A., Pasqual, V., Tomasi, F., & Vitali, F. (2024). On assessing weaker logical status claims in Wikidata cultural heritage records. Semantic Web, 15, 2395–2417. https://doi.org/10.3233/SW-243686

D'Ignazio, C., & Klein, L. F. (2020). Data Feminism. The MIT Press. https://doi.org/10.7551/mitpress/11805.001.0001

Doerr, M., Ore, C.-E., & Fafalios, P. (2023). Definition of the crminf. https://www.cidoc-crm.org/crminf/sites/default/files/CRMinf%20v1.0%28site%29.pdf

Doerr, M., Ore, C.-E., & Stead, S. (2007). The CIDOC conceptual reference model-A new standard for knowledge sharing. 26th International Conference on Conceptual Modeling (ER 2007), 83, 51–56. https://doi.org/10.13140/2.1.1420.6400

Dörk, M., Carpendale, S., & Williamson, C. (2011). The information flaneur: A fresh look at information seeking. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 1215–1224. https://doi.org/10.1145/1978942.1979124

Drucker, J. (2011). Humanities Approaches to Graphical Display. Digital Humanities Quarterly, 5(1). https://www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html

Drucker, J. (2013). Performative Materiality and Theoretical Approaches to Interface. Digital Humanities Quarterly, 007(1).

Drucker, J. (2014). Graphesis: Visual forms of knowledge production. Harvard University Press.

Elli, T., Briones Rojas, M., Gobbo, B., Parigini, M., Giustetto, V., Cavalloro, V., & Mauri, M. (2021). Metaphors as Knowledge Activators in Data Visualizations: The case of the Archipelago of Calvino's literary works. CUMULUS CONFERENCE PROCEEDINGS SERIES, 2, 906–924.

Fisher, D., Popov, I., Drucker, S., & Schraefel, M. C. (2012). Trust me, I'm partially right: Incremental visualization lets analysts explore large datasets faster. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 1673–1682. https://doi.org/10.1145/2207676.2208294

Gershon, N., & Page, W. (2001). What storytelling can do for information visualization. Communications of the ACM, 44(8), 31–37. https://doi.org/10.1145/381641.381653

Gitelman, L. (2013). 'Raw data' is an oxymoron. MIT press.

Haghighatkhah, P., El-Assady, M., Fekete, J.-D., Mahyar, N., Paradis, C., Simaki, V., & Speckmann, B. (2022). Characterizing Uncertainty in the Visual Text Analysis Pipeline. 2022 IEEE 7th Workshop on Visualization for the Digital Humanities (VIS4DH), 25–30. https://doi.org/10.1109/VIS4DH57440.2022.00010

Haider, S. (2022). Verzeichnis der den oberösterreichischen Raum betreffenden gefälschten, manipulierten oder verdächtigten mittelalterlichen Urkunden. Oberösterreichisches Landesarchiv.

Hinrichs, U., Forlini, S., & Moynihan, B. (2016). Speculative Practices: Utilizing InfoVis to Explore Untapped Literary Collections. IEEE Transactions on Visualization and Computer Graphics, 22(1), 429–438. https://doi.org/10.1109/TVCG.2015.2467452

Khulusi, R., Kusnick, J., Meinecke, C., Gillmann, C., Focht, J., & Jänicke, S. (2020). A Survey on Visualizations for Musical Data. Computer Graphics Forum, 39(6), 82–110. https://doi.org/10.1111/cgf.13905

Kosara, R., & Mackinlay, J. (2013). Storytelling: The Next Step for Visualization. Computer, 46(5), 44–50. https://doi.org/10.1109/MC.2013.36

Ma, R., & Xiao, F. (2021). What is a Good Visualization for Digital Humanities Researchers? An Exploratory Study. Proceedings of the Association for Information Science and Technology, 58(1), 484–489. https://doi.org/10.1002/pra2.483

Mariani, F. (2023). Introducing VISU: Vagueness, Incompleteness, Subjectivity, and Uncertainty in Art Provenance Data. Computational Methods in the Humanities 2022: Proceedings of the Workshop on Computational Methods in the Humanities 2022 Lausanne, Switzerland, June 9–10, 2022., 63–84. http://fox.leuphana.de/portal/en/publications/introducing-visu(3370ecc3-6940-467c-8388-6d9dbc2f93ef)/export.html

Mayr, E., Federico, P., Miksch, S., Schreder, G., Smuc, M., & Windhager, F. (2016). Visualization of cultural heritage data for casual users. IEEE VIS Workshop on Visualization for the Digital Humanities, 1.

Panagiotidou, G., Lamqaddam, H., Poblome, J., Brosens, K., Verbert, K., & Moere, A. V. (2022). Communicating Uncertainty in Digital Humanities Visualization Research. IEEE Transactions on Visualization and Computer Graphics, 1–11. https://doi.org/10.1109/TVCG.2022.3209436

Pasqual, V. (2025). The critical inquiry in humanities knowledge graphs: Challenges, methods and innovations [Phd, alma]. https://amsdottorato.unibo.it/id/eprint/12142/

Pasqual, V., & Tomasi, F. (2022). Linked open data per la valorizzazione di collezioni culturali: Il dataset mythLOD. AIB Studi, 62(1), 149–168. https://doi.org/10.2426/aibstudi-13301

Piotrowski, M. (2023). Uncertainty as Unavoidable Good. https://doi.org/10.4119/UNIBI/2983506

Porter, D., & Campagnolo, A. (2021). VisColl: Modeling and Visualizing the Physical Construction of Codex Manuscripts. University of Pennsylvania Libraries. https://viscoll.org.

Renda, G., Daquino, M., & Presutti, V. (2023). Melody: A Platform for Linked Open Data Visualisation and Curated Storytelling. Proceedings of the 34th ACM Conference on Hypertext and Social Media, 1–8. https://doi.org/10.1145/3603163.3609035

Sartini, B., Baroncini, S., van Erp, M., Tomasi, F., & Gangemi, A. (2023). ICON: An ontology for comprehensive artistic interpretations. Journal of Computational Cultural Heritage, 16(3), 59–76. https://doi.org/10.1145/3594724

Shan, X., Wang, D., & Li, J. (2022). Research on Data Storytelling Strategies for Cultural Heritage Transmission and Dissemination. In C. Stephanidis, M. Antona, S. Ntoa, & G. Salvendy (Eds), HCI International 2022 – Late Breaking Posters (pp. 344–353). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-19679-9_43

Therón Sánchez, R., Benito Santos, A., Santamaría Vicente, R., & Losada Gómez, A. (2019). Towards an Uncertainty-Aware Visualization in the Digital Humanities. Informatics, 6(3), 31. https://doi.org/10.3390/informatics6030031

Whitelaw, M. (2015). Generous Interfaces for Digital Cultural Collections. Digital Humanities Quarterly, 9(1), 1–16.

Windhager, F., Abdul-Rahman, A., Bludau, M.-J., Hengesbach, N., Lamqaddam, H., Meirelles, I., Speckmann, B., & Correll, M. (2024). Complexity as Design Material. 2024 IEEE Evaluation and Beyond - Methodological Approaches for Visualization (BELIV), 71–80. https://doi.org/10.1109/BELIV64461.2024.00013

Windhager, F., Federico, P., Schreder, G., Glinka, K., Dork, M., Miksch, S., & Mayr, E. (2019). Visualization of Cultural Heritage Collection Data: State of the Art and Future Challenges. IEEE Transactions on Visualization and Computer Graphics, 25(6), 2311–2330. https://doi.org/10.1109/TVCG.2018.2830759

Windhager, F., Filipov, V. A., Salisu, S., & Mayr, E. (2018). Visualizing Uncertainty in Cultural Heritage Collections. In K. Lawonn, N. Smit, L. Linsen, & R. Kosara (Eds), EuroVis Workshop on Reproducibility, Verification, and Validation in Visualization (EuroRV3). The Eurographics Association. https://doi.org/10.2312/eurorv3.20181142

Windhager, F., Salisu, S., & Mayr, E. (2019). Exhibiting Uncertainty: Visualizing Data Quality Indicators for Cultural Collections. Informatics, 6(3), 29. https://doi.org/10.3390/informatics6030029

Windhager, F., Salisu, S., Schreder, G., & Mayr, E. (2019). Uncertainty of what and for whom-and does anyone care? Propositions for cultural collection visualization. 4th IEEE Workshop on Visualization for the Digital Humanities (VIS4DH), 1–5.

Iconclass as a Semantic Framework for the Analysis of the Sacred Space

POLINA VORONOVA

BIBLIOTHECA HERTZIANA – MAX PLANCK INSTITUTE OF ART HISTORY, ROME, ITALY – POLINA.VORONOVA@BIBLHERTZ.IT

ABSTRACT (ENGLISH)

Rooted in the study of medieval liturgical furnishings, this work examines the application of the Iconclass system to three-dimensional objects, an area that has so far received little attention. Given the dispersed nature of these objects, representing them in a structured, machine-readable form enhances comparability, interoperability, and systematic analysis. The paper highlights the semantic structure implemented on the ResearchSpace platform, which allows the coexistence of two descriptive fields—one based on Iconclass notation and the other on natural language—thereby enabling richer iconographical representations. It also presents a dataset of 120 Iconclass-coded entries extracted from the Mapping Sacred Spaces database. Using SPARQL querying, the dataset enables cross-regional comparison of iconographic themes in the regions of southern Italy. While primarily methodological, the study demonstrates how semantic modelling supports structured description and lays the groundwork for future computational approaches, including clustering and vectorisation, to reveal patterns and regional trends in medieval ecclesiastical art.
Keywords: Digital iconography; Iconclass; Liturgical furnishings; Semantic Web; SPARQL

ABSTRACT (ITALIANO)

Iconclass come framework semantico per l'analisi dello spazio sacro
Il contributo analizza l'applicazione del sistema Iconclass agli arredi liturgici medievali, estendendo l'uso agli oggetti tridimensionali, ambito finora poco esplorato. Data la natura dispersa di questi manufatti, la loro rappresentazione in forma strutturata e leggibile dalle macchine consente una maggiore comparabilità, interoperabilità e analisi sistematica. Il lavoro mette in evidenza una struttura semantica, implementata sulla piattaforma ResearchSpace, che permette la coesistenza di due modalità descrittive — una basata sulla notazione Iconclass e l'altra sul linguaggio naturale — consentendo così una rappresentazione iconografica più ricca. Inoltre, presenta un dataset di 120 voci Iconclass, estratte dal database *Mapping Sacred Spaces*. Attraverso interrogazioni SPARQL, il dataset consente il confronto transregionale dei temi iconografici nell'Italia meridionale. Pur mantenendo un'impostazione prevalentemente metodologica, il contributo mostra come la modellazione semantica supporti una descrizione strutturata degli arredi liturgici e costituisca una base per futuri approcci computazionali, inclusi clustering e vettorializzazione, finalizzati all'individuazione di pattern e tendenze regionali nell'arte ecclesiastica medievale.
Parole chiave: Iconografia digitale; Iconclass; Arredi liturgici; Web semantico; SPARQL

1 INTRODUCTION

1.1 Iconography of liturgical furnishings

In medieval sacred space, iconographical motifs beyond mere decoration articulated theological and liturgical concepts. The ambo (see Fig. 1), above all other liturgical furnishings, became a significant vehicle for visual expression. It operated simultaneously as the place of proclamation, as a symbol and a metaphor, and as a medium of liturgical jubilation (Gianandrea, 2006; Scirocco, 2015).
Iconographic analysis has long been central to the study of sacred space, yet the task is complicated by the fact that most medieval liturgical furnishings were dismantled, displaced, or survive only in part (Longo & Scirocco, 2025). This fragmentary survival makes comparative study and long-term synthesis particularly difficult. Semantic data modelling offers a way to overcome these challenges by representing iconographic information in a structured, machine-actionable form. Using a controlled vocabulary such as

Iconclass[1], available as Linked Open Data (LOD), can help researchers disambiguate figures, scenes, and motifs across liturgical elements.
Moreover, the structured, semantic representation of iconographic data is particularly relevant for southern Italy, a region characterised by intense cultural interaction between Byzantine, Norman, Latin, and Islamic traditions. The complexity and hybridity of the iconographic repertoire found in liturgical elements from this area make them especially suitable for digital classification. In this respect, the *Mapping Sacred Spaces* project provides a methodological context in which traditional art-historical inquiry meets semantic data modelling approaches (Adamou & Voronova, 2025). It has directly informed the present research by highlighting the potential of structured iconographic data for analysing ecclesiastical spaces and their visual programs.



Figure 1[2]. Part of ambo of Teggiano Cathedral, Campania, Italy (Bibliotheca Hertziana – Max Planck Institute for Art History, photo: Enrico Fontolan)

1.2 Iconclass

Iconclass is an alphanumeric system that classifies iconographic subjects. Its hierarchical structure enables precise, comparable descriptions across objects, media, and regions. The present approach aims to demonstrate the potential of the Iconclass system beyond its most common application to the analysis of pictorial imagery. When applied to the iconography of medieval furnishings, Iconclass supports a more holistic interpretation of medieval ecclesiastical contexts and facilitates integrated analysis that bridges architecture, liturgy, and iconography by providing a shared classificatory language for these different dimensions. Combined with structured metadata on materials, techniques, provenance, and ritual function, Iconclass-based representations can support relational and network-oriented analyses, contributing to the reconstruction of artistic connections, workshop practices, and patterns of cultural exchange.

2. RELATED WORK

Research relevant to the present study can be grouped into three main strands: large-scale iconographic indexing initiatives, semantic modelling frameworks for cultural heritage data, and recent attempts to extend the expressive capacity of Iconclass beyond subject retrieval.

[1] https://iconclass.org/ (cons. 06/02/2026).
[2] Figure 1 illustrates the iconographic themes of the ambo in the Teggiano Cathedral. The sculptural program includes an angel holding a book, the symbol of Saint Matthew; an eagle, the symbol of Saint John the Evangelist; a man carrying a hare on his shoulders, among several other figures.

A first strand concerns the use of controlled vocabularies for the cataloguing and retrieval of visual material. The Index of Medieval Art[3] at Princeton University remains a major reference point for thematic and iconographic documentation of medieval artworks, offering a large controlled terminology linked to external authority systems such as the Getty AAT[4]. Comparable principles inform repositories such as Bildindex der Kunst und Architektur[5], where Iconclass supports iconographic access and image retrieval. These projects demonstrate the long-standing value of controlled vocabularies for standardisation, discoverability, and cross-collection comparison. Their primary focus, however, lies in indexing two-dimensional visual artefacts rather than modelling complex liturgical furnishings.

A second strand concerns semantic interoperability and knowledge representation. Projects such as SARI[6] and Linked.Art[7] have shown how CIDOC CRM-based approaches can model artworks, provenance, events, and institutional relationships in machine-readable form. More directly relevant is Sartini's IICONGRAPH (2024), which proposes the formal representation of iconographic and iconological statements through knowledge graphs and LOD principles. By embedding Iconclass concepts within explicit semantic structures, this framework enables the modelling of interpretative claims, relationships between motifs, and provenance of assertions. These developments are important because they move Iconclass from a classificatory vocabulary toward a richer semantic layer.

A third strand evaluates the limits and possible enrichment of Iconclass itself. Adamou and Picca (2024) examined whether Iconclass can encode emotional dimensions in classical imagery, showing both its usefulness as a basis for semantic expansion and its limits in representing fine-grained affective content. Baroncini and Mele (2025) similarly demonstrate how iconographic vocabularies may be integrated with ontological models in the specific domain of monastic imagery.

Taken together, this scholarship suggests that Iconclass is most effective when used not as a closed classification system, but as a semantic foundation that can be extended through linked data, ontologies, and interpretative layers. Yet these applications remain centred largely on pictorial media. The structured modelling of three-dimensional liturgical furnishings—objects combining sculpture, architecture, ritual function, and distributed iconographic programs—remains underdeveloped. The present study addresses this gap by investigating how Iconclass can support the comparative semantic analysis of medieval sacred space.

3. APPROACH

3.1 Iconclass integration into semantic model

Iconclass enables the transformation of liturgical furnishings from purely descriptive objects into analysable data structures. This approach has been successfully implemented in the *Mapping Sacred Spaces* project, where it has proven useful for art historians actively working with the database. In most cases, Iconclass categories adequately satisfy researchers' needs, even when addressing relatively detailed iconographic subjects, such as the category 92LL7141: fighting centaurs, centauromachy (in general) – LL – female centaurs (in general) or 11B13: Trinity represented by three animals sharing one head.

At the same time, the project complements iconographic references with a field for free-text iconographic description. Its principal function is to provide vertical enrichment[8], refining cases in which available Iconclass categories remain too broad. Domain experts may therefore record more detailed subject descriptions, local variants, relationships between figures, or symbolic interpretations. In addition, the field may contain observations relevant to iconographic analysis that function as horizontal enrichment, for example, the spatial position of motifs, their compositional arrangement, or their relationship to liturgical use. These annotations remain distinct from the object metadata already recorded in the model, such as material, technique, chronology, or provenance.

[3] https://theindex.princeton.edu/ (cons. 06/02/2026).
[4] https://www.getty.edu/research/tools/vocabularies/aat (cons. 06/02/2026).
[5] https://www.bildindex.de/ (cons. 06/02/2026)
[6] https://docs.swissartresearch.net/ (cons. 06/02/2026)
[7] https://linked.art/ (cons. 06/02/2026)
[8] In the present study, vertical enrichment designates the refinement of broad Iconclass categories through greater iconographic specificity, while horizontal enrichment refers to contextual, spatial, or relational information linked to the represented motif.

This is evident in the case of the Holy Water font capital of the Santissima Trinità di Venosa. The broad Iconclass category 31A: the (nude) human figure is accompanied by the free-text field: "The representation shows pairs of nude figures caught in the jaws of feral masks, a detail that may simultaneously evoke the theme of sin and the redemption of the human soul." (Audizi, Mapping Sacred Spaces database, 2026). Here, the annotation refines the generic classification, records relations between motifs, and preserves symbolic interpretation. Together, controlled vocabulary and textual annotation form a layered semantic model that balances standardisation with interpretive nuance.
Additionally, the project distinguishes between iconographic themes and ornamental motifs. While Iconclass categories already include certain geometric and decorative subjects, ornamental features are more often recorded in a dedicated ornamental motifs field, implemented separately from the iconographic field and functioning as a complementary thesaurus within the database. Considered together, these three fields provide a flexible semantic framework for modelling the heterogeneous visual registers present on medieval liturgical furnishings.

3.2 SPARQL

The ability to query a knowledge graph by semantic structure, e.g., via the SPARQL query language, enables domain experts to formulate and issue complex research questions across iconographic, spatial, chronological, and provenance-related dimensions of the data. One of the central research aims of *Mapping Sacred Spaces* is to analyse scattered evidence and visualize meaningful relationships among entities. Through these queries, researchers can explore the distribution, depth, and frequency of Iconclass entries across item types (components, objects), condition (lost, decontextualized), geographical location (provinces, regions), and chronological placement. For example, reconstructing the full liturgical ensemble of a dismantled medieval church requires combining elements preserved nowadays across different conservation sites. In response to this challenge, SPARQL querying serves as an ideal tool, providing a systematic framework for organising and representing data in a clear, structured manner that supports further art-historical analysis.
Stemming from the abovementioned project, we have created a dataset[9] comprising 120 Iconclass entries, corresponding to the full set of iconographical themes of liturgical furnishings currently documented in the *Mapping Sacred Spaces* database. The data were extracted using a SPARQL query that combines the endpoint of the liturgical furnishings dataset with that of Iconclass[10], ensuring direct alignment between iconographic classifications and their semantic definitions. The resulting dataset was organised into the following categories: value (Iconclass code), label (corresponding meaning), location (cathedral and its region), and number of occurrences.
This structure enables a comparative analysis of iconographic distributions by examining the relative frequency of Iconclass categories across different regions of southern Italy. From a methodological perspective, the use of SPARQL supports interoperability and data reuse, allowing iconographic data to be queried and recombined across independent semantic resources. At the same time, the approach is replicable and scalable, as it can be extended to different parameters as the database grows.

4. IMPLEMENTATION

The project's semantic catalogue is implemented in *ResearchSpace (RS),* a platform for managing the full lifecycle of semantic data (Oldman and Tanase, 2018), and is grounded on the *CIDOC Conceptual Reference Model (CRM)*. RS allows page template generation based on entity types, provides a SPARQL API for querying, and features a knowledge pattern (KP) system that dynamically maps between different data models at runtime. KPs use SPARQL queries and updates to map domain-level feature definitions not present in CRM itself to CRM concepts.
The Iconclass vocabulary is published as LOD, which provides the following three major advantages: first, its concepts can be directly incorporated into semantic data models; second, its notations function as stable, dereferenceable identifiers within RDF-based frameworks, supporting their integration into CIDOC CRM–compliant models; third, its interoperability across datasets is facilitated.

[9] The dataset, released in advance of the Mapping Sacred Spaces project, is accessible in CSV format at PoliVoro/Iconclass-for-liturgical-furnishings. RDF exports of the entire dataset will follow as the project data platform is opened to the public.
[10] Mirrored through an institutional knowledge graph, (data.biblhertz.it).

```
$subject crm:P128_carries ?iconography .
?iconography a crm:E90_Symbolic_Object
    ; crm:P138_represents $value .
$value a skos:Concept
    ; skos:prefLabel ?label .
```

Figure 2. Iconclass incorporation through SKOS.

```
$subject crm:P128_carries ?iconography .
?iconography a crm:E90_Symbolic_Object
    ; crm:P129i_is_subject_of ?info_obj .
?info_obj a crm:E73_Information_Object
```

Figure 3. Free text iconographic description

Iconographic information is integrated through two distinct KPs: Iconographical Themes**,** which enrich the semantic model by incorporating Iconclass as a controlled vocabulary, and Iconographic Description**,** which provides an additional descriptive note when the former is not sufficient to capture the iconographic content. Both share an entity of CRM type `E90_Symbolic_Object`, which represents the overall iconography of that item, such as a complex scene. Iconclass concepts are connected through a SKOS[11] taxonomy and linked to the data structure through CRM property `P138_represents` (see Fig. 2). The second free-text KP further documents the iconography by attaching to it an `E73_Information_Object` that encapsulates a natural-language description (see Fig. 3). `$subject` and `$value` are special variables that RS resolves to the entity under scrutiny (e.g., the ambo) and the value to be entered or retrieved (e.g., the Iconclass notation), respectively.

5. CONCLUSION AND FUTURE WORK

This study aims to demonstrate the applicability of the Iconclass system to three-dimensional objects within medieval sacred space in the semantic framework ResearchSpace. While Iconclass categories generally provide a robust classification, the addition of a free-text field allows researchers to record observations that complement formal encoding. By combining structured codes with textual annotations, the approach establishes a flexible semantic model that supports both comparative study and semantic interoperability, with potential applications in large language models and 3D modelling of liturgical spaces.

The database Mapping Sacred Spaces remains under development aiming at a late-2026 public release and is continuously being populated with new entries by domain experts. Consequently, comprehensive results could not yet be rendered to fully illustrate the method. Nevertheless, the methodology has been proposed, and the existing dataset of 120 Iconclass-coded entries has been made publicly available to support reproducibility and further research.

Future work will focus on the systematic expansion of the dataset and the application of computational techniques, such as clustering and vectorisation, to detect patterns and relationships across iconographic programs and regions. It will also explore AI-assisted methods for suggesting Iconclass descriptors from textual object descriptions, to support expert annotation workflows and improve scalability. In parallel, further research will investigate the suitability of the Iconclass paradigm for the direct annotation of visual representations through International Image Interoperability Framework (IIIF) protocols.

These approaches enable the identification of recurring motifs and regional variations, while allowing art historians to uncover meaningful correlations between architectural, liturgical, and iconographic features, offering significant potential for comparative and semantic analyses. In the longer term, it establishes a foundation for structured representation of medieval ecclesiastical art, transforming descriptive cataloguing into interoperable, analysable data, and supporting scalable, reproducible, and comparative research across dispersed and fragmentary material.

[11] https://www.w3.org/TR/skos-reference/ (cons. 01/02/2026)

REFERENCES

Adamou, A., & Picca, D. (2024). Assessing the expressivity of Iconclass to embody emotional features in classical iconography. In *CEUR Workshop Proceedings (Vol. 3809).* https://www.scopus.com/pages/publications/85210267927

Adamou, A., Voronova, P. (2025). Mediating art history data models for native linked data construction using ResearchSpace. In S. Campana, D. Ferdani, H. Graf, G. Guidi, Z. Hegarty, S. Pescarin, & F. Remondino (Eds.), *Digital Heritage*. The Eurographics Association.

Baroncini, F., & Mele, G. (2025). Describing monastic iconography using semantic data. *AIUCD 2025* https://aiucd2025.dlls.univr.it/assets/pdf/papers/76.pdf

Gianandrea, M. (2006). *La scena del sacro: L'arredo liturgico nel basso Lazio tra XI e XIV secolo* (pp. 35–80). Rome: Viella.

Longo, R., & Scirocco, E. (2025). Fragmented objects/Fragmented data: Quantitative methods and digital approaches in the project *Mapping Sacred Spaces: Forms, Functions, and Aesthetics in Medieval Southern Italy*. In M. Pilutti Namer, A. Auconi, & G. A. Bordi (Eds.), *Networks and quantitative methods in archaeology and art history: Essays from the Venice Symposium (5–6 December 2024)*, *Archeologia e Calcolatori*, 36(2), 85–102. https://doi.org/10.19282/ac.36.2.2025.08

Oldman, D., & Tanase, D. (2018). Reshaping the knowledge graph by connecting researchers, data, and practices in ResearchSpace. In D. Vrandecic, K. Bontcheva, M. C. Suárez-Figueroa, V. Presutti, I. Celino, M. Sabou, L. Kaffee, & E. Simperl (Eds.), *The Semantic Web – ISWC 2018: 17th International Semantic Web Conference, Monterey, CA, USA, October 8–12, 2018, Proceedings, Part II* (Lecture Notes in Computer Science, Vol. 11137, pp. 325–340). Springer. https://doi.org/10.1007/978-3-030-00668-6_20

Sartini, B. (2024). IICONGRAPH: Improved iconographic and iconological statements in knowledge graphs. In *The Semantic Web: 21st International Conference, ESWC 2024, Hersonissos, Crete, Greece, May 26–30, 2024, Proceedings, Part II* (pp. 57–74). Springer-Verlag. https://doi.org/10.1007/978-3-031-60635-9_4

Scirocco, E. (2010). *Arredi liturgici dei secoli XI–XIII in Campania: Le cattedrali di Salerno, Ravello, Amalfi, Caserta Vecchia, Capua*. Napoli.

Scirocco, E. (2015). Jonah, the whale and the ambo: Image and liturgy in medieval Campania. In I. Foletti & Z. Frantová (Eds.), *The Antique Memory and the Middle Ages* (pp. 87–123). Rome: Viella.

[7/B] GIS e spazializzazioni digitali

Gli *Historical GIS* e la storia forestale: potenzialità e applicazioni

VINCENZO COLAPRICE[1]

[1] UNIVERSITÀ DEGLI STUDI DI TORINO, ITALIA – VINCENZO.COLAPRICE@UNITO.IT

ABSTRACT (ITALIANO)

Il contributo illustra le potenzialità applicative degli *Historical GIS* (HGIS) nell'ambito del progetto ERC INWOOD (*European Industrialisation as Seen from the Forests*), volto a indagare il ruolo del legno nei processi di industrializzazione europea tra il 1870 e il 1914. Il paper propone un approccio metodologico basato sull'integrazione di fonti storiche eterogenee all'interno di ambienti GIS, finalizzato all'analisi congiunta delle dinamiche economiche, sociali ed ecologiche legate allo sfruttamento delle risorse forestali. L'analisi si concentra su tre ambiti principali: la ricostruzione delle geografie degli investimenti e dei flussi commerciali del legname attraverso la spazializzazione dei dati ricavati dalla stampa europea; la *network analysis* della mobilità dei lavoratori coinvolti nell'economia del legno; la comparazione diacronica della copertura forestale mediante la metodologia del filtraggio cartografico. I casi studio selezionati (Val di Fiemme e Cadore nell'area italiana, Galizia orientale e Transilvania) consentono di sviluppare un'analisi spaziale multiscala all'interno di un comune contesto politico-istituzionale. I risultati presentati hanno carattere preliminare e mirano a illustrare scenari applicativi e potenzialità metodologiche degli HGIS in relazione alla storia forestale.
Parole chiave: Historical GIS; storia forestale; cartografia digitale; industrializzazione.

ABSTRACT (ENGLISH)

Historical GIS and Forest History: Potential and Applications. This paper illustrates the potential applications of Historical GIS (HGIS) within the ERC project INWOOD (European Industrialisation as Seen from the Forests), which investigates the role of wood in European industrialisation processes between 1870 and 1914. It proposes a methodological approach based on the integration of heterogeneous historical sources within GIS environments, aimed at the joint analysis of economic, social, and ecological dynamics related to the exploitation of forest resources. The analysis focuses on three main areas: the reconstruction of the geographies of investments and timber trade flows through the spatialisation of data derived from European periodical publications; the network analysis of the mobility of workers involved in the wood economy; and the diachronic comparison of forest cover through the cartographic filtering methodology. The selected case studies (Val di Fiemme and Cadore in the Italian area, Habsburg Galicia, and Transylvania) enable the development of a multi-scale spatial analysis within a common political-institutional framework. The results presented are preliminary and aim to illustrate the methodological potential and application scenarios of HGIS in relation to forest history.
Keywords: Historical GIS; Forest History; Digital Cartography; Industrialization.

1. INTRODUZIONE

Con l'affermarsi dello spatial turn (Lefebvre, 1991; Massey, 1995; Tuan, 2002), l'impiego di metodologie di ricerca geospaziali è divenuto crescente nell'ambito delle discipline umanistiche (Murrieta-Flores & Martins, 2019). In particolare, nel corso dell'ultimo trentennio si sono affermate una varietà di innovazioni tecnologiche e metodologiche, oggi comprese nel campo della Spatial History, che hanno trasformato e ampliato la relazione tra geografia e storia, contribuendo a superare il divario epistemologico esistente tra le due discipline (Knowles & Hillier, 2008, p. 3; DeBats et al., 2018, p. 1). A tale scenario ha contributo anche l'introduzione dei sistemi informativi geografici (GIS) nell'ambito della ricerca storica, le cui metodologie di applicazione sono comprese sotto la denominazione di Historical GIS (HGIS). La loro implementazione prevede la creazione e l'uso di database relazionali contenenti dati storici e geografici adoperati all'interno di un ambiente GIS, consentendo agli utenti di conservare, consultare, analizzare e visualizzare dati georeferenziati (Gregory & Geddes, 2014, p. XI; DeBats et al., 2018, p. 1). Pertanto, gli HGIS si presentano come degli strumenti epistemologici che consentono di interrogare criticamente le fonti e formulare nuove ipotesi interpretative (Grava et al., 2020, p. 3). A tal

proposito, sono ormai note varie modalità di applicazione degli HGIS nei vari campi della ricerca storica, contemplando l'utilizzo di dati e fonti di carattere quantitativo e qualitativo (Knowles, 2005; Gregory & Ell, 2007; Von Lünen & Travis, 2013; Gregory et al., 2018b; Grava et al., 2020). Uno degli ambiti in cui gli HGIS hanno avuto un'applicazione più diffusa è rappresentato dalla storia ambientale. Essa, infatti, si rivela particolarmente predisposta alla luce della possibilità di localizzare e visualizzare i dati relativi a fenomeni ecologici e ambientali manifestatisi entro un dato arco di tempo, rispondendo a domande di ricerca spesso caratterizzate da una natura intrinsecamente eografica (Gregory et al., 2018a, p. 351). Nell'ambito della storia forestale, tra le applicazioni più comuni degli HGIS figura la ricostruzione degli usi e della copertura del suolo su scala diacronica (cfr. Munteanu et al., 2015; Statuto et al., 2017; Bruzzone et al., 2019; Gabellieri & Sarzotti, 2024). In tal caso, la fonte maggiormente utilizzata e analizzata è la cartografia storica, opportunamente reinterpretata attraverso l'estrazione, analisi e riproduzione di informazioni georeferenziate relative allo sfruttamento del suolo e alla distribuzione e qualità della vegetazione. Nondimeno, la molteplicità di aspetti che caratterizza qualsiasi fenomeno storico rende problematico basare l'analisi su un solo tipo di fonte. Un approccio che si limiti alla sola cartografia rischia infatti di ridurre la complessità dei processi storici alle loro manifestazioni spaziali, trascurando le dinamiche economiche, sociali e istituzionali che ne costituiscono il presupposto. Parimenti, la ricerca storica richiede un lavoro di costante comparazione tra fonti differenti e di diverso tipo. Come ricondurre, dunque, questa complessità all'interno degli HGIS? Superando qualsiasi approccio di tipo positivistico, l'impiego dei GIS nella ricerca storica deve essere concepito come parte di un processo epistemologico più ampio, nel quale la traduzione delle informazioni storiche in database, attributi e campi analitici non rappresenti un semplice esercizio di standardizzazione dei dati, ma uno strumento per interrogare criticamente la pluralità e l'eterogeneità delle fonti. La formalizzazione dei dati storici all'interno di un HGIS comporta infatti il rischio di silenziare ambiguità, sfumature interpretative e discontinuità documentarie, elementi che costituiscono invece componenti essenziali dell'analisi storica (Bodenhamer, 2013, p. 6). Per sviluppare degli HGIS capace di restituire questa profondità interpretativa, è dunque necessario affiancare all'analisi cartografica l'integrazione sistematica di fonti differenti. Solo attraverso un approccio che combini cartografia, documentazione testuale, dati economici e fonti locali è possibile comprendere non soltanto gli esiti spaziali di un fenomeno storico, ma anche i processi economici, sociali e istituzionali che ne hanno determinato l'evoluzione. Questa prospettiva costituisce il punto di partenza del presente contributo, collocandosi nel soldo di recenti esperienze di sviluppo di HGIS basati sull'utilizzo di fonti eterogenee nell'ambito della storia forestale (Colaprice, 2025). A partire da tale impostazione metodologica, il contributo illustra diverse modalità di applicazione degli HGIS in relazione alla molteplicità di fonti storiche adoperate nell'ambito del progetto ERC INWOOD (*European Industrialisation as Seen from the Forests*), volto a comprendere le dinamiche economiche, sociali ed ecologiche che hanno caratterizzato il rapporto tra industrializzazione e sfruttamento delle risorse forestali in Europa tra il 1870 e il 1914. In particolare, il progetto pone la sua attenzione sul ruolo avuto dalla più importante risorsa preindustriale, il legno, evidenziando come la diffusione di nuove fonti energetiche e nuovi materiali non abbia comportato una sua sostituzione immediata, inaugurando, piuttosto, una fase di coesistenza (Iriarte-Goñi, 2013; Fressoz, 2023). Nei decenni a cavallo tra XIX e XX secolo, il legno continuò a occupare una posizione centrale sia come fonte energetica sia come materia prima impiegata in numerosi settori produttivi: dalla costruzione di infrastrutture e miniere allo sviluppo delle reti ferroviarie, fino alla produzione della cellulosa e alla diffusione del telegrafo (Braudel, 1967; Brett, 2019; Fressoz & Bonneuil, 2017; Knoll, 2006; Särkkä, 2021). La persistenza e la riorganizzazione dei processi produzione e commercializzazione del legno accompagnarono l'emergere di diversi fenomeni già evidenziati dalla storiografia, tra cui la ridefinizione dei diritti di accesso alle risorse forestali (Bonan, 2019; Balogh, 2026), il progressivo sfruttamento di aree forestali sempre più distanti dai luoghi di consumo, secondo la dinamica di espansione della timber frontier (Grewe, 2014; Lotz, 2015; Daheur, 2016), nonché il manifestarsi della forest transition, ovvero il passaggio da fasi di deforestazione a periodi di riforestazione (Kaplan et al., 2009; Loran et al., 2016; Gingrich et al., 2019). Comprendere le interazioni tra trasformazioni economiche, dinamiche sociali e cambiamenti dei paesaggi forestali rappresenta una delle principali sfide della ricerca attuale. In questa prospettiva, l'integrazione tra fonti storiche eterogenee e HGIS consente di analizzare tali processi in modo differenziato e multi-scalare.

2. METODOLOGIA E FONTI

Le attività di ricerca si articolano attorno a tre principali direttrici: 1) l'analisi delle dinamiche economiche su scala macro, con particolare attenzione alle trasformazioni dei consumi di legname, alla diffusione di nuove fonti energetiche e alla ridefinizione della geografia dei flussi e delle filiere di approvvigionamento; 2) lo studio delle dinamiche sociali su scala micro, volto a comprendere l'impatto dei cambiamenti dell'economia del legno sulle comunità locali e sulle scelte degli attori coinvolti nello sfruttamento delle risorse forestali; 3) l'indagine delle dinamiche ecologiche su scala multipla, finalizzata a ricostruire l'evoluzione della copertura forestale europea e le principali trasformazioni qualitative dei paesaggi boschivi. Nel dare seguito a questi orientamenti di ricerca, l'attenzione è ricaduta sul territorio dell'Impero austro-ungarico. Considerato nel suo complesso, esso risulta tra i meno studiati nell'ambito della storia ambientale, al netto della letteratura tendente a focalizzarsi su singole aree e regioni dell'Impero (Daheur & Lučić, 2026, pp. 3–5). Al tempo stesso, la sua notevole estensione territoriale e l'elevato grado di diversità interna consentono di osservare come dinamiche analoghe possano produrre esiti eventualmente differenti in contesti diversi, rilevando somiglianze, divergenze e discontinuità nei processi osservati all'interno di una stessa unità politico-istituzionale. Pertanto, sono state individuate tre aree all'interno del territorio asburgico che fungono da casi studio di scala locale (Fig. 1): l'area tra la val di Fiemme e il Cadore per l'ambito italiano; i possedimenti boschivi della Chiesa greco-cattolica di Perehinsko, nell'attuale Galizia ucraina; l'area compresa tra le località di Tălmaciu (Nagytalmács) e Zăbala (Zabola) nella Transilvania ungherese, oggi parte della Romania.

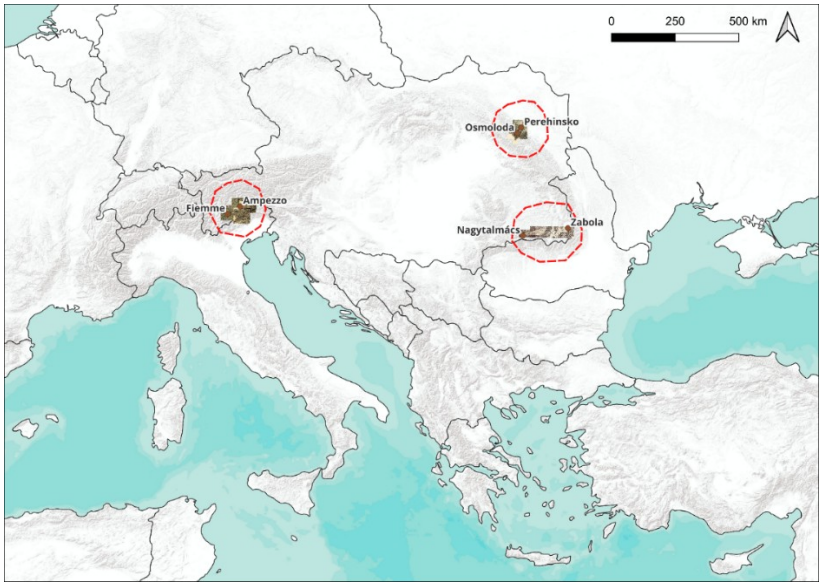


Figura 1. Posizione dei casi studio all'interno del territorio austro-ungarico. Confini d'Europa al 1886. Fonte: elaborazione dell'autore in QGIS.

Le metodologie di ricerca adottate per sviluppare le tre direttrici menzionate prevedono l'impiego integrato di fonti qualitative e quantitative, tutte caratterizzate da potenziale applicativo nell'ambito degli HGIS. Per quanto riguarda le dinamiche economiche, le fonti privilegiate sono costituite dalle riviste forestali in lingua tedesca pubblicate in Europa tra la fine dell'Ottocento e l'inizio del Novecento, dalle quali è possibile ricavare informazioni di carattere statistico ed economico relative ai flussi commerciali del legname, nonché dati sulla compravendita di proprietà forestali tra Europa centrale e orientale. In questo contesto, si prevede la costruzione di un HGIS di base volto a evidenziare le relazioni spaziali tra sedi societarie e proprietà forestali acquisite, attraverso la visualizzazione delle informazioni contenute nel database delle acquisizioni e della loro successione cronologica. Per quanto concerne le dinamiche sociali, una parte delle attività di ricerca è dedicata alla ricostruzione della migrazione di lavoratori italiani verso i territori orientali dell'Impero austro-ungarico. Le fonti utilizzate comprendono la letteratura esistente, documentazione demografica e quotidiani italiani rivolti ai lavoratori emigrati. In questo caso, l'applicazione dell'HGIS prevede l'impiego di un database contenente le informazioni relative ai movimenti dei lavoratori tra Italia e Impero austro-ungarico, con l'obiettivo di sviluppare una *network analysis* capace di evidenziare le dinamiche della circolazione del lavoro specializzato nell'Europa della seconda

metà dell'Ottocento. Infine, per quanto riguarda le dinamiche ecologiche, l'approccio HGIS è finalizzato alla ricostruzione della copertura forestale nei tre casi studio attraverso l'analisi e la comparazione diacronica e regressiva della cartografia storica pertinente all'arco cronologico del progetto, secondo la metodologia del filtraggio cartografico (Cevasco, 2007, pp. 74–83; Dai Prà et al., 2022; Dossche & Van Eetvelde, 2022). A tal fine, gli estremi della serie cartografica considerata sono rappresentati dal dataset Corine Land Cover 2018, rilasciato in formato vettoriale dal Copernicus Land Monitoring Service (scala 1:100.000), e dai fogli di mappa del Second Military Survey dell'Impero asburgico (scala 1:28.800), prodotti tra il 1806 e il 1869. La serie sarà integrata con ulteriore cartografia storica relativa alle fasi cronologiche intermedie e di diversa tipologia (cartografia militare, catastale e locale).

3. APPLICAZIONE DI SOLUZIONI HGIS: PRIMI SCENARI E IPOTESI

La presente sezione illustra l'applicazione delle metodologie HGIS in relazione alle tre direttrici di ricerca che animano il progetto. I risultati qui discussi hanno carattere preliminare e non definitivo, poiché la ricerca è tuttora in corso e il progetto si colloca nelle sue fasi iniziali. Con riferimento al primo ambito di applicazione, attraverso la georeferenziazione delle informazioni contenute nelle notizie relative alla compravendita di tenute forestali, all'acquisizione di diritti di sfruttamento e agli investimenti delle imprese del legname, è stato possibile creare una base HGIS in cui visualizzare i dati ricavati da riviste forestali di lingua tedesca pubblicate tra il 1876 e il 1913 (Fig. 2)[1].

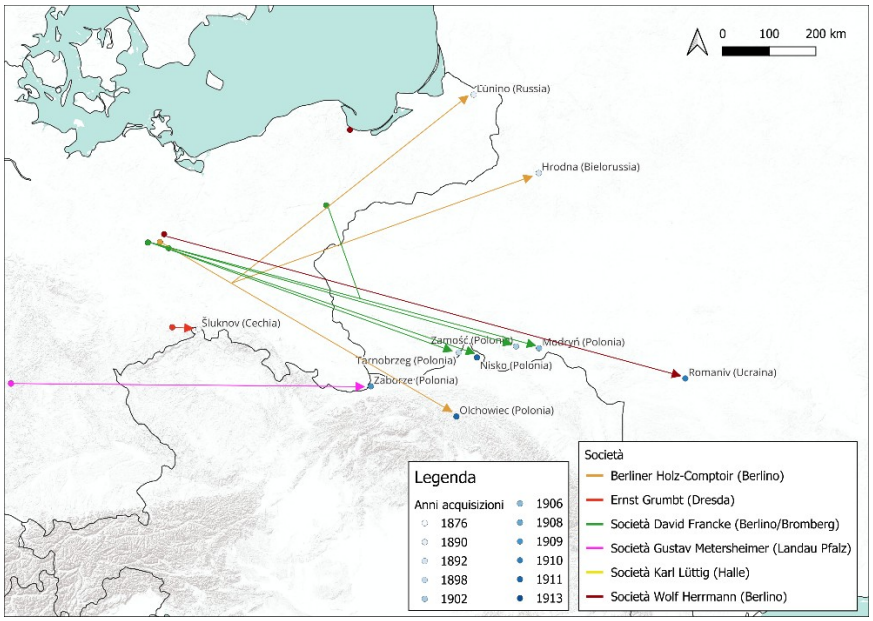


Figura 2. Visualizzazione dei dati relativi alla compravendita di tenute forestali. Fonte: elaborazione dell'autore in QGIS.

I dati raccolti finora valicano l'area dell'Impero austro-ungarico, evidenziando l'attività di società con sede nell'Impero tedesco, i cui investimenti si concentrano in diverse aree dell'Europa orientale e in particolar modo in Russia. Questo primo risultato consente di individuare alcune direttrici ricorrenti di espansione del mercato e di ipotizzare il sussistere di una relazione tra la localizzazione degli investimenti e la qualità e accessibilità delle risorse forestali, a partire dalla presenza e dallo sviluppo delle infrastrutture ferroviarie. Al contempo, permangono alcuni limiti dovuti alla natura non omogenea delle fonti, le cui informazioni risultano talvolta incomplete. Per quanto riguarda la *network analysis* della circolazione dei lavoratori italiani verso l'Impero austro-ungarico, la loro presenza risulta particolarmente significativa nella seconda metà dell'Ottocento nelle aree interessate dall'attività di imprese forestali collegate al conte Ármin Mikes a Gelence e Zabola (contea di Háromszék) e della società Feltrinelli a Nagytalmács (contea di Sibiu). Sebbene questo caso applicativo sia ancora in fase di sviluppo, i dati raccolti in un database GIS dovrebbero consentire di individuare nodi relazionali connessi ai movimenti migratori, nonché tendenze

[1]Le riviste da cui sono stati estratti i dati sono le seguenti: Dresdner Nachrichten, Kölnische Zeitung, Sächsischer Landes-Anzeiger, Berliner Börsen-Zeitung, Österreichische Forst-Zeitung, General-Anzeiger der Stadt Mannheim und Umgebung, Berliner Tageblatt, Hamburger Fremdenblatt. Tale lavoro di spoglio si deve a Cristiano La Lumia.

specifiche che colleghino luoghi di origine, rotte di migrazione, lavoratori e aree di insediamento. Infine, per quanto riguarda le dinamiche ecologiche, l'analisi diacronica della copertura forestale è stata avviata attraverso la comparazione della cartografia storica mediante la metodologia del filtraggio cartografico. Nei tre casi studio di scala locale, il lavoro ha permesso di estrarre i dati relativi al Corine Land Cover 2018 e di confrontarli con le superfici forestali ricavate dal Second Military Survey asburgico, consentendo di individuare alcune prime linee di tendenza nell'evoluzione della copertura boschiva tra XIX e XXI secolo (Fig. 3).

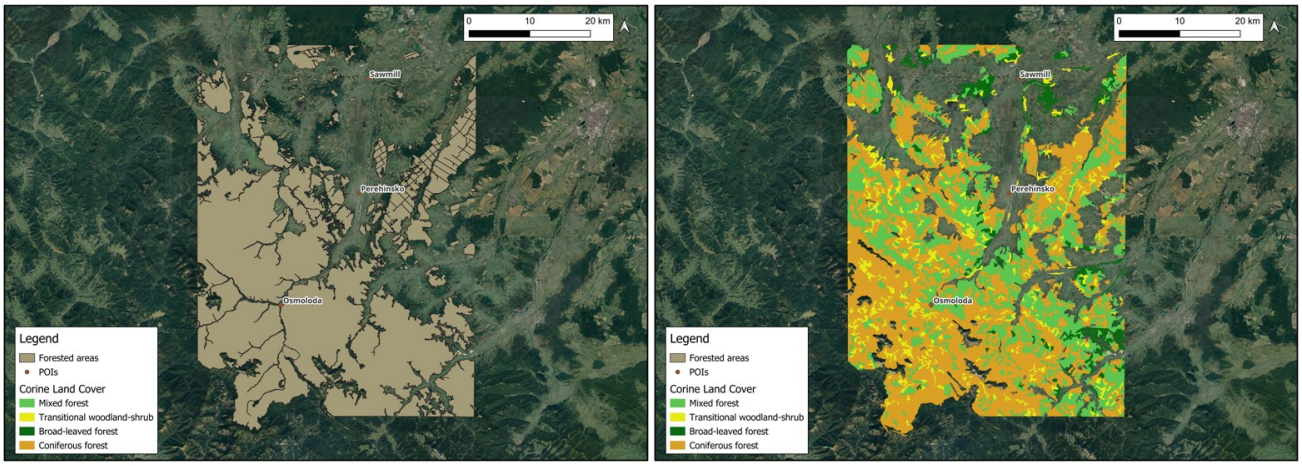


Figura 5. Comparazione tra le superfici forestali estratte dal Second Military Survey (1853-1870) e dal Corine Land Cover (2018) per il caso studio della Galizia. Fonte: elaborazione dell'autore in QGIS.

Nel caso della Galizia, i fogli di mappa prodotti tra il 1861 e il 1864 indicano una superficie boschiva di 1.449,96 km², aumentata a 1.789,58 km² nel 2018; in Transilvania, le carte risalenti al periodo 1853-1870 indicano una superficie di 2.867,56 km², sostanzialmente stabile rispetto ai 2.649,25 km² del 2018. Infine, nell'area alpina orientale italiana, i fogli di mappa risalenti al periodo 1816-1829 indicano una superficie di 3.672,36 km², inferiore rispetto ai 4.113,76 km² registrati nel 2018. Come si può notare, questi dati fanno emergere delle tendenze differenziate tra i casi studio, rendendo necessario integrare la serie cartografica utilizzata con fonti risalenti a fasi intermedie in modo da approfondire il livello di ricostruzione dell'evoluzione della copertura forestale, rilevando eventuali linee di continuità e discontinuità rispetto al fenomeno della *forest transition*. Parimenti, occorre considerare la forte eterogeneità dei fogli cartografici utilizzati, realizzati in epoche differenti e secondo criteri di rappresentazione non sempre conformi agli standard indicati dal *Muster-Blätter für die Darstellung des Terrains in militärischen Aufnahms-Plänen*, il quale costituisce la guida interpretativa al Second Military Survey. Tale variabilità impone cautela nella vettorializzazione delle superfici forestali e ricorda la necessità di interpretare criticamente i dati GIS, evitando di considerarli come certi e definitivi nel processo di comparazione diacronica delle fonti cartografiche. Al termine del completamento di tali soluzioni HGIS, è prevista la loro integrazione all'interno di un web-GIS che sarà ospitato sul sito del progetto[2]. La creazione del web-GIS prevede l'integrazione tra tecnologie open source, come Leaflet, e l'archiviazione della documentazione necessaria su un repository GitHub. I dati geospaziali, lavorati in QGIS e espostati in formati aperti e interoperabili (GeoPackage, GeoJSON, GeoTIFF), saranno pubblicati secondo i principi FAIR e rilasciati attraverso Zenodo al fine di garantirne l'accessibilità nel lungo periodo attraverso l'associazione di DOI.

4. RIFLESSIONI CONCLUSIVE

Il contributo ha inteso presentare le diverse possibilità applicative degli HGIS nel contesto delle attività di ricerca del progetto INWOOD, delineando vari approcci metodologici attraverso cui integrare fonti eterogenee (cartografiche, testuali e archivistiche). In tal modo è possibile contribuire a una ricostruzione multi-scalare dei processi economici, sociali ed ecologici legati all'economia del legno tra Ottocento e Novecento. Le applicazioni presentate mostrano come gli HGIS possano costituire non soltanto strumenti di visualizzazione dei dati, ma dispositivi analitici in grado di mettere in relazione fenomeni storici

[2]https://www.inwood.unito.it/ (cons. 15/04/2026)

differenti, consentendo di osservare scale locali, regionali e continentali. I risultati illustrati in queste pagine hanno carattere preliminare. Nei prossimi mesi, l'ampliamento delle serie documentarie, l'integrazione di ulteriore cartografia storica e il perfezionamento dei modelli di analisi spaziale consentiranno di affinare le interpretazioni qui proposte, rendendo possibile una comparazione più solida tra i casi studio e una maggiore precisione nella ricostruzione delle dinamiche osservate. In questa prospettiva, il ricorso agli HGIS non costituisce un punto di arrivo, ma un processo metodologico in progressiva costruzione, destinato a integrare nuove pluralità di fonti e dati.

RINGRAZIAMENTI

Il lavoro di ricerca esposto in queste pagine è finanziato dall'Unione Europea attraverso il progetto ERC StG 2023 'Industrial Wood: European Industrialisation as Seen from the Forests (1870-1914)' (INWOOD, grant agreement no. 101115916).

BIBLIOGRAFIA

Balogh, R. (2026). An Environmental History of Knowledge and Politics Forestry in Nineteenth- and Twentieth-Century Hungary. In *An Environmental History of Knowledge and Politics*. Central European University Press. https://doi.org/10.5117/9789633868430_ch02

Bodenhamer, D. J. (2013). Beyond GIS: Geospatial Technologies and the Future of History. In A. Von Lünen & C. Travis (A c. Di), *History and GIS: Epistemologies, Considerations and Reflections* (pp. 1–14). Springer Netherlands. https://doi.org/10.1007/978-94-007-5009-8

Bonan, G. (2019). *The State in the Forest: Contested Commons in the Nineteenth Century Venetian Alps*. The White Horse Press.

Braudel, F. (1967). *Civilisation Matérielle et Capitalisme (XVe-XVIIIe Siècle)*. Librairie Armand Colin.

Brett, A. (2019). Railways and the Exploitation of Victoria's Forests, 1880s–1920s. *Australian Economic History Review*, *59*(2), 159–180. https://doi.org/10.1111/aehr.12158

Bruzzone, R., Cevasco, R., Gabellieri, N., Montanari, C., Moreno, D., Pescini, V., & Traldi, C. (2019). "Volta la carta". Cartografia storica e ricerca multidisciplinare: La caratterizzazione storico-ambientale dei paesaggi rurali. Casi studio dalla Liguria. In F. Salvatori, *L'apporto della Geografia tra rivoluzioni e riforme* (pp. 541–547). AGeI. https://unige.iris.cineca.it/handle/11567/975888?mode=full

Cevasco, R. (2007). *Memoria verde. Nuovi spazi per la geografia*. Edizioni Diabasis.

Colaprice, V. (2025). Un Historical GIS per la ricostruzione della copertura boschiva tra Val Trebbia e Val d'Aveto (Liguria, XIX-XXI sec.). In N. Gabellieri & G. Cristina (A c. Di), *Biografie di paesaggi boschivi. Nuove ricerche di geografia e storia* (print, pp. 181–204). Viella editrice. https://doi.org/10.52056/9791257010492

Daheur, J. (2016). Exporting Environmental Burdens into the Central-European Periphery: Christmas Tree Trade and Unequal Ecological Exchange Between Germany and Habsburg Galicia around 1900. *HISTORYKA. Studia Metodologiczne*, *46*, Articolo hal-03081236.

Daheur, J., & Lučić, I. (2026). Towards the Writing of an Environmentally Inspired History of the Late Habsburg Empire. In J. Daheur & I. Lučić (A c. Di), *Habsburg Natures. Imperial Governance and Environment in Central Europe, 1850-1918* (1ª ed., pp. 1–24). Berghahn Books. https://doi.org/10.3167/9781836952275

Dai Prà, E., Gabellieri, N., & Scanu, N. (2022). Dalla mappa al geodatabase: Un modello di raccolta, digitalizzazione e analisi sincronica e diacronica in ambiente GIS del patrimonio toponomastico del territorio trentino da fonti cartografiche storiche (XIX-XXI secolo). *Il Capitale Culturale. Studies on the Value of Cultural Heritage*, (25), 603–633.

DeBats, D. A., Gregory, I., & Lafreniere, D. (2018). Spatial History, History, and GIS. In I. Gregory, D. A. DeBats, & D. Lafreniere, *The Routledge Companion to Spatial History* (pp. 1–6). Routledge.

Dossche, R., & Van Eetvelde, V. (2022). Land Abandonment and Its Impact on the Landscape Character of Val Borbera (Northern Apennines). In E. Borgogno-Mondino & P. Zamperlin (A c. Di), *Geomatics and Geospatial Technologies* (Vol. 1507, pp. 366–384). Springer International Publishing. https://doi.org/10.1007/978-3-030-94426-1_27

Fressoz, J.-B. (2023). Bois et charbon: Une histoire symbiotique de l'industrialisation. *Histoire & mesure*, *XXXVIII*(1), 157–186. https://doi.org/10.4000/histoiremesure.19153

Fressoz, J.-B., & Bonneuil, C. (2017). Growth Unlimited: The Idea of Infinite Growth from Fossil Capitalism to Green Capitalism. In I. Borowy & M. Schmelzer (A c. Di), *History of the Future of Economic Growth* (pp. 52–68). Routledge.

Gabellieri, N., & Sarzotti, E. (2024). Forest planning, rural practices, and woodland cover in an 18th-century Alpine Valley (Val di Fiemme, Italy): A geohistorical and GIS-based approach to the history of environmental resources. *AIMS Geosciences*, *10*(4), 767–791. https://doi.org/10.3934/geosci.2024038

Gingrich, S., Lauk, C., Niedertscheider, M., Pichler, M., Schaffartzik, A., Schmid, M., Magerl, A., Le Noë, J., Bhan, M., & Erb, K. (2019). Hidden emissions of forest transitions: A socio-ecological reading of forest change. *Current Opinion in Environmental Sustainability*, *38*, 14–21. https://doi.org/10.1016/j.cosust.2019.04.005

Grava, M., Berti, C., Gabellieri, N., & Gallia, A. (2020). *Historical GIS: Strumenti digitali per la geografia storica in Italia*. EUT, Edizioni Università Trieste.

Gregory, I. N., DeBats, D. A., & Lafreniere, D. (2018a). Introduction to Part IV. In I. N. Gregory, D. A. DeBats, & D. Lafreniere, *The Routledge Companion to Spatial History* (pp. 351–352). Routledge.

Gregory, I. N., DeBats, D. A., & Lafreniere, D. (2018b). *The Routledge Companion to Spatial History*. Routledge.

Gregory, I. N., & Ell, P. S. (2007). *Historical GIS: Technologies, Methodologies, and Scholarship*. Cambridge University Press.

Gregory, I. N., & Geddes, A. (A c. Di). (2014). *Toward Spatial Humanities: Historical GIS and Spatial History*. Indiana University Press.

Grewe, B.-S. (2014). Forest History. In F. Uekötter (A c. Di), *The Turning Points of Environmental History* (pp. 44–54). University of Pittsburgh Press.

Iriarte-Goñi, I. (2013). Forests, Fuelwood, Pulpwood, and Lumber in Spain, 1860–2000: A Non-Declensionist Story. *Environmental History*, *18*(2), 333–359. JSTOR.

Kaplan, J. O., Krumhardt, K. M., & Zimmermann, N. (2009). The prehistoric and preindustrial deforestation of Europe. *Quaternary Science Reviews*, *28*(27–28), 3016–3034. https://doi.org/10.1016/j.quascirev.2009.09.028

Knoll, M. (2006). Urban Needs and Changing Environments. Regensburg' s Wood Supply between the Early Modern Period and Industrialization. *Bulletin of the German Historical Institute Washington*, (3), 7–101.

Knowles, A. K. (2005). Emerging Trends in Historical Gis. *Special Issue: Emerging Trends in Historical GIS. Historical Geography*, (33), 7–13.

Knowles, A. K., & Hillier, A. (2008). Preface. In A. K. Knowles (A c. Di), *Placing History: How Maps, Spatial Data, and GIS Are Changing Historical Scholarship* (1st ed, pp. 1–25). ESRI Press.

Lefebvre, H. (1991). *The Production of Space*. Blackwell.

Loran, C., Ginzler, C., & Bürgi, M. (2016). Evaluating forest transition based on a multi-scale approach: Forest area dynamics in Switzerland 1850–2000. *Regional Environmental Change*, *16*(6), 1807–1818. https://doi.org/10.1007/s10113-015-0911-1

Lotz, C. (2015). Expanding the Space for Future Resource Management: Explorations of the Timber Frontier in Northern Europe and the Rescaling of Sustainability During the Nineteenth Century. *Environment and History*, *21*(2), 257–279.

Massey, D. (1995). *Spatial Divisions of Labour*. Macmillan Education UK. https://doi.org/10.1007/978-1-349-24059-3

Munteanu, C., Kuemmerle, T., Keuler, N. S., Müller, D., Balázs, P., Dobosz, M., Griffiths, P., Halada, L., Kaim, D., Király, G., Konkoly-Gyuró, É., Kozak, J., Lieskovsky, J., Ostafin, K., Ostapowicz, K., Shandra, O., & Radeloff, V. C. (2015). Legacies of 19th Century Land Use Shape Contemporary Forest Cover. *Global Environmental Change*, *34*, 83–94. https://doi.org/10.1016/j.gloenvcha.2015.06.015

Murrieta-Flores, P., & Martins, B. (2019). The Geospatial Humanities: Past, Present and Future. *International Journal of Geographical Information Science*, *33*(12), 2424–2429. https://doi.org/10.1080/13658816.2019.1645336

Särkkä, T. (2021). *Paper and the British Empire: The quest for imperial raw materials, 1861-1960*. Routledge.

Statuto, D., Cillis, G., & Picuno, P. (2017). Using Historical Maps within a GIS to Analyze Two Centuries of Rural Landscape Changes in Southern Italy. *Land*, *6*(3), 65. https://doi.org/10.3390/land6030065

Tuan, Y.-F. (2002). *Space and Place: The Perspective of Experience*. University of Minnesota press.

Von Lünen, A., & Travis, C. (A c. Di). (2013). *History and GIS: Epistemologies, Considerations and Reflections*. Springer Netherlands. https://doi.org/10.1007/978-94-007-5009-8

Le nazioni genovesi nell'Atlante digitale di Storia Marittima del Regno di Sardegna (ASMSA). Lavori in corso

GIAMPAOLO SALICE[1], GIOMMARIA CARBONI[2]
[1] UNIVERSITÀ DEGLI STUDI DI CAGLIARI, ITALIA, GIAMPAOLO.SALICE@UNICA.IT
[2] UNIVERSITÀ DEGLI STUDI DI CAGLIARI, ITALIA, GIOMMARIA.CARBONI@UNICA.IT

ABSTRACT (ITALIANO)

Dalla fine del Quattrocento, precisi accordi tra la Repubblica di Genova e i Sovrani cattolici consentivano alla Superba di nominare propri consoli negli avamposti della monarchia iberica di maggiore rilevanza strategica e commerciale. Fra Cinque e Seicento, l'ufficio consolare si stabilizzava anche nel Regno di Sardegna, entrando in piena operatività nelle sette città regie. Qui il console diventava presto il perno di un sistema reticolare dai significati sociale, economico e politico, intorno al quale la presenza ligure assumeva forme istituzionalizzate, sottoposte a continua negoziazione. Le comunità genovesi sono state sinora studiate singolarmente, senza essere inserite in un quadro organico capace di valorizzare il ruolo dei consoli in un più ampio disegno di controllo della Repubblica sui sudditi genovesi all'estero. L'esame delle corrispondenze tra il console a Cagliari e il Senato di Genova, lette insieme ad altre fonti (ecclesiastiche, processuali e notarili), consente di ricostruire le reti relazionali che definivano gli equilibri di forza tra la "nazione" genovese cittadina, la sua arciconfraternita e il console, aprendo a una rilettura di tale esperienza di radicamento forestiero in terra d'altri.
Questo contributo dà conto degli impatti interpretativi scaturiti dall'impiego dell'ecosistema digitale ASMSA (Atlante digitale di Storia Marittima del Regno di Sardegna) per la metadazione, la spazializzazione e la visualizzazione della documentazione summenzionata. L'ASMSA consente di ordinare semanticamente, nello spazio e nel tempo, i dati estratti dalle diverse tipologie di fonti disponibili, integrandone l'interpretazione con strumenti di network analysis digitale prodotti anche con l'ausilio della AI. Ne derivano soluzioni analitiche che aumentano la capacità di cogliere relazioni e che offrono elementi utili alla problematizzazione e all'interpretazione del fenomeno studiato.
Parole chiave: Repubblica di Genova; Regno di Sardegna; Storia marittima; Storia digitale; console; nazione; network analysis

ABSTRACT (ENGLISH)

The Genoese Nations within the Digital Atlas of Maritime Sardinian history (ASMSA): A Work in Progress.
From the late fifteenth century onwards, specific agreements between the Republic of Genoa and the Catholic Monarchs allowed La Superba to appoint its own consuls in the most strategically and commercially significant outposts of the Iberian monarchy. Between the sixteenth and seventeenth centuries, the consular office also became firmly established in the Kingdom of Sardinia, reaching full operation in the seven royal cities and the principal maritime ports. There, the consul soon became the pivot of a networked system with social, economic, and political implications, around which the Ligurian presence assumed institutionalized forms subject to constant negotiation. Genoese communities have so far been studied individually, without being placed within an organic framework capable of highlighting the role of the consulates within a broader strategy of control exercised by the Republic over Genoese citizens abroad. The examination of the consular correspondence between Cagliari and the Senate of Genoa, read alongside other sources (ecclesiastical, judicial, and notarial), makes it possible to reconstruct the relational networks that defined the balance of power between the Genoese "nation," its archconfraternity, and the consul, opening the way to a reinterpretation of this experience of foreign settlement in another land.
The article also discusses the interpretative implications arising from the use of the ASMSA digital ecosystem (Digital Atlas of Maritime History of the Kingdom of Sardinia) for the metadata enrichment, spatialization, and visualization of the research. ASMSA enables the semantic, spatial, and chronological organization of data extracted from different types of available sources, integrating their interpretation with digital network analysis tools. The development of these tools combined the research group's workflow with new analytical solutions obtained through artificial intelligence, increasing the capacity to

* Sono da attribuire a G. Carboni i paragrafi 1,3,4; a G. Salice il paragrafo 2

identify relationships and providing useful elements for the problematization and interpretation of the phenomenon under study..
Keywords: Repubblica di Genova; Regno di Sardegna; Storia marittima; Storia digitale; console; nazione; network analysis

1. IL TEMA DI RICERCA

Sin dal Medioevo l'espansione mercantile di Genova ha permesso il proliferare per tutto il Mediterraneo e sino al Mar Nero di insediamenti di mercanti organizzati per colonie o "nazioni" sotto la guida diretta o indiretta del governo genovese (Balard, 2004; Petti Balbi, 2005). Mercanti e armatori genovesi erano presenti a vario titolo anche in Sardegna. Tuttavia, è solo a seguito della stipula dei trattati di pace del 5 agosto 1493 tra Comune di Genova e Re cattolici, e in particolare i capi XIV e XX,[1] confermati successivamente il 20 marzo 1519,[2] che si avviava la liberalizzazione degli scambi mercantili coi genovesi, ai quali era inoltre riconosciuta la facoltà di nominare un console competente nelle cause riguardanti i nazionali genovesi (Lisciandrelli, 1960: 176–177). È quindi solo dalla fine del Quattrocento, ma è un processo che si stabilizza pienamente entro la fine del secolo successivo, che le nazioni genovesi in Sardegna prendono forma come comunità mercantili organizzate attorno a un loro rappresentante e a una chiesa nazionale (Carboni, 2025 e bibliografia precedente) in corrispondenza delle città e dei principali porti caricatori dell'isola (per un elenco dei consolati genovesi si veda Vitale, 1934). I rapporti tra Genova e la monarchia asburgica si faranno più stretti a partire dall'alleanza con Andrea Doria; così, dal 1528 le sorti di Genova e dei suoi *hombre de negocios* rimarranno legate a doppio filo alla parabola del sistema imperiale spagnolo (Álvarez Nogal, 2001; Herrero Sánchez, 2001; Franch, 2005). Tuttavia, già nel 1533 il sovrano Carlo V doveva raccomandare al viceré a Cagliari il rispetto dei capitoli che disciplinavano il diritto di nomina e rimozione dei consoli genovesi, evidentemente frustrati dagli abusi dell'amministrazione regia.[3] È la spia di un confronto giurisdizionale frequente, tra gli oggetti principali delle missive che i consoli spedivano al Senato di Genova e oggi custodite all'Archivio di Stato della città ligure.[4]
Lo spoglio della documentazione menzionata insieme con fonti notarili, giudiziarie e di ambito ecclesiastico (registri parrocchiali, fondi confraternali e fondi capitolari) sta consentendo di sia definire in maniera puntuale le prerogative e il perimetro d'azione dell'ufficio consolare e dei suoi rapporti con le istituzioni locali sia di affrontare alcune questioni storiografiche significative: la figura del console può essere usata come cartina di tornasole per la comprensione dei processi di fabbricazione delle comunità nazionali? Studiarla può consentirci di cogliere la natura polimorfa, relazionale e negoziata della comunità nazionale che egli rappresenta, nonché i diversi status di "cittadinanza" che si acquistano stabilendosi in terra d'altri? Tempi e strumenti della fabbricazione di comunità o di negoziazione degli statuti dei forestieri sono temi già affrontati dalla storiografia, ma in rapporto ad altre comunità, come quelle greche (Grenet, 2010), o in riferimento a spazi portuali come Livorno (Calafat, 2012). Mancano studi analoghi alla scala del Regno di Sardegna ed è da questa lacuna che muove il presente lavoro.

2. L'ATLANTE DIGITALE DI STORIA MARITTIMA DEL REGNO DI SARDEGNA (ASMSA)

Per affrontare la ricerca si è scelto di operare all'interno dell'Atlante Digitale per la Storia Marittima della Sardegna (ASMSA).[5] ASMSA è un progetto di storia digitale che, adottando una prospettiva multidisciplinare, punta alla ricostruzione della dimensione marittima della Sardegna nella lunga durata. Nato nel 2020 grazie a un finanziamento della Fondazione di Sardegna, il progetto è sviluppato in seno al DH UniCa - Centro Interdipartimentale per l'Umanistica Digitale dell'Università degli Studi di Cagliari[6], sotto la supervisione di Giampaolo Salice. L'ASMSA invita a superare la frammentarietà delle fonti e degli studi sulla storia marittima sarda, offrendo un'unica piattaforma digitale pensata per integrare bibliografia, documentazione archivistica, cartografia storica e materiali multimediali (anche prodotti da

[1] Archivio di Stato di Genova (d'ora in avanti ASGe), *Archivio Segreto, Materie Politiche*, 2733, mazzo 14, fascicolo 61.
[2] ASGe, *Archivio Segreto, Materie Politiche*, 2734, mazzo 15, fascicolo 30.
[3] ASGe, *Archivio Segreto, Materie Politiche*, 2734, mazzo 15, fascicolo 81.
[4] Si vedano le serie "Lettere consoli, Sardegna" (unità 2668, 2669) del fondo dell'Archivio Segreto e "Consoli nazionali ed esteri" (unità 1-6) del fondo della Giunta di Marina dell'Archivio di Stato di Genova.
[5] https://storia.dh.unica.it/asmsa/it (cons. 07/02/2026).
[6] https://dh.unica.it/ (cons. 07/02/2026).

filoni di indagine differenti e a diverse scale spaziali e temporali) al fine di favorire la loro interoperabilità, comparabilità e riuso.
Sotto il profilo infrastrutturale l'Atlante opera attraverso tre applicativi: Drupal, Omeka-S (che consente di descrivere documenti archivistici, pubblicazioni a stampa, immagini e/o fotografie storiche, audiovideo, carte geografiche ecc.) e GeoNode (che pubblica i progetti cartografici elaborati con QGIS). Drupal permette invece di comporre pagine web arricchite coi contenuti richiamati da Omeka-S, organizzabili in gallerie, linee temporali e li può spazializzare sulle mappe interattive pubblicate da GeoNode., sulle quali è inoltre possibile sovrapporre uno o più layer WMS esterni, ampliando così le possibilità di rappresentazione, interrogazione e contestualizzazione spaziale dei dati (Salice, 2023).
Tutto questo è possibile grazie a una serie di *widget* – sviluppati nel DH UNICA - che estendono le funzionalità native del *Layout Builder* di Drupal. L'ecosistema è basato su tecnologie *Docker* e *Git* ed è concepito in conformità ai principi dell'*Infrastructure as Code*, che hanno consentito di predisporre ambienti di lavoro versionati, replicabili e facilmente manutenibili. L'Atlante si avvale anche della suite di strumenti hardware del DH UniCa che consentono di dematerializzazione fonti bibliografiche, archivistiche, cartografiche anche in forma massiva.[7] L'intera documentazione tecnica relativa allo sviluppo dell'infrastruttura DOG è pubblicamente accessibile nel repository GitHub del DH UniCa.[8]
Pertanto l'ASMSA si configura non tanto quale strumento di disseminazione, quanto soprattutto come ecosistema strutturato, fondato su specifici modelli di rappresentazione, di standard descrittivi e interpretativi. All'interno di questa cornice, ASMSA ha progressivamente incorporato i risultati di percorsi di ricerca dottorale[9] e post-dottorale[10], configurandosi come ambiente digitale di elaborazione, organizzazione e restituzione di dati storici complessi. L'atlante è inoltre utilizzato nella Scuola di Specializzazione in Beni Architettonici e del Paesaggio (SSBAP) UniCa, per attività formative dedicate allo studio e alla documentazione delle torri costiere. L'integrazione di tali ricerche con quelle messe in campo da ricercatori esperti ha ampliato il campo applicativo dell'Atlante, confermandone la funzione come infrastruttura scientifica capace di sostenere processi di modellazione, interoperabilità e valorizzazione dei dati e di rafforzare il legame tra ricerca avanzata, didattica specialistica e pratiche di conoscenza del patrimonio territoriale. L'Atlante consente, infatti, di attivare programmi di Storia pubblica finalizzati al coinvolgimento attivo delle comunità nella produzione condivisa di diverse forme di conoscenza. Tale processo può comprendere il recupero della microtoponomastica tramandata dalla memoria collettiva, la raccolta di fotografie storiche conservate negli archivi privati, nonché l'identificazione diretta di elementi architettonici e di pratiche legate alla cultura del mare, tanto nei suoi aspetti materiali quanto in quelli immateriali.

3. VISUALIZZARE LE NAZIONI GENOVESI NELLO SPAZIO. UN PROGETTO IN DIVENIRE

In questa sede discuteremo di come e perché la ricerca dottorale[11] di cui al paragrafo 1 è stato incorporato nell'Atlante e dei primi risultati che questa azione sta conseguendo.
Se le nazioni genovesi nel Regno di Napoli sono state già ampiamente indagate (Musi, 1996; Brancaccio, 2001; Ben Yessef Garfia, 2025), e se per Palermo disponiamo della pubblicazione integrale dei privilegi del consolato genovese (Basile, 2007), più numerosi e tuttavia più frammentari risultano gli studi sui genovesi in Sardegna, a partire dalla pubblicazione degli statuti inediti dell'Arciconfraternita dei Genovesi di Cagliari (Zedda Macciò, 1974), dalle ricerche sull'economia sarda del Seicento e sull'attività mercantile

[7] https://dh.unica.it/scanner/ (cons. 08/02/2026)
[8] https://github.com/Dh-unica (cons. 08/02/2026)
[9] Tesi dottorali di Beatrice Schivo, *Schiavi. Una piattaforma digitale per la storia della schiavitù nel Regno di Sardegna (secc. XVII-XVIII)* e di Filippo Astori, *Dalla Sardegna all'Oceano Indiano: il corallo nelle reti di scambio globale ispano-genovesi (secc. XVI-XVIII)*, Dottorato Internazionale in Storia, Beni Culturali e Studi Internazionali, Indirizzo Studi Storici Euromediterranei, dall'antichità all'età Contemporanea, Università degli Studi di Cagliari, Ciclo XXXVII. Cfr. rispettivamente https://storia.dh.unica.it/asmsa/it/esplora-latlante/schiave-e-schiavi; https://storia.dh.unica.it/asmsa/it/corallo-e-corallatori (cons. 08/02/2026)
[10] Nell'ambito della Ayudas "Margarita Salas" (2021-2023), Álvaro Casillas ha condotto tra Università di Cagliari e Complutense de Madrid una ricerca dal titolo *El reino de Cerdeña bajo asedio. Corsarismo, información y defensa en el Mediterráneo occidental (1618-1648)*, impiegando l'ASMSA per esaminare i riferimenti alla guerra corsara contenuti in sei portolani prodotti in ambito ispanico nella prima metà del XVII secolo. Cfr.https://storia.dh.unica.it/asmsa/it/rotte-e-portolani/portolani-della-monarchia-ispanica (cons. 08/02/2026).
[11] *Le colonie genovesi di Cagliari e Napoli. Due comunità mercantili a confronto (secc. XVII-XVIII)*, Dottorato Internazionale in Storia, Beni Culturali e Studi Internazionali, Indirizzo Studi Storici Euromediterranei, dall'antichità all'età Contemporanea, Università degli Studi di Cagliari, Ciclo XL.

dei liguri (Sorgia, 1983; Anatra, 1983; Plaisant, 1989), confluite nel volume Genova in Sardegna (Saiu Deidda, 2000), e dagli studi più recenti sulle nazioni genovesi (Mele, 2014) e sulla rete commerciale ligure (Mele, 2011). Manca ancora, per la Sardegna, uno studio che prenda in considerazione la figura del console quale elemento di raccordo ufficiale fra colonie in un'ottica translocale, come già ampiamente individuato negli studi sui consoli genovesi quali agenti di informazione nel Mediterraneo (Lo Basso, 2013; Beri, 2016).

Per colmare questa lacuna si è deciso in primis di visualizzare su mappa le nazioni genovesi, così da aumentare la capacità di lettura dei processi storici a diverse scale analitiche, dalla dimensione locale a quella translocale, dal singolo evento a riflessioni sul processi di media e lunga durata. Si è inoltre adottato un approccio quantitativo per la ricostruzione della consistenza numerica delle colonie nel tempo, da affiancare a un'analisi qualitativa dei dati: se gli archivi diocesani permettono di quantificare la presenza genovese suddivisa per parrocchie, le stesse presenze inserite nello spazio cittadino consentono di visualizzarne la distribuzione rispetto alla chiesa "nazionale", alle botteghe artigiane e agli spazi delle istituzioni. Ancora, l'analisi dei nominativi attraverso database relazionali possono mostrarci il grado di partecipazione dei genovesi ai sodalizi, ai culti, alle diverse forme di socialità sia nazionali che locali. La quantificazione delle interazioni individua reti relazionali grazie agli strumenti della *network analysis*.
Per la visualizzazione e l'analisi di questi dati è stato sviluppato, con l'ausilio dell'intelligenza artificiale, uno strumento digitale interattivo che che consente di esplorare la rete sociale della comunità genovese a Cagliari attraverso un'interfaccia web. I dati sono raccolti e organizzati in due fogli di calcolo: uno relativo ai registri matrimoniali delle quattro parrocchie storiche della città, l'altro ai verbali delle congregazioni dell'Arciconfraternita dei Santi Martiri Giorgio e Caterina. La scelta del matrimonio non è casuale: il matrimonio è uno dei dispositivi con cui il forestiero si naturalizza, inserendosi nelle reti di parentela e di prossimità della società locale. Il registro matrimoniale si configura insomma come un osservatorio privilegiato sui meccanismi di radicamento nella comunità. Il confronto con la documentazione confraternale, e in particolare i registri delle adunanze, permette così di individuare quali genovesi appartengono al sodalizio e quali ne sono esclusi. Inoltre, il calcolo delle percentuali di provenienza dei confratelli – disaggregate per congregazione e incrociate con i dati matrimoniali – apre la strada a riflessioni sulle dinamiche conflittuali e negoziali interne alla nazione genovese e consente di individuare le "patrie" di cui si compone: quella alassina, quella genovese in senso stretto, quella delle comunità della Riviera di Ponente, ciascuna con le proprie reti di *patronage*, le proprie strategie matrimoniali e i propri percorsi di radicamento nel contesto cagliaritano.
Dal punto di vista tecnico, lo strumento è distribuito come applicazione web autonoma – un singolo file HTML – che carica i dati dinamicamente attraverso le API pubbliche di Google Drive, garantendo che qualsiasi aggiornamento dei dati caricati sui fogli di calcolo si rifletta immediatamente sulla visualizzazione. La lettura dei file Excel è affidata alla libreria SheetJS, mentre la costruzione della rete si avvale di D3.js, che implementa un algoritmo di forze per disporre nodi e archi secondo criteri di leggibilità e densità relazionale. Le persone identificate nei documenti costituiscono i nodi del grafo – la cui dimensione è proporzionale al numero di relazioni intrattenute – mentre i legami documentati formano gli archi, distinti per tipo (matrimoniale, testimoniale, confraternale), direzione e colore. Un sistema di filtri consente di interrogare la rete per parrocchia, congregazione, luogo di origine e arco cronologico, con una soglia di compatibilità biografica di quarant'anni per escludere automaticamente i casi di omonimia.

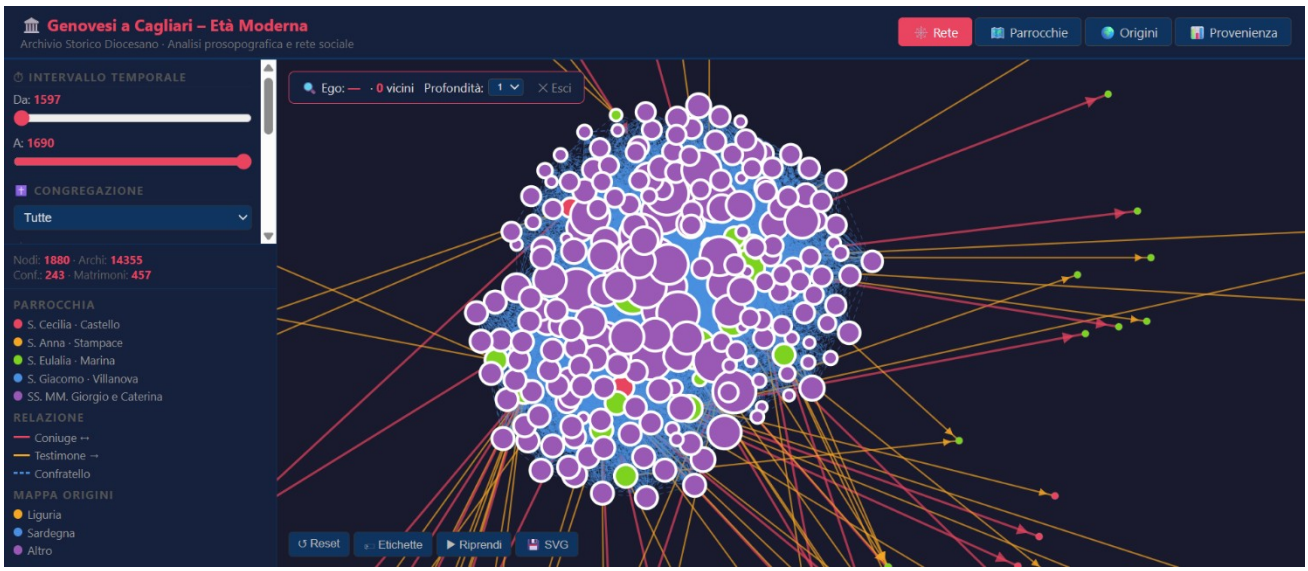


Fig. 1 – La rete relazionale della comunità genovese a Cagliari (vista "Rete")

La componente cartografica, realizzata con Leaflet.js su base OpenStreetMap, geolocalizza la comunità sulle chiese storiche di Cagliari, con le sedi delle cinque istituzioni religiose rilevanti posizionate mediante coordinate di precisione.

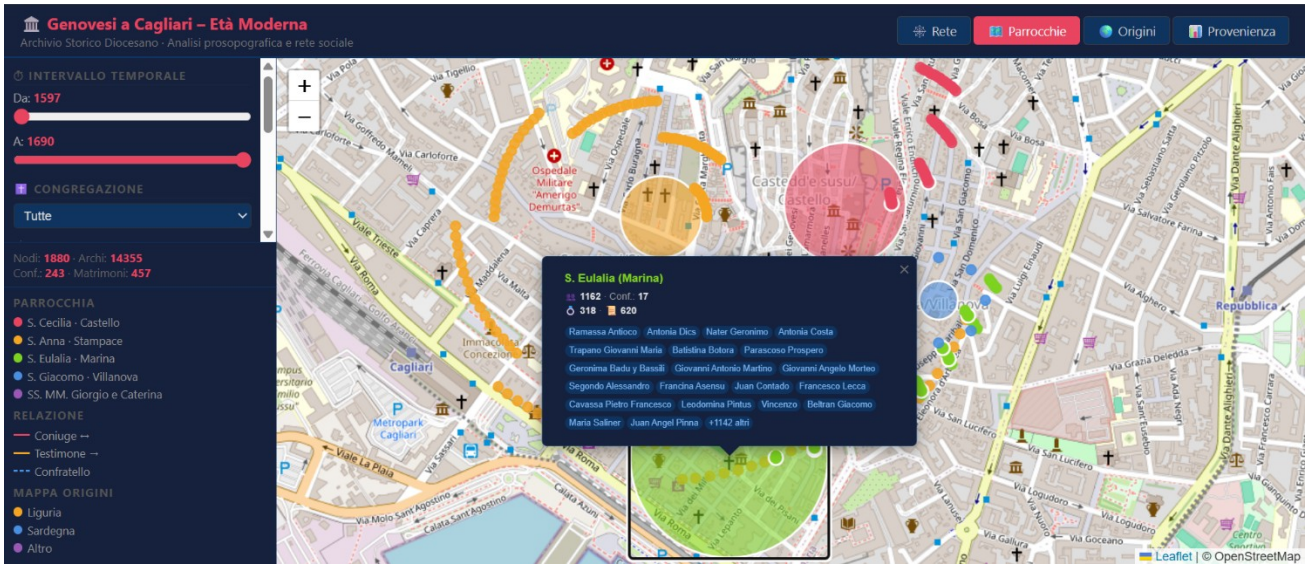


Fig. 2 – La distribuzione spaziale della presenza genovese nel tessuto urbano di Cagliari (vista "Parrocchie")

La stessa componente visualizza su scala mediterranea la distribuzione geografica delle provenienze, restituendo una mappa delle origini dei membri della comunità genovese cagliaritana.

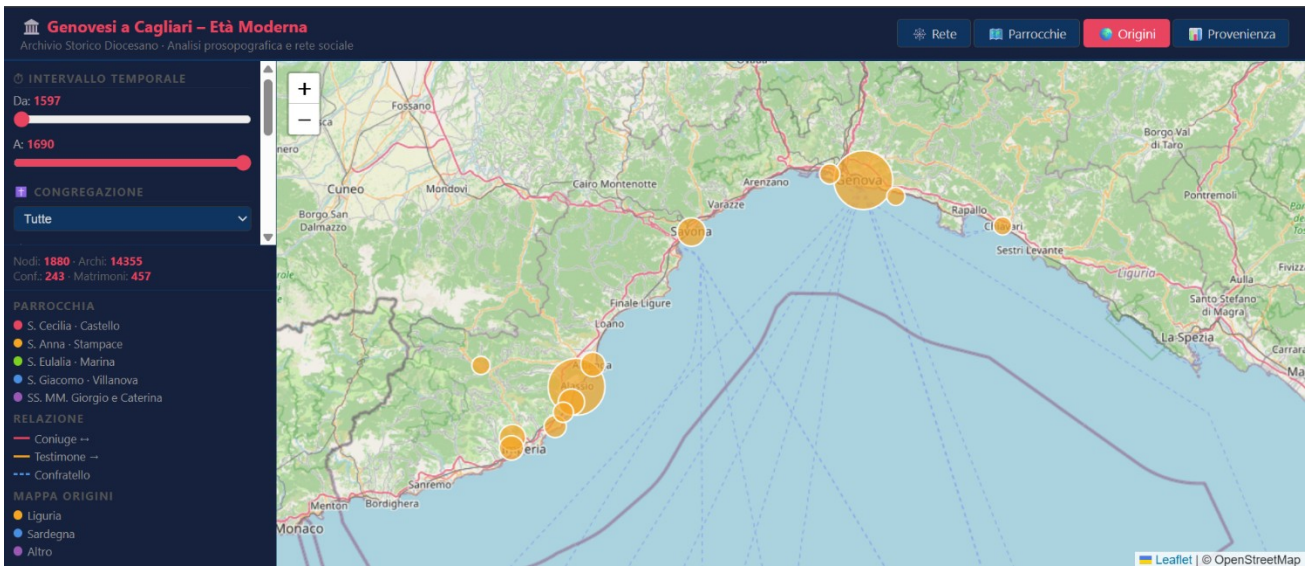


Fig. 3 – La distribuzione geografica delle provenienze su scala mediterranea (vista "Origini")

Prima della costruzione del grafo, i dati sono stati sottoposti a una normalizzazione automatica dei toponimi, che riconduce le varianti grafiche documentarie alle denominazioni attuali. Per colmare le lacune nella documentazione in rapporto alla provenienza geografica dei confratelli – registrata sistematicamente solo in una delle quindici congregazioni finora caricate – i dati confraternali sono stati incrociati con quelli matrimoniali attraverso un algoritmo di *matching* che valuta la compatibilità temporale tra le attestazioni e classifica i risultati per livello di certezza, segnalando esplicitamente i casi in cui la copertura è insufficiente per trarre conclusioni statisticamente robuste.



Fig. 4 – L'analisi delle provenienze per congregazione con incrocio dei dati matrimoniali (vista "Provenienza")

Lo strumento è progettato per integrarsi pienamente nell'ecosistema ASMSA: i dati elaborati confluiscono in schede di metadatazione prodotte in Omeka-S, i riferimenti spaziali diventano interoperabili con i *layer* geografici pubblicati in Geonode, e l'ambiente Drupal permette di incorporare lo strumento come componente interattiva dell'Atlante, collegandolo ad altri progetti e *dataset* già presenti nella piattaforma. Lo strumento si configura dunque come parte di un sistema più ampio, tanto più utile nel quadro d'insieme che punta allo studio della dimensione marittima del Regno di Sardegna, in cui i liguri ebbero un ruolo cruciali: nelle concessioni per l'esportazione del grano, per la pesca del corallo[12] o per lo

[12] https://storia.dh.unica.it/asmsa/it/corallo-e-corallatori (cons. 08/02/2026).

sfruttamento di tonnare,[13] peschiere[14] e saline,[15] tutti ambiti a cui l'Atlante dedica una specifica sezione di approfondimento. L'ASMSA è, in definitiva, un ecosistema di ricerche interconnesse: l'inserimento di nuovi dati e rappresentazioni produce un impatto euristico più ampio, favorendo confronti, sovrapposizioni e riletture trasversali.

4. PRIMI RISULTATI

I dati finora caricati – le quindici congregazioni dell'Arciconfraternita dei Santi Martiri Giorgio e Caterina comprese tra il giugno 1599 e l'ottobre 1604, e gli atti matrimoniali relativi a persone genovesi nelle quattro parrocchie storiche – consentono di formulare alcune considerazioni preliminari, in attesa dell'inserimento dei dati relativi alle restanti congregazioni del Seicento. I risultati qui presentati hanno pertanto carattere provvisorio.
La congregazione del 4 marzo 1600, unica per la quale la fonte registra sistematicamente la provenienza, mostra una composizione dominata dagli alassini (61.3%, 46 individui su 75 identificati), seguiti da Genova città (13.3%) e da contingenti minori di Laigueglia, Oneglia, Diano Marina e altri centri della Riviera di Ponente. Nei registri matrimoniali, la distribuzione per parrocchia vede la concentrazione maggiore di genovesi a Sant'Eulalia (18 atti su 30), dato coerente con la funzione portuale e commerciale del quartiere Marina e con la prossimità alla chiesa "nazionale", sede dell'Arciconfraternita. Sul piano delle provenienze estrapolate dagli atti matrimoniali, la metà degli sposi è registrata con la sola formula "genovese" – indicante genericamente la provenienza dai territori della Repubblica, senza specificazione della località di origine – rendendo difficile qualsiasi confronto diretto coi registri confraternali.
Il peso preponderante di Alassio nell'Arciconfraternita è un dato rilevante, che apre interrogativi sulla natura di questa presenza: se si tratti cioè di un primato numerico casuale o piuttosto del riflesso di una capacità organizzativa e di una rete di solidarietà interna particolarmente consolidata, tale da garantire agli alassini un controllo sulle strutture di rappresentanza della nazione. Il completamento del caricamento dei dati confraternali del resto del Seicento, incrociato con i registri matrimoniali sull'intero arco cronologico, consentirà di verificare la tenuta di questi equilibri nel tempo e di restituire un quadro più preciso delle dinamiche interne alla nazione genovese a Cagliari.

BIBLIOGRAFIA

Álvarez Nogal, C. (2001). I genovesi e la monarchia spagnola tra Cinque e Seicento. Atti della Società Ligure di Storia Patria, n.s., XLI (CXV)(2), 107–121.

Anatra, B. (1983). Aspetti della congiuntura seicentesca in Sardegna. Studi di storia moderna e contemporanea, Annali della Facoltà di Magistero dell'Università degli Studi di Cagliari. Istituto di Studi Storici, (23), 5–44.

Balard, M. (2004). Vendere nel dominio e fuori: Botteghe di città e colonie mercantili. In D. Puncuh (A c. Di), Storia della cultura ligure: II (pp. 99–116). Società Ligure di Storia Patria.

Basile, M. (2007). Una natio straniera nella Sicilia medievale e moderna. I Privilegi del Consolato di Genova a Palermo. Rubbettino.

Ben Yessef Garfia, Y. R. (2025). Ser noble y genovés en Nápoles. Formas de servicio y de identidad en la Monarquía Hispánica (siglos XVI-XVII). In A. Cogné & B. Maréchaux (A c. Di), Le élites italiane e la Monarchia ispanica (secoli XVI-XVII). Servizio, mobilità e poteri (pp. 179–205). Publications de l'École française de Rome.

Beri, E. (2016). I consoli genovesi del Tirreno, agenti d'informazione (1640-1797). RiMe. Rivista dell'Istituto di Storia dell'Europa Mediterranea, 17(2), 153–188.

Brancaccio, G. (2001). Nazione genovese. Consoli e colonia nella Napoli moderna. Guida Editori.

Calafat, G. (2012). Être étranger dans un port franc. Droits, privilèges et accès au travail à Livourne (1590-1715), «Cahiers de la Méditerranée»/84, pp. 103–122.

Carboni, G. (2025). I Genovesi di Cagliari. Arciconfraternita o Nazione? In G. T. Colesanti, F. Scaduto, & M. Schirru (A c. Di), Le Confraternite nell'Europa medioevale e moderna. Architettura – Archivi – Arte – Storia – Tutela – Valorizzazione del patrimonio (pp. 99–106). Editoriale Scientifica.

[13] https://storia.dh.unica.it/asmsa/it/tonnare-nel-regno-di-sardegna (cons. 08/02/2026).
[14] https://storia.dh.unica.it/asmsa/it/esplora-latlante/pesca-e-pescatori (cons. 08/02/2026).
[15] https://storia.dh.unica.it/asmsa/it/esplora-latlante/sale-saline-salinieri (cons. 08/02/2026).

Franch, R. (2005). Los Genoveses en la España moderna: Finanzas, comercio y actividad laboral de los protagonistas de un intenso flujo migratorio. In L. Gallinari (A c. Di), Genova. Una «porta» del Mediterraneo: II (Vols. 1–2, pp. 643–683). Brigati.

Grenet, M. (2016). La fabrique communautaire. Les Grecs à Venise, Livourne et Marseille (1770-1840). École Française d'Athènes- École Française de Rome.

Herrero Sánchez, M. (2001). Una república mercantil en la órbita de la monarquía católica (1528-1684). Hegemonía y decadencia del agregado hispano-genovés. In B. Anatra & F. Manconi (A c. Di), Sardegna, Spagna e Stati italiani nell'età di Carlo V (pp. 183–200). Carocci.

Lisciandrelli, P. (1960). Trattati e negoziazioni politiche della Repubblica di Genova (958-1797). Regesti: I. Società Ligure di Storia Patria.

Lo Basso, L. (2013). Consoli e corsari. La corrispondenza consolare come fonte per la guerra corsara nel Mediterraneo tra XVII e XVIII secolo. In M. Aglietti, M. Herrero Sánchez, & F. J. Zamora Rodríguez (A c. Di), Los cónsules de extranjeros en la Edad Moderna y a principios de la Edad Contemporánea (pp. 177–186). Doce Calles.

Mele, G. (2011). La rete commerciale ligure in Sardegna nella prima metà del XVII secolo. In M. Herrero Sánchez, Y. R. Ben Yessef Garfia, C. Bitossi, & D. Puncuh (A c. Di), Génova y la Monarquía Hispánica (1528-1713): I (pp. 203–218). Società Ligure di Storia Patria. https://www.storiapatriagenova.it/Scheda_vs_info.aspx?Id_Scheda_Bibliografica=1710

Mele, G. (2014). Formaggi e corallo. La colonia ligure di Bosa nel XVII secolo. Bollettino di Studi Sardi, 7, 87–110. https://doi.org/10.13125/bss-4817

Musi, A. (1996). Mercanti genovesi nel Regno di Napoli. Edizioni scientifiche italiane.

Petti Balbi, G. P. (2005). Negoziare fuori patria. Nazioni e genovesi in età medievale. CLUEB.

Plaisant, M. L. (1989). Attività mercantili e imprenditoriali dei Genovesi in Sardegna nel secolo XVII. In R. Belvederi (A c. Di), Rapporti Genova, Mediterraneo, Atlantico nell'eta moderna. Atti del III° Congresso internazionale di studi storici, Genova 1987 (pp. 519–529). s.e.

Saiu Deidda, A. (A c. Di). (2000). Genova in Sardegna: Studi sui genovesi in Sardegna fra Medioevo ed età contemporanea. CUEC. (Biblioteca comunale Generale e di Studi Sardi).

Salice, G. (2023). ASMSA. Atlante digitale per la storia marittima della Sardegna. Umanistica Digitale, (15), Articolo 15. https://doi.org/10.6092/issn.2532-8816/16660

Sorgia, G. (1983). Genova, Sardegna e Spagna nel Cinquecento. In R. Belvederi (A c. Di), Rapporti Genova, Mediterraneo, Atlantico nell'eta moderna. Atti del Congresso Internazionale di studi storici (pp. 113–128). s.e.

Vitale, V. (1934). Diplomatici e consoli della Repubblica di Genova: LXIII. Società Ligure di Storia Patria.

Zedda Macciò, I. (1974). L'Arciconfraternita dei Genovesi in Cagliari nel sec. XVII (da documenti inediti dei sec. XVI e XVII). Arciconfraternita dei Genovesi.

Integrating Romanian Toponymy in a GIS Background: dictionary forest names vs. OpenStreetMap forest data

ROXANA PATRAS[1], ANA ODOCHICIUC[2], MIHAI-BOGDAN ATANASIU[3], CONSTANTIN RĂCHITĂ[4]
[1]"ALEXANDRU IOAN CUZA" UNIVERSITY OF IAŞI, INSTITUTE OF INTERDISCIPLINARY RESEARCH, ROXANA.PATRAS@UAIC.RO
[2]"ALEXANDRU IOAN CUZA" UNIVERSITY OF IAŞI, INSTITUTE OF INTERDISCIPLINARY RESEARCH, ANA.HONCU@UAIC.RO
[3]"ALEXANDRU IOAN CUZA" UNIVERSITY OF IAŞI, INSTITUTE OF INTERDISCIPLINARY RESEARCH, BOGDAN.ATANASIU@UAIC.RO
[4]"ALEXANDRU IOAN CUZA" UNIVERSITY OF IAŞI, INSTITUTE OF INTERDISCIPLINARY RESEARCH, CONSTANTIN.RACHITA@UAIC.RO

ABSTRACT (ENGLISH)

In this paper, we use a sequence of toponymical data that has been extracted for a larger Balkan criminality project (focused on hajduks and hajduk life in late 18th and 19th centuries), namely, 5225 geo-referenced dendronyms (forest names) drawing from various Romanian thesauri. By overlapping this dendronym sub-set with contrasting OSM data we will be able to answer the following research questions: 1. How are points, clusters and buffers (applied to fictional criminality hot spots) positioned in comparison with the polygons vectorialized in OSM (outside area vs. inside area)? What dendronyms fall outside the vectorialized areas? Which of the two reasons - deforestation or lack of forest coverage in the current OSM database – is more plausible for inside/ outside positionings? Forests have always been a (hajduk) criminality area, which justifies a research interest in the way dendronyms correspond to the currently geo-referentiated areas. For several reasons, Romania's geographical territory provides a complex staging for this confrontation between linguistically-bound, static forest-data and spatially-bound, dynamic forest-data. One of them is that natural areas have not been vectorialized completely, which, with the growing contributions in OSM, lays seeds for a long-term inspection on phenomena such as deforestation or, even if more unlikely, forestation in various areas of Romania. Methodology-wise, we use QGIS functions (buffer, clip, raster, select location/ feature etc.) and PGAdmin platform for PostgreSQL data to carry out a spatial analysis whose hotspots have been determined by giving primacy to literary resources such as the Hai-Ro corpus. However, the present paper offers a valuable quantitative insight into the phenomenon of deforestation along modernity and provides an analysis that will serve as a preparatory argument for enlarging the toponymical database on Balkan criminality.
Keywords: GIS; forest vectorialization; dendronyms; Balkan criminality; hajduk novels

ABSTRACT (ITALIANO)

In questo articolo utilizziamo una sequenza di dati toponimici estratti da un più ampio progetto sulla criminalità balcanica (incentrato sugli hajduk/ briganti e sulla vita degli hajduk tra la fine del XVIII e il XIX secolo), ovvero 5225 dendronimi georeferenziati (nomi di foreste) tratti da vari thesauri rumeni. Sovrapponendo questo sottoinsieme di dendronimi con dati OSM contrastanti, saremo in grado di rispondere alle seguenti domande di ricerca: 1. Come si posizionano punti, cluster e buffer (applicati a punti caldi di criminalità fittizia) rispetto ai poligoni vettorializzati in OSM (area esterna vs. area interna)? Quali dendronimi ricadono al di fuori delle aree vettorializzate? Quale delle due ragioni — deforestazione o assenza di copertura forestale nell'attuale database OSM — è più plausibile per i posizionamenti interni/esterni? Le foreste sono sempre state un'area di criminalità (degli briganti), il che giustifica un interesse di ricerca nel modo in cui i dendronimi corrispondono alle aree attualmente georeferenziate. Per diverse ragioni, il territorio geografico della Romania offre una complessa messa in scena per questo confronto tra dati forestali statici di natura linguistica e dati forestali dinamici di natura spaziale. Una di esse è che le aree naturali non sono state vettorializzate completamente, il che, con i contributi crescenti in OSM, getta le basi per un'analisi a lungo termine di fenomeni come la deforestazione o, seppur più improbabile, il rimboschimento in varie aree della Romania. Dal punto di vista metodologico, utilizziamo le funzioni di QGIS (buffer, clip, raster, selezione per posizione/attributo, ecc.) e la piattaforma PGAdmin per i dati PostgreSQL, al fine di condurre un'analisi spaziale i cui hotspot sono stati determinati attribuendo la priorità a risorse letterarie come il corpus Hai-Ro. Il presente articolo offre tuttavia una

preziosa visione quantitativa del fenomeno della deforestazione nel corso della modernità e fornisce un'analisi che servirà come argomento preparatorio per l'ampliamento del database toponimico sulla criminalità balcanica.
Parole chiave: GIS; vettorializzazione forestale; dendronimi; criminalità balcanica; romanzi dei *haiduci*/ briganti

1. INTRODUCTION

The data used in this paper are part of a wider initiative that aims to scaffold a *factuality* framework for the toponymic NER of Hai-Ro, a literary corpus of 47 hajduk novels issued during the long nineteenth century (1850-1920), and counting approximatively 1.7 million words (Patraş *et al*, 2020). We state from the very beginning that our resources represent the raw materials for a forthcoming database, whose computational and infrastructural aspects, still in progress, will be clarified in terms of management and accessibility very shortly. The GIS analytical layers derived from these locally produced data and further integrations with OSM data (https://www.openstreetmap.org) and PostgreSQL data (https://www.postgresql.org/) will be made available in a WebGIS application created with React Leaflet, Node. js and Visual Studio Code as in the following example that represents the core for a more advanced interactive map on Balkan criminality: https://anaodo.github.io/webgis_Haiduci/.
For the moment, the mock version of this dataset, containing CSV, GeoJSON, txt files, is available here: DATABASE ON HAJDUK CRIMINALITY, https://doi.org/10.5281/zenodo.18712335. Since data contained therein regard engaging socio-cultural issues like deforestation, historical zones of criminality, road networking, demographics, trade values and goods and other aspects (e.g. hybrid heroism, factuality), the envisaged audiences for such database are scholarly, educational, civic, and institutional. More on the project's conceptual premises, variety of toponyms, and preliminary results is available in our recent paper "Structuring Resources for a Database of Balkan Criminality in the Long Nineteenth Century", DOI: https://doi.org/10.5334/johd.515. Contextualizing the Romanian hajduk fiction represented one of the desiderata of the Hai-Ro Project (PN-III-P3-3.1-PM-RO-FR-2019-0063), chiefly through the operationalization of concepts such as *literary spatiality* (Galleron *et al*, 2021, Piatti *et al.*, 2009: 178) and *literary factuality* (Lavocat, 2020; Rajewsky, 2020). Placed at the crossroads of mapping, georeferencing and literary studies, the concept of *factuality* is defined through toponymical data exclusively. By and large, (literary) toponyms function as proxies of real referents and as a shared set of nodes for complementary (factual) toponyms.
In the initial stages of our research, we visualized in QGIS (https://www.qgis.org) the spatial distribution of our dataset and the positional accuracy of approx. 6500 datapoints that have been organized in the following categories: *General Toponyms*, *Crime Toponyms (Crime Department(s))*, *Crime Toponyms (Foreign Travelers in Romania)*, *Dwelling Toponyms*, *Money-Hoarding Toponyms*, *Post Stations Toponyms, Inn/ Pubs Toponyms*, *Wood/ Forest Toponyms* (Dendronyms) (*see* Digital Humanities Laboratory, 2026).
Since the variety of collected toponyms and the current area of Romania are really large, in this paper we decided to explore only the posionality and the quantitative distribution of the geo-referenced 5225 dendronyms (forest names) that have been extracted from the Romanian thesauri. Dendronyms were overlapped with OSM data, that were extracted in their turn with Overpass Turbo (https://overpass-turbo.eu). After a careful inspection of our dendronym subset, we decided for narrowing the spatial focus to the province of Moldova, more precisely Iaşi, Vaslui, Neamş and Vrancea counties, currently situated in the Eastern part of Romania. The final count of dendronyms for the entire region of Moldova is 909 datapoints (*MD-Subset*), while for the 4 counties selected the count is 676 (henceforth abbreviated as *MDC-Subset*). Such strategy represents a step-by-step checking of positional accuracy and of empirical grounding limitations (see *Fig. 1*).

Through this approach, we will be able to answer the following research questions: 1. How are dendronym points positioned in comparison with the polygons vectorialized in OSM (outside area vs. inside area) and what are the retreat distances of these positionings? What is the density of inside dendronyms in relation to the current forest areas in four case counties? What are the centroids of the current forest areas and what are the distances between them and dendronym points? Can we hypothesize a causal relationship between deforestation (inside/ outside positionings of dendronyms) and the gradual disappearance of hajduk/ criminal life?

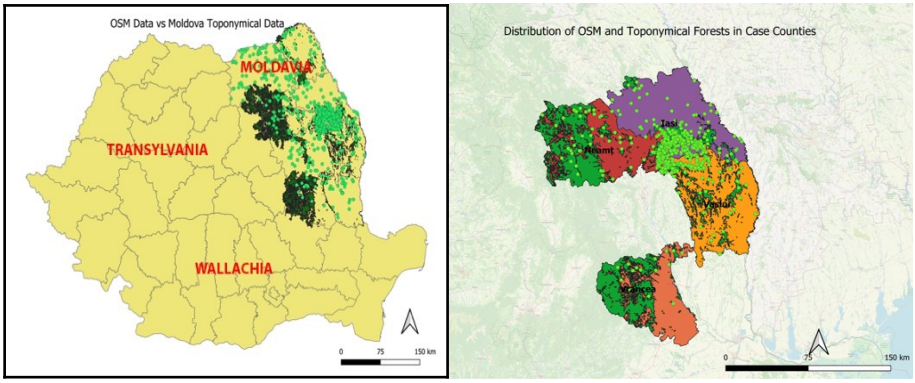


Figure 1. Distribution of MD and MDC data subsets vs. OSM data per 4 case counties

2. PROPOSAL

During the long nineteenth century, the hajduk fiction was one of the most productive subgenres of the Romanian novel. A wide and socially-diverse readership and a horde of industrious authors were common traits of the emerging literary market in both Principalities, Moldavia and Wallachia, that united, after the Treaty of Paris, in 1859 (Patraş, 2019, Patraş *et al*, 2020, Patraş, 2021, Pascariu *et al*, 2022, Olteanu, 2023). As a region with a Romanian majority, sharing the same language and cultural tradition (chiefly folklore and national myths) with the Romanian Principalities and neighboring the wider influence of the Balkan area, Transylvania was also part of the hajduk game; located between the Danube and the Black Sea, and assembled to the United Principalities of Romania in 1878, after the Russo-Turkish War, albeit the closest site to the Ottoman border, Dobruja was a land of the hajduk circulation too. In a nutshell, the actual geographical area of Romania circumscribes the hajduk life as a way of living or as a "form of life" (Agamben, 2000: 3-15). The idea of creating an interdisciplinary database and its rationale for a multitude of research endeavours draw from the discovery that, considered as a "form of life," hajduk criminality [in Romanian *haiducia*]—usually referring to grand route hijacking or to robbery in blind spots such as woods, pubs, inns, monasteries, country mansions (Gârleanu, 1969; Cvetkova, 1987)—requires a complex approach.

In order to check the nineteenth-century maps positionality and the preciseness of descriptive geographical references in the Romanian toponymy resources, conceptual streamlining is, for the moment, a must-do. However, oversimplification is not a research risk, since toponymical data (literary *y compris*) will be enriched so as to supply a nineteenth-century Balkan criminality database. By criminality we refer to crimes defined in juridical sources from the long nineteenth century, chiefly to varieties of theft, injury, abduction, rape and murder (hajduk criminality will be used as a preparatory case-study). By Balkan area we defined not only a strictly clipped geographical territory, but also regions that were influenced, via the Ottoman Empire's administration, by the Balkan culture, trade and societal types. Being a *longue durée* periodization, *long nineteenth* century—considered, in our previous projects, as extending from 1800[s] to the end of the First World War—was more convenient for the broadening scope of our research endeavor.

The present paper provides an analysis that will serve as a preparatory argument for enlarging the toponymical database on Balkan criminality, for enriching the attribute table of various layers, and for anchoring literary factuality into spatial referentiality. Forests have always been a (hajduk) criminality area (Zlatanov, 2018), which justifies a research interest in the way dendronyms correspond to the currently geo-referentiated areas. For several reasons, Romania's geographical territory provides a complex staging for this confrontation between linguistically-bound forest-data and spatially-bound forest-data: 1. Forests (as natural areas), on the one hand, and Roads and Cities (as humanly-built areas), on the other, represent the scaffolding of *stationing/ hiding vs. movement/ assault*, that is, of *haiducie* as a "form of life"; 2. A static perspective (self-implied in the act of *naming*) is tuned in a dynamic perspective (spatially *referential*, thus *factual*); 3. Natural areas have not been vectorialized completely, which, with the growing of contributions in OSM, lays seeds for long-term inspection on phenomena such as deforestation or, even if more unlikely, forestation.

Previous research has shown that in certain areas, particularly around Bucharest (Wallachia), Roman (Moldavia), or other historically significant regions, certain toponym types (forests, roads, crimes) tend to cluster and to form hot spots. This suggests a spatial correlation or co-occurrence, not necessarily a

direct causal relationship. Therefore, the current analysis is a zoomed-in, detailed mapping of previously hypothesized phenomena in the areas of 4 Moldavian case-counties. As in the case of *grand routes*, these new insights about *forest* positionality and distribution is meant to support the theoretical claim on discriminating between the sub-typologies of the national hajduks as popular fiction characters: the most evident Hajduk hubs in Wallachia are closer rather to forests than to roads (post stations); in stark contrast, hajduk hubs in Moldavia are closer to roads and to locations listed in the geographic dictionary as related to highway robbery. Such mapping enables us to demonstrate, literary-wise, that we need to differentiate between the two types of literary hajduk spatiality—one more factual (Moldavia) and the other less factual (Muntenia) (Atanasiu *et al*, 2025; Odochiciuc *et al*, 2025). Moreover, it opens new questions about the causes of deforestation, thus of landscape fragmentation. In certain areas of Romania, namely in the 4 Moldavian counties, a lot of place names have once referred to woods and forests but currently are not tagged among the OSM attributes as "forest landuse".

3. METHODS, TOOLS, AND METRICS

Our approach to data analysis is thus straightforward. Toponymical data (dendronyms), organized as csv files, are imported in QGIS (and stored as GEOjson files, see repo above) with EPSG4326: WGS84 and then integrated with OSM and PostgreSQL data. QGIS was used for producing maps by using various analyses performed with PGAdmin queries:

A. *Persistence analysis* checks the overlap between point and polygon geometries (https://doi.org/10.5281/zenodo.19694803). The calculus of the number of points in *MDC-Subset* which is situated inside the OSM polygons resulted in 290 ghost forests (dendronyms that overlap with areas outside current forest polygons) and 386 in-place forests (dendronyms that overlap with areas inside current forest polygons). Per each county, the figures are as follows: Neamş (60 ghost forests); Vaslui (46 ghost forests); Vrancea (5 ghost forests); Iaşi (*179 ghost forests*). The figures also show a proportion of 57 % (overlap) vs. 43 % (non-overlap), which means a massive deforestation in the focused counties, chiefly in Iaşi. Nearly a half of the places named as forests have now other land attributes.

B. *Nearest neighbor analysis* produces a list of retreat distances (in meters) between the current OSM forest polygons and *MDC-Subset* (https://doi.org/10.5281/zenodo.19694803). The list has been partitioned in three types of retreats: a. *critical retreat distance* (over 5 km); b. *major retreat distance* (between 1-5 km); c. *marginal retreat distance* (under 1 km). The distances resulted from the query calculation ran from 23,600 m to 1,5m, with the biggest values situated in the Iași county: 29 out of 31 critical forest retreats and 115 out of 149 major forest retreats. A limit of our analysis is the relative positionality of the *MDC-Subset*, which asks for more careful inspection of the 129 marginal retreats, and generally of the long tail of small figures, under 2000 m, which require a careful statistic processing. The standard deviation is quite big, therefore a typical value for forest distancing is 1,400 m, a score that will be taken as reference for further buffer analyses applied to roads and cities neighboring forest polygons.

C. *Dendronymic density* analyses provide insight into the density of dendronyms with respect to the areas of current forests and each case-county's administrative surface (https://doi.org/10.5281/zenodo.19694803). For instance, the Fishnet/Grid Spatial Analysis quantifies and maps the historical deforestation in the 4 case-counties by cross-referencing the two heterogeneous datasets. This study area is tessellated into a regular grid of 5×5 km cells (25 km² each), ensuring uniform areal units for inter-regional statistical comparison. For each cell, two core indicators are computed: a. *total dendronyms* = total number of historically recorded forest names); b. *current forest surface area* (Forest surface km² = derived from OSM polygons with "forest" attribute). High scores identify forest memory hotspots — areas where dense historical forest nomenclature persists in the absence of extant woodland, interpreted as evidence of significant anthropogenic landscape transformation (primarily agricultural conversion). The spatial distribution of scores reveals a marked east–west gradient: the Moldavian Plateau (Iaşi–Vaslui corridor, the road Vaslui-

Roman, via Negreşti) emerges as the principal zone of landscape discontinuity, characterized by dendronymic richness and near-complete forest loss, while the relationship between the Carpathian foothills (Neamş, Vrancea) and correspondent dendronymic record looks unbalanced, which indicates data biases and limits that should be researched further (see *Table 1*, *Fig. 2*).

County	Total Dendronyms	County surface (km2)	**Total Dendro Density**	In-Place Dendronyms	Forest surface (km2)	**In-place Dendro Density**
Iași	303	5477.41	**0.0553**	123	584	**0.2106**
Neamț	239	5897.18	**0.0405**	179	2758	**0.0649**
Vaslui	119	5316.89	**0.0224**	73	777	**0.0939**
Vrancea	15	4853.96	**0.0031**	10	2018	**0.0050**

Table 1. Dendronymic density scores

D. *Heart of forest* analysis is a spatial metric that employs the PostGIS function ST_PointOnSurface to compute an interior reference point, guaranteed to fall within the polygon boundary (see query https://doi.org/10.5281/zenodo.19694803). This central point represents a proxy for the geometric core of each extant forest massif, at the same time distinguishing between stable interior habitat (core) and more exposed peripheral zones (edge). By calculating the Euclidean distance between each georeferenced dendronym and the geometric heart of its corresponding OSM forest polygon, our analysis quantifies the degree of dendronymic centrality: dendronyms proximate to the core indicate historically persistent, spatially stable woodland nuclei, whereas dendronyms displaced toward or beyond the forest edge signal progressive fragmentation and spatial retreat of the forest boundary over time. From a landscape change perspective, it provides a measure of forest contraction: if the majority of historical forest names are now situated at the periphery of or outside extant woodland, this constitutes evidence that deforestation has proceeded centrifugally, eroding the forest from its margins inward. From a historical criminality perspective, the analysis offers a geographically grounded argument for understanding hajduk spatial behavior: dendronyms associated with outlaw activity (e.g., Pădurea Tâlhăreşti, Vaslui County) that cluster within the geometric core of forest massifs demonstrate a deliberate exploitation of maximum geographical isolation, as interior forest zones were least accessible to pursuing authorities (*potera*). The distance from the forest edge to the centroid thus functions as a quantifiable surrogate for the historical safety gradient of forested refugia, linking linguistic memory (toponymy), spatial analysis (GIS), and historical ecology within a single analytical framework (see *Fig. 3*).

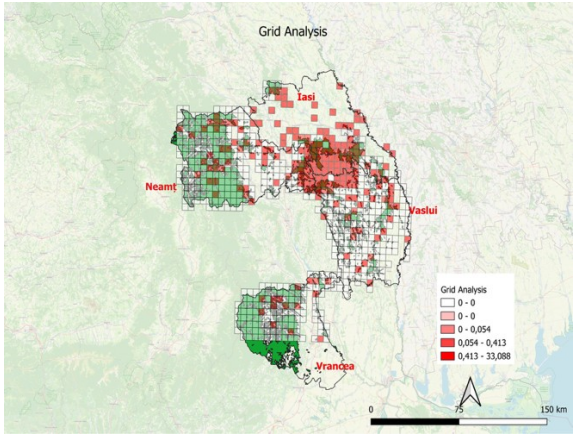


Figure 2: Fishnet analysis – in depth density values and positions per 4 case counties

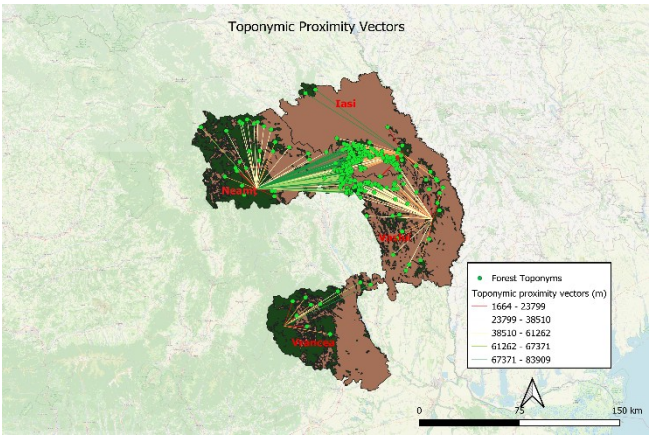


Figure 3: Heart of forests analysis - vectors between the heart of the forest per each case county and in-place datapoints

4. RESULTS AND DISCUSSION

In the Iași county, we found a record density and landscape fragmentation (0.519 dendronyms /km²). Although it has the smallest current forest area (584 km²), it was enlisted with the highest absolute number of dendronyms (303). Thus, Iaşi has an extremely rich "forest memory" relative to a small woodland area. It is the county with the greatest historical anthropic pressure and the highest degree of habitat fragmentation. Considering the densities calculated above as well as the OSM landuse attributes for this county, Vaslui has now a dominant agricultural landscape (0.153 dendronyms/km²). It preserves a somewhat larger forest area than Iaşi, but far fewer dendronyms have been enlisted in toponymical resources (119). The average density suggests a county where forests were massively converted into agricultural land (characteristic of the Bârlad Plateau). A comparison between the scores of Iași and Vaslui may suggest either weaker historical documentation in the thesaurus, or that the forests disappeared so long ago that even their names have been lost from collective memory. OSM landuse attributes support the finding that the Neamş county and ensuing dendronym densities are influenced by mountain massiveness (0.087 dendronyms/km²). In fact, this is the largest forest area (2,758 km²) overlapped by a relatively high number of dendronyms (239). Despite having so many forest names, the calculated density is low. This is typical of mountain areas with compact and continuous forest massifs. Within a large massif, a single dendronym (e.g. "The Tarcău Mountains") can cover an enormous area. Neamț represents the stability and continuity of the forest heritage. The Vrancea county clearly indicates a thesauri data gap (0.007 dendronyms/km²), because a vast forest area (2,018 km²) has only 15 recorded dendronyms. However, the current state of data leads to the interpretation that the forests of Vrancea are so vast and uninterrupted that they were never "fragmented" toponymically: in the conditions of one single immense woodland, there was no need for different dendronyms for each landform or administrative partition.

While data extraction, customized metrics (densities, centroids) and test mapping will be fine-tuned with supplementary forest data that OSM and other institutions may provide in the future, our analysis taps in previous findings on the hajduk's literary space (Galleron *et al*, 2021), namely cities/ zones and grand routes that are also marked in fiction as indicatives of hajduk life.

REFERENCES

Agamben, G. (2000). Means without end: Notes on politics (V. Binetti & C. Casarino, Trans.). University of Minnesota Press.

Atanasiu, M.-B., Răchită, C., Odochiciuc, A., & Patraş, R. (2025). Crossing the grand routes: Highwaymen on the roads of the Romanian principalities in the long nineteenth century. Linguaculture, 16(1), 140–173. https://doi.org/10.47743/lincu-2025-1-0410.

Cvetkova, B. (1987). Les sources ottomanes concernant le mouvement des haïdouks au XVe et XVIIIe siècles. In G. Kara (Ed.), Between the Danube and the Caucasus, 7–23.

Digital Humanities Laboratory. (2026). Database on hajduk criminality [Data set]. Zenodo. https://doi.org/10.5281/zenodo.18233708.

Galleron, I., Patraş, R., Grădinaru, C., & Mélanie-Becquet, F. (2021). À la recherche des haïdouks: Annoter les entités spatiales dans des romans roumains du XIXe siècle. Humanités numériques, 3. https://doi.org/10.4000/revuehn.1399.

Gârleanu, S. I. (1969). Haiducie şi haiduci. Bucureşti.

Lavocat, F. (2020). Pseudofactual narratives and signposts of factuality. In M. Fludernik & M.-L. Ryan (Eds.), Narrative factuality: A handbook. De Gruyter, 577–592. https://doi.org/10.1515/9783110486278.

Olteanu, A. (2023). The ideological instrumentalization of literary myths in 19th century Romanian literature. Meridian Critic, 43(2), 123–136.

Odochiciuc, A., Atanasiu, M.-B., Răchită, C., & Patraş, R. (2025). Georeferencing criminality hubs in late 18th and 19th-century Romanian principalities and Transylvania [Data set]. Zenodo. https://doi.org/10.5281/zenodo.15909806.

Pascariu, L., Olteanu, A., Tican, M., Ionescu, O., & Babii, R. (2022). Romanul haiducesc al lui N. D. Popescu. Meridian Critic, 39(1), 121–136.

Patraş, R. (2019). Hayduk novels in the nineteenth-century Romanian fiction: Notes on a sub-genre. Swedish Journal of Romanian Studies, 2(1), 24–33. https://doi.org/10.35824/sjrs.v2i1.18769.

Patraş, R. (2021). Is this a man's world? Hajduk heroines, from eighteenth-century ballads to nineteenth-century Romanian fiction. In C. Duşu (Ed.), Eurofringes: Translating texts, translating cultures.

Patraş, R., Galleron, I., Grădinaru, C., & Mélanie-Becquet, F. (2020). HAIRO: Corpus de romane haiduceşti (1850–1950) [Data set]. Nakala. https://www.nakala.fr/collection/10.34847/nkl.eaf5321x.

Patraş, R., Răchită, C., Atanasiu, M.-B., & Odochiciuc, A. (2026). Structuring Resources for a Database of Balkan Criminality in the Long Nineteenth Century. Journal of Open Humanities Data, 12: 49, pp. 1–11. DOI: https://doi.org/10.5334/johd.515.

Piatti, B., Bär, H. R., Reuschel, A.-K., & Cartwright, W. (2009). Mapping literature: Towards a geography of fiction. In W. Cartwright, G. Gartner, & A. Lehn (Eds.), Cartography and art (Lecture Notes in Geoinformation and Cartography, 177–192). Springer. https://doi.org/10.1007/978-3-540-68569-2_15.

Rajewsky, I. (2020). Theories of fictionality and their real other. In M. Fludernik & M.-L. Ryan (Eds.), Narrative factuality: A handbook. De Gruyter, 29–50. https://doi.org/10.1515/9783110486278-002.

Zlatanov, A. (2018). Forest and outlaws in the Bulgarian lands during the eighteenth century. In F. Mérot, T. Rentet, & C. Buchet (Eds.), Le bois dans tous ses états: De la représentation aux réalités quotidiennes de la forêt européenne (Xe–XIXe siècles), 105–115.

OpenStreetMap. (n.d.). OpenStreetMap. https://www.openstreetmap.org.

Overpass Turbo. (n.d.). Overpass Turbo. https://overpass-turbo.eu.

QGIS.org. (n.d.). QGIS geographic information system. https://www.qgis.org.

PostgreSQL. (n.d.). https://www.postgresql.org/.

Digital Humanities per narrare i territori: un WebGIS sui porti minori della Sardegna dal medioevo alla contemporaneità. Una proposta di implementazione collaborativa[1]

LUIGI SERRA[1], SEBASTIANA NOCCO[2]

[1] CONSIGLIO NAZIONALE DELLE RICERCHE – ISTITUTO DI STORIA DELL'EUROPA MEDITERRANEA, LUIGI.SERRA@CNR.IT

[2] CONSIGLIO NAZIONALE DELLE RICERCHE – ISTITUTO DI STORIA DELL'EUROPA MEDITERRANEA, SEBASTIANA.NOCCO@CNR.IT

ABSTRACT (ITALIANO)

Il contributo analizza metodi e strumenti per ricostruire e rappresentare con cartografia digitale la rete dei porti minori della Sardegna dal Medioevo all'età contemporanea. Considerati in relazione al paesaggio costiero, alle dinamiche insediative e alle reti di mobilità marittima e terrestre, i porti minori costituiscono un osservatorio privilegiato per leggere le trasformazioni delle coste sarde in una prospettiva di lungo periodo. La ricerca integra fonti testuali, iconografiche e materiali, indispensabili per una mappatura diacronica della rete e per la ricostruzione geo-storica dei siti e dei legami tra porti, insediamenti e aree limitrofe. I dati su porti storici e attuali sono stati schedati e organizzati in banche dati pseudo relazionali, quindi georeferenziati e restituiti in ambiente GIS.
L'uso del GIS consente di descrivere la rete portuale non come un elenco statico di punti, ma come una configurazione dinamica, sensibile a cambiamenti politici, economici e ambientali. Esempi applicativi mostrano come l'Historical GIS integri scale diverse (locale, regionale, mediterranea) e renda visibili relazioni, gerarchie e percorsi difficilmente ricostruibili con approcci tradizionali.
In questa prospettiva, il GIS funge sia da infrastruttura aggiornabile di raccolta dati, sia da strumento di valorizzazione dei porti minori, supportando narrazioni identitarie e proposte per la promozione del patrimonio culturale e paesaggistico, la sostenibilità e la tutela delle aree costiere, anche in chiave partecipativa e turistica. In quest'ottica il presente contributo propone una possibile declinazione dinamica del GIS che potrebbe trarre giovamento da un popolamento partecipato ovvero nella coniugazione tra citizen science ed heritage sharing per un reale public engagement.
Parole chiave: Porti minori della Sardegna; Sistemi Informativi Geografici, Collaborazione Digitale; Fonti geo-storiche; Turismo culturale

ABSTRACT (ENGLISH)

Digital Humanities to narrate the territories: a WebGIS on the minor ports of Sardinia from the Middle Ages to the present day; results and a proposal for collaborative implementation
The paper analyzes methods and tools to reconstruct and represent with digital cartography the network of minor ports of Sardinia from the Middle Ages to the contemporary age. Considered in relation to the coastal landscape, settlement dynamics and maritime and land mobility networks, the minor ports offer a privileged observatory for slight transformations of the Sardinian coasts in a long-term perspective. The research integrates textual, iconographic and material sources, indispensable for a diachronic mapping of the network and for the geo-historical reconstruction of the sites and the links between ports, settlements and neighboring areas. The data on historical and current ports were filed and organized in pseudo-relational databases, then georeferenced and returned in a GIS environment. The use of GIS allows the port network to be described not as a static list of points, but as a dynamic configuration, sensitive to political, economic and environmental changes. Application examples show how Historical GIS integrates different scales (local, regional, Mediterranean) and makes visible relationships, hierarchies and paths that are difficult to reconstruct with traditional approaches. In this perspective, the GIS acts both as an updatable data collection infrastructure and as a tool for the enhancement of minor

[1] Sebbene il contributo sia frutto di ricerche e riflessioni comuni dei due autori, sono da attribuire a Luigi Serra i paragrafi 3, 4, 5, 6, a Sebastiana Nocco i paragrafi 1 e 2.

ports, supporting identity narratives and proposals for the promotion of cultural and landscape heritage, sustainability and protection of coastal areas, also in a participatory and tourist key. In this perspective, this contribution proposes a possible dynamic declination of GIS that could benefit from a participatory population, i.e. in the combination of citizen science and heritage sharing for real public engagement.
Keywords: Minor ports of Sardinia; GIS; Digital Collaboration; Geo-Historical Sources; Cultural Tourism

1. INTRODUZIONE

Il presente contributo intende presentare alcuni risultati e le possibili implementazioni di una ricerca condotta nell'ambito del progetto PRIN 2022 PNRR *Recovering and representing the identity of minor ports in Southern Italy (peninsular and islands) between the Middle Ages and the Modern Age for the inclusion and sustainable development of coastal areas*. Al progetto, coordinato da F.P. Tocco dell'Università degli Studi di Messina, hanno collaborato le unità di ricerca dell'Università della Calabria, dell'Università del Salento e dell'ISEM-CNR di Cagliari. L'area indagata è, dunque, quella delle due isole maggiori, Sicilia e Sardegna, oltre a Calabria e Puglia (Tocco, 2026). All'interno delle varie unità di ricerca hanno collaborato principalmente storici medievisti, storici modernisti e geografi, con l'aggiunta in alcuni casi anche di storici dell'arte, archeologi, informatici ed esperti di *digital humanities*.
Il progetto si è articolato in varie fasi che hanno previsto la catalogazione e mappatura della rete dei porti minori delle regioni coinvolte con l'obiettivo di individuare i siti di interesse; la ricostruzione storica e geo-storica di questi siti, anche nei loro rapporti con l'entroterra e con l'esterno; l'approfondimento di alcuni casi studio analizzati nel lungo periodo, fino alla contemporaneità. Infine, i dati raccolti nel corso della ricerca sono stati analizzati per l'elaborazione di proposte e azioni finalizzate alla promozione del patrimonio culturale, alla sostenibilità e alla tutela delle aree costiere e dei siti portuali attraverso lo sviluppo di contenuti di supporto e strumenti di lavoro derivati dalla ricerca storica e geografica.
Quest'ultima fase, tuttora in corso, vede concretizzarsi tali proposte anche in operazioni integrate e collaborative di sviluppo locale per il benessere umano e lo sviluppo sostenibile, con una sinergia tra stakeholder pubblici e privati, unitamente ad attività di circolazione e comunicazione delle informazioni, attraverso la realizzazione, tra l'altro, di un WebGIS.
All'interno di questa cornice si è inserito il contributo dell'unità di ricerca dell'ISEM-CNR, all'interno della quale hanno collaborato cinque ricercatori della sede di Cagliari dell'Istituto (L. Gallinari, G. Serreli, M.G.R. Mele, L. Panetta, S. Nocco, quest'ultima anche nel ruolo di coordinatrice dell'UR) con competenze diverse che spaziano dalla storia medioevale e moderna, all'archeologia, alla geografia storica. L'unità di ricerca si è concentrata sullo studio della rete dei porti minori della Sardegna tra Medioevo ed Età Moderna, integrando fonti documentarie, cartografiche, archeologiche e strumenti digitali per la ricostruzione storica e la rappresentazione georeferenziata dei siti portuali. Lo staff di ricerca è stato affiancato da un ingegnere informatico (L. Serra) che ha fornito supporto per l'armonizzazione dei dati raccolti dai collaboratori del progetto, per la normalizzazione dei dati relativi ai porti, per la costituzione del database delle informazioni collezionate e loro integrazione, per la georeferenziazione digitale dei porti storici della Sardegna e di quelli ancora in attività.
Obiettivo del lavoro presentato in questa sede - frutto della collaborazione sinergica tra competenze, metodi e approcci differenti (geografico-storico e informatico) - è proporre alcuni possibili sviluppi futuri della ricerca condotta dai due autori nell'ambito del progetto PRIN. In particolare, si intendere sfruttare le potenzialità del GIS e soprattutto del web GIS per veicolare l'esito delle ricerche storiche e geo-storiche e renderle accessibili a un più ampio pubblico, contribuendo così a trasformare la conoscenza del passato in una narrazione identitaria, capace di rafforzare il legame delle comunità con il proprio territorio e a promuovere processi partecipativi di coinvolgimento delle comunità locali, chiamate a riconoscersi nei propri paesaggi storici e ad essere partecipi della loro narrazione. Allo stesso tempo si intende attrarre forme di turismo attente alla storia e all'ambiente e politiche di turismo sostenibile che superino la stagionalità e la monocultura balneare.

2. I PORTI MINORI DELLA SARDEGNA TRA MEDIOEVO E CONTEMPORANEITÀ: DALLA RICERCA GEO-STORICA ALLA VALORIZZAZIONE DEI TERRITORI

Il caso studio qui analizzato è quello dei porti minori della Sardegna, talvolta trascurati dalla storiografia, che ha privilegiato i grandi scali urbani, in particolare Cagliari.
Eppure, i porti minori hanno svolto un ruolo fondamentale nella strutturazione delle connessioni locali e sovraregionali dell'isola, fungendo da nodi importanti per la circolazione di uomini, merci e saperi nel

corso dei secoli. Questi scali, spesso semplici approdi naturali o legati a insediamenti costieri, hanno contribuito alla costruzione dell'identità insulare, fungendo da snodi vitali per la pesca, il commercio locale, la mobilità interna e la difesa territoriale (Simbula, 2014).
Da qui l'importanza di un'indagine geo-storica che abbracci il periodo compreso tra il Basso Medioevo e l'età contemporanea, con l'obiettivo di ricostruire il ruolo di questi scali nella formazione dell'identità insulare e nella connessione della Sardegna con il Mediterraneo (Quaini, 1974; Nocco, 2026).
Fin dall'antichità, i porti sardi hanno svolto funzioni fondamentali nel connettere l'isola alle grandi civiltà mediterranee: dai Fenici ai Cartaginesi, dai Romani ai Bizantini, agli Arabi, fino ai Genovesi, Pisani, Aragonesi e Spagnoli. Questi contatti hanno lasciato tracce visibili sul paesaggio urbano e culturale delle aree costiere.
L'evoluzione storica di queste realtà è strettamente legata a processi economici, ma anche politici, militari e urbanistici. Porti e approdi erano collegati non solo alle reti marittime, ma anche all'entroterra isolano, creando sistemi economici integrati che mettevano in relazione risorse interne – agro-pastorali (penso a grano, vino, formaggio, legname), minerarie, artigianali – con le esigenze del commercio e della navigazione mediterranea. Assai importante era altresì il commercio di preziose risorse naturali quali il sale e il corallo, minerali quali l'ossidiana o l'argento, ma anche la loro funzione in chiave difensiva (Serreli, 2024 e 2025; Nocco, 2025; Mele, 2026; Panetta, 2026; Serreli, Cocco, 2026).
Fin dal primo momento è emersa l'importanza del lavoro di individuazione e mappatura di porti e approdi minori attestati in epoca medioevale e moderna, al quale è stata dedicata la prima fase del progetto. Inizialmente è stata condotta una ricerca bibliografica sui testi generali e su quelli riguardanti argomenti specifici – quali le istituzioni, l'economia, le politiche difensive, la navigazione, il commercio, la cartografia storica, le indagini archeologiche – che fornissero elementi utili a identificare i piccoli porti e a definirne le caratteristiche.
Fondamentale per la mappatura diacronica della rete dei porti minori, per la ricostruzione geo-storica dei siti e delle relazioni tra porti, insediamenti e aree limitrofe, che ha caratterizzato la seconda fase del progetto, è stata l'analisi integrata di fonti di varia natura – testuali, iconografiche e materiali.
L'approccio è quello geostorico, che – adottando la necessaria trans-scalarità - si propone di integrare dimensioni spaziali e temporali nella ricostruzione storica (Guarducci, Rombai 2017).
La ricerca si è basata dunque sull'integrazione di una pluralità di fonti, che dialogano tra loro e permettono una ricostruzione multilivello della realtà portuale: si tratta di documentazione politico-amministrativa e istituzionale, fonti economico-fiscali, fonti testuali e cartografiche legate ad esigenze difensive, queste ultime in particolare per l'età moderna (Nocco, 2025; Serreli, 2024 e 2025). Ad essi si aggiungono inoltre fonti letterarie e testi narrativi di vario genere, cronache, diari e resoconti di viaggio; fonti iconografiche e cartografiche, quali carte nautiche e topografiche, vedute urbane, immagini pittoriche.
Infine, ma non certo ultimo per importanza, l'apporto dei portolani che, con i loro elenchi di porti, distanze tra scali, condizioni di approdo, orientamenti utili alla navigazione, si rivelano fonti imprescindibili per ricostruire le reti marittime medievali, analizzare l'importanza strategica dei porti nel sistema mediterraneo; studiare la toponomastica storica costiera, come sottolineato da Massimo Quaini fin dal 1969 (Quaini, 1974; Nocco, 2026).
Come previsto dal progetto, le informazioni sui porti storici e su quelli attuali sono state schedate e raccolte in banche dati pseudo relazionali, georeferenziate e rappresentate attraverso cartografia digitale. Le schede descrittive che si aprono su ogni sito forniscono le informazioni utili alla sua localizzazione e conoscenza geostorica, con diversi gradi di dettaglio in base alla ricchezza di notizie reperite, fornendo un utile strumento di lavoro (almeno speriamo) anche in futuro.
L'integrazione di fonti scritte, iconografiche e materiali è altresì fondamentale per restituire una lettura stratificata del porto, non come luogo isolato, ma come spazio vissuto, trasformato, narrato, illustrato e rappresentato.
Questa molteplicità di fonti ha permesso di:

- Ricostruire la morfologia e l'evoluzione dei porti e delle aree urbane circostanti;
- Comprendere le dinamiche storiche che hanno determinato l'evoluzione delle aree portuali in una prospettiva di lunga durata;
- Analizzare le connessioni fra porto ed entroterra, sia a livello materiale (merci, infrastrutture, mobilità), sia immateriale (idee, culture, lingue);
- Offrire strumenti per una valorizzazione consapevole e sostenibile di questi spazi.

Lo studio geostorico dei porti sardi non si limita, infatti, alla ricostruzione del passato: ambisce a fornire strumenti utili per il governo del territorio e la sua valorizzazione culturale e turistica in chiave sostenibile e partecipativa (Grava, Berti, Gabellieri & Gallia, 2020; Masetti, Spadafora, 2024).
In quest'ottica, il GIS sulla rete portuale (che sarà oggetto di ulteriori approfondimenti nei paragrafi successivi) assolve alla funzione di raccolta di dati georeferenziati con potenzialità di aggiornamento e implementazione, e anche di strumento di valorizzazione dei porti minori.
Una volta terminato, il GIS potrà essere oggetto di presentazioni pubbliche e divulgato alle istituzioni, alle amministrazioni e alle comunità locali, così come previsto e auspicato dal PNRR che ha finanziato il progetto.

3. IL WEB GIS SUI PORTI MINORI DELLA SARDEGNA

Lo sviluppo di un Web GIS statico, ovvero una rappresentazione cartografica digitale dei dati raccolti sugli approdi e i porti minori della Sardegna dal medioevo ad oggi, è stato previsto come risultato finale di progetto. La sua realizzazione si è basata sulle attestazioni documentali, con i riferimenti alle fonti e ai relativi contenuti, successivamente acquisite su schede descrittive sintetiche e rese fruibili mediante consultazione web. Il sistema informativo geografico realizzato ha previsto il posizionamento dei punti individuati su un sistema cartografico georeferenziato e una rappresentazione su mappa interattiva dei porti e degli approdi del medioevo, dell'età moderna, le loro eventuali sovrapposizioni e riutilizzo stratificato, e gli attuali porti e marine in attività ancora ai giorni nostri.
Il prodotto, ancora in fase di perfezionamento, è stato installato su un server di prova in rete locale, fruibile pubblicamente in modo selettivo, ma non ancora ottimizzato esteticamente e dal punto di vista della sicurezza informatica (DNS, SSL Certificate, HTTPS Protocol).
Le soluzioni GIS sono molto diffuse perché forniscono informazioni stratificate e inerenti diversi aspetti legati al territorio: analisi statistiche, demografiche, economiche e nel nostro caso, geografiche e storiche, salvo altre. Amministrazioni pubbliche e soggetti privati si avvalgono di questa tecnologia perché la rappresentazione su un sistema informativo geografico consente di avere una visione panoramica tematica del territorio in esame (Grava, M., Berti, C., Gabellieri, N., & Gallia, A., 2020). La scelta di implementare una piattaforma web GIS cerca di dare una risposta integrata a quelle esistenti su differenti piattaforme cartacee e web: la frammentazione delle informazioni ha suggerito la realizzazione di una raccolta organica delle informazioni sui porti minori della Sardegna affinché queste fossero reperibili su un'unica piattaforma web-based di semplice ed immediata consultazione, sulla base di tematismi predefiniti e selezionabili a piacere. Tale approccio, sebbene con declinazioni e tassonomie differenti, è stato utilizzato nella realizzazione di piattaforme analoghe di cui riportiamo alcuni esempi accademici: Atlante digitale di Storia Marittima del Regno di Sardegna del Centro Interdipartimentale per l'Umanistica Digitale DH dell'Università degli Studi di Cagliari[2]; "envlandscapes" Envisioning landscapes: geohistorical travel sources and GIS-based approaches for participative territorial management and enhancement PRIN 2022 PNRR dell'Università di Genova, Università di Roma Tre, Università di Trento[3]; "LOC-GLOB" The local connectivity in an age of global intensification PRIN 2017 delle Università di Bergamo, Università di Sassari, Università di Torino, Università di Udine[4]. La scelta di una rappresentazione cartografica digitale fornisce uno strumento di consultazione interattiva per la fruibilità delle informazioni presenti sulla piattaforma. L'approccio visuale mediante schede contestuali consente la lettura delle informazioni essenziali direttamente sui siti di interesse, sia dal punto di vista storico geografico che dal punto di vista turistico. Il sistema multilivello (multilayer) è pensato per la selezione della carta tematica di interesse e dei relativi sottolivelli in modo cumulativo, alternato o selettivo per la visualizzazione delle informazioni desiderate.
A titolo d'esempio è possibile visualizzare: i porti medievali, i porti di età moderna e loro sovrapposizioni laddove presenti; i porti e le marine attuali con i riferimenti di contatto e le principali informazioni; Presidi territoriali della Guardia Costiera e recapiti coerenti con le informazioni del sito istituzionale del Corpo delle Capitanerie di Porto e Guardia Costiera; altri tematismi come le Aree SNAI della Sardegna aggiornate all'anno 2025; La dislocazione delle Bandiere Blu in Sardegna per l'anno 2025; La rappresentazione dei porti in relazione ai parchi naturali regionali nazionali e le Aree Marine Protette.

[2] https://storia.dh.unica.it/asmsa/it/esplora-latlante/asmsa-carta-di-sintesi
[3] https://envlandscapes.prin.unige.it/WebGIS
[4] https://loc-glob.unibg.it/

Tutte queste informazioni possono essere inserite con livelli aggiuntivi relativamente al loro aggiornamento nel tempo, per un eventuale confronto temporizzato delle stesse categorie oltre ad eventuali altri tematismi.

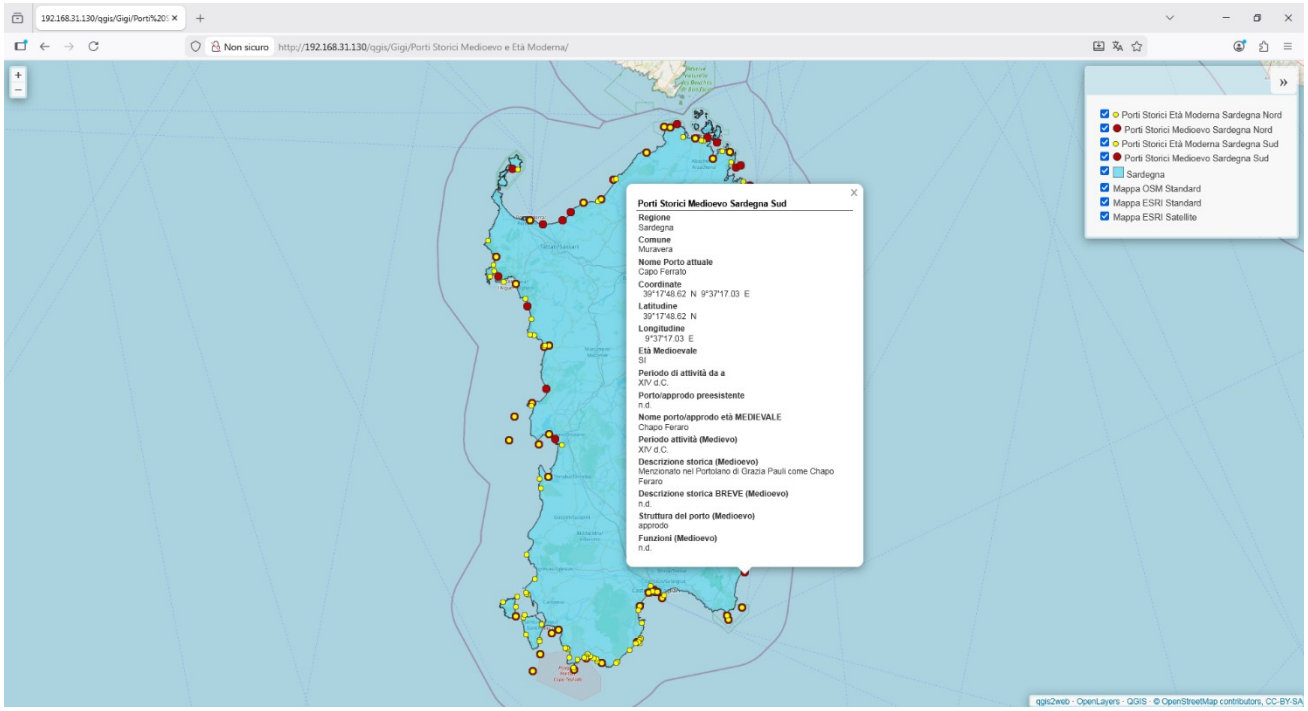


Figura 1. Historical Web GIS sui porti minori della Sardegna di Medioevo ed Età Moderna

4. STRUMENTI E METODI

La metodologia adottata si è basata su paradigmi Open-Source[5] Web-Based: la quasi totalità dei software impiegati nella realizzazione del prodotto è a sorgente aperto o accessibile gratuitamente, per una agevole ridistribuzione dei contenuti e riproducibilità scalabile della soluzione realizzata. Per una scelta di progetto si sono utilizzati file di dati, impropriamente definiti database, costituiti da fogli di calcolo Microsoft Excel come "template", poi condivisi tra tutte le UR, per il collezionamento di informazioni mirate allo studio geo-storico dalle fonti medievali e di età moderna, e per lo studio turistico per i porti e le marine attuali in relazione al territorio, alla viabilità e all'offerta culturale. I siti individuati sono stati dapprima posizionati su mappa utilizzando Google Earth[6] per una successiva importazione su QGIS[7] e creazione dei geopackage relativi a tutti i siti. La facilità di manipolazione dei dati strutturati in forma tabellare, ha agevolato il popolamento del foglio dati da parte di tutto il gruppo di lavoro costituito prevalentemente da umanisti. Gli informatici coinvolti hanno trattato i dati estrapolati dalle fonti e memorizzati nel file Excel mediante opportuna normalizzazione ed esportazione in formato CSV, sono poi stati correttamente importati sul software QGIS per la creazione delle carte tematiche. Gli shape territoriali sono stati prelevati da geoportali istituzionali della regione Sardegna[8], dal sito ISTAT[9] e da altri repository web condivisi in formato WMS e WMTS. L'utilizzo di vari plugin di QGIS ha consentito la manipolazione e il rendering delle informazioni GIS direttamente in formato web. Il plugin base utilizzato per questa esportazione è stato qgis2web[10], i file sono stati poi modificati per una resa grafica ed estetica personalizzata. È stato installato un server web basato su Apache 2[11] ospitato su una piattaforma Linux Ubuntu 22.04[12] per la fruizione dei contenuti via browser web. La piattaforma è tutt'ora in versione *beta* e

[5] https://opensource.org/osd (cons. 22/04/2026)
[6] https://www.google.it/intl/it/earth/index.html (cons. 22/04/2026)
[7] https://qgis.org/ (cons. 22/04/2026)
[8] https://www.sardegnageoportale.it/ (cons. 22/04/2026)
[9] https://www.istat.it/statistiche-per-temi/focus/informazioni-territoriali-e-cartografiche/ (cons. 22/04/2026)
[10] https://github.com/qgis2web/qgis2web (cons. 22/04/2026)
[11] https://www.apache.org/licenses/LICENSE-2.0 (cons. 22/04/2026)
[12] https://releases.ubuntu.com/jammy/ (cons. 22/04/2026)

ospitata su un server on premise senza risoluzione DNS[13] né certificato SSL per connessioni sicure https[14]. È pertanto raggiungibile selettivamente in http non crittografato da utenti ed indirizzi IP specifici. In accordo con le altre UR che hanno realizzato il proprio web GIS regionale, verrà pubblicato a breve su server pubblico e messo in sicurezza informatica per una fruibilità estesa e senza limitazioni.

5. PROPOSTA DI IMPLEMENTAZIONE COLLABORATIVA

Il sito Web GIS realizzato, anche se fruibile interattivamente, è di per sé statico: le informazioni inserite sono estrapolate da file statici popolati in precedenza da più fonti e la generazione delle viste è legata ai dati inseriti in queste tabelle predefinite di cui abbiamo parlato in precedenza. Questo si traduce in una "staticità" relativa ai file utilizzati per generare le mappe stesse. Tale approccio è collaudato e a prova di inconsistenza delle informazioni inserite perché attentamente vagliate; tuttavia, una possibile declinazione dinamica di questo sistema potrebbe trarre giovamento da un popolamento partecipato ovvero nella coniugazione tra *citizen science* ed *heritage sharing* per un reale *public engagement.* Si propone a tal fine l'implementazione di un portale "mediato" per l'inserimento di dati sul Web GIS costituito da informazioni ed integrazioni documentali che partano dal basso come citazioni a fonti non indagate, fotografie, carte inedite, et alia. Una analoga proposta di piattaforma collaborativa è stata presentata con altra declinazione (Serra, 2019), ma sempre in ambito storico turistico. Un sistema di autenticazione "on-demand" o federato consentirebbe l'accesso autenticato al portale, concesso dagli amministratori su richiesta dell'utente, per il successivo inserimento di informazioni preformattate su apposite maschere, funzionali al popolamento del database dedicato. Le informazioni memorizzate sulle tabelle del db, una volta estrapolate, consentirebbero la successiva generazione delle schede descrittive dinamicamente, seppure preimpostate dagli accademici, dai ricercatori o dagli operatori del settore. Le diverse contestualizzazioni, storica, turistica o altre, verrebbero arricchite così dagli studiosi interessati ad ampliare i dati scientifici o dagli operatori economici che volessero essere presenti sulla piattaforma con l'inserimento autonomo delle informazioni che li riguardano a garanzia di correttezza e costante aggiornamento. Tutte le informazioni, poiché introdotte con accesso autenticato e vagliate dagli specialisti che amministrano e supervisionano la piattaforma, riporterebbero un costante riferimento al compilatore a garanzia della scientificità dei dati e della veridicità delle informazioni aggiunte.

6. CONCLUSIONI

I primi risultati ottenuti forniscono un panorama piuttosto variegato e raccontano una vivacità inattesa nella fruizione capillare delle coste sarde a partire dal medioevo, come si evince dalla continuità e numerosità dei punti GIS rappresentati sul perimetro costiero.
La piattaforma realizzata è scalabile e ulteriormente implementabile, le informazioni storiche collezionate potranno essere integrate man mano che ulteriori fonti dovessero far emergere dati e informazioni sulla realtà portuale e marittima della Sardegna. Il solido fondamento storico propone allo studioso una base da cui partire per ulteriori approfondimenti sul tema e si candida come guida di consultazione per i porti medievali e di età moderna. Le informazioni acquisite sui porti e le marine attuali forniscono un ulteriore strumento a supporto del diportismo e della nautica odierna, offrendo informazioni essenziali per i naviganti e i diportisti. L'attenzione rivolta alle realtà portuali esistenti in stretta relazione con il territorio nel quale insistono, ma legati alla viabilità e alla rete dei trasporti terrestri e aerei, consente di integrare queste realtà costiere in funzione della fruibilità del territorio e dell'offerta culturale interna. L'analisi geografica in sinergia con le proiezioni economiche e sociali che tali infrastrutture forniscono, offre strumenti di valutazione dei diversi contesti per adottare scelte strategiche di valorizzazione del territorio e del patrimonio culturale tangibile ed intangibile, per una fruizione turistica consapevole e una destagionalizzazione del turismo che oltre alla componente balneare, si arricchisce con la componente culturale.

RINGRAZIAMENTI

Ricerca finanziata dal progetto PRIN 2022 PNRR *Recovering and representing the identity of minor ports in Southern Italy (peninsular and islands) between the Middle Ages and the Modern Age for the inclusion and sustainable development of coastal areas* (30 novembre 2023 - 31 gennaio 2026),

[13] https://www.ietf.org/rfc/rfc1035.txt (cons. 22/04/2026)
[14] https://datatracker.ietf.org/doc/html/rfc2818 (cons. 22/04/2026)

BIBLIOGRAFIA

Grava, M., Berti, C., Gabellieri, N., & Gallia, A. (2020). Historical GIS. Strumenti digitali per la geografia storica in Italia. Trieste, Italy: Edizioni Università di Trieste.

Guarducci, A., Rombai, L. (2017). Paesaggio e territorio, il possibile contributo della geografia. Concetti e metodi. *Scienze Del Territorio*, 5, 19–25, https://doi.org/10.13128/Scienze_Territorio-22226

Masetti C., Spadafora G. (2024) (a cura di), *Digital Humanities, patrimonio culturale e applicazioni geostoriche*, Roma, Laboratorio geocartografico "Giuseppe Caraci" ("Dalla mappa al GIS", n. 7).

Mele, M.G.R. (2026). *Il porto di Oristano in età medioevale*, in Tocco, F.P. (2026) (a cura di). *Porti minori di Mezzogiorno, Sicilia e Sardegna dal Medioevo a oggi. Fisionomie storiche e prospettive di riqualificazione*, Palermo, Plumelia edizioni: 247-262.

Nocco, S. (2025). *Progettare la difesa di porti e approdi del Mediterraneo occidentale in età moderna: cartografie e documenti d'archivio*, in *Defensive Architecture of the Mediterranean*, vol. XVIII, Proceedings of the *International Conference on Fortifications of the Mediterranean Coast. FortMed* 2025 (Caserta, 10-11-12 april 2025), O. Zerlenga, V. Cirillo (Eds.), © DADI_PRESS (Department of Architecture and Industrial Design, University of Campania Luigi Vanvitelli) ISBN 978-88-85556-35-5 / edUPV (Universitat Politècnica de València) ISBN 978-84-1396-322-8, DOI volume: https://doi.org/10.4995/Fortmed2025.2025.20440: 229-236, DOI contributo: https://doi.org/10.-4995/Fortmed2025.2025.20250.

Nocco, S. (2026). *Porti e approdi della Sardegna nei portolani*, in Tocco, F.P. (2026) (a cura di). *Porti minori di Mezzogiorno, Sicilia e Sardegna dal Medioevo a oggi. Fisionomie storiche e prospettive di riqualificazione*, Palermo, Plumelia edizioni: 231-245.

Panetta, A. (2026). *Il porto di Bosa dal Medioevo all'Età Contemporanea. Le molte vite e le molte morti di un sistema portuale fluviale*, in *Tocco, F.P. (2026) (a cura di). Porti minori di Mezzogiorno, Sicilia e Sardegna dal Medioevo a oggi. Fisionomie storiche e prospettive di riqualificazione, Palermo, Plumelia edizioni:* 263-282.

Quaini, M. (1974). *Catalogna e Liguria nella cartografia nautica e nei portolani medievali*, in *Atti del I Congresso storico (Liguria-Catalogna-Ventimiglia-Bordighera-Albenga-Finale-Genova, 14-19 ottobre 1969)*, Bordighera: 550-571, poi ripubblicato sotto il titolo di *Immagini del Mediterraneo nella cartografia e nei portolani medievali*, in M. Quaini, *Tra geografia e storia. Un itinerario della geografia umana*, Bari 1992, 41-62.

Serra, L. (2019). "Open Monuments Engine", proposta per un database aperto a supporto dei GIS, dei sistemi di navigazione satellitare e della Storia. Conferenza GARR 2018 - Data (R)evolution - Selected Papers, 119-124, DOI: https://doi.org/10.26314/GARR-Conf18-proceedings

Serreli, G. (2024). *Il territorio di Teulada e il sud ovest della Sardegna nella* Relatione de tutti li territorii, et costa marina et luochi nominati cargatori ... *di questo Regno (1581-1582)*, in *Teulada. Storia e società. Nel 400° anniversario della fondazione (1624-2024)*, a cura di Salvatore Loi e Giovanni Murgia, Edizioni Grafica del Parteolla, Dolianova: 133-152.

Serreli, G. (2025). *Territori e Costa di Mare del Regno di Sardegna alla fine del Cinquecento*, in P. Sardina (a cura di), *La terra e il mare. Signori, feudi e porti nel Mezzogiorno medievale e moderno*, Collana "Culture e Società" 17, Palermo University Press, Palermo: 209-241.

Serreli, G., Cocco A. (2026). *Porti minori del Regno di Sardegna alla fine del Cinquecento. Alcuni casi studio: Carbonara e Teulada*, in *Tocco, F.P. (2026) (a cura di). Porti minori di Mezzogiorno, Sicilia e Sardegna dal Medioevo a oggi. Fisionomie storiche e prospettive di riqualificazione, Palermo, Plumelia edizioni:* 315-350.

Simbula, F. (2014). *I porti nello sviluppo economico della Sardegna medievale*, in *Attività economiche e sviluppi insediativi nell'Italia dei secoli XI-XV. Omaggio a Giuliano Pinto*, cur. E. Lusso, Cherasco: 269-305.

Tocco, F.P. (2026) (a cura di). *Porti minori di Mezzogiorno, Sicilia e Sardegna dal Medioevo a oggi. Fisionomie storiche e prospettive di riqualificazione*, Palermo, Plumelia edizioni.

[7/C] Modelli, Metodi e Prospettive dell'Edizione Digitale

Roman and Justinianic Legal Terminology in Thirteenth-Century Western European Diplomatic Sources

TAMÁS KOVÁCS[1], ANGUELOS NICOLAOU[2], JOHANNES LAROCHE[3], GEORG VOGELER[4]

[1] DEPARTMENT OF DIGITAL HUMANITIES, UNIVERSITY OF GRAZ, AUSTRIA – TAMAS.KOVACS@UNI-GRAZ.AT
[2] DEPARTMENT OF DIGITAL HUMANITIES , UNIVERSITY OF GRAZ, AUSTRIA – ANGELOS.NICOLAOU@UNI-GRAZ.AT
[3] DEPARTMENT OF DIGITAL HUMANITIES , UNIVERSITY OF GRAZ, AUSTRIA – JOHANNES.LAROCHE@UNI-GRAZ.AT
[4] DEPARTMENT OF DIGITAL HUMANITIES , UNIVERSITY OF GRAZ, AUSTRIA – GEORG.VOGELER@UNI-GRAZ.AT

ABSTRACT

Monasterium.net[1] hosts a digital corpus of more than 600,000 European charters, offering an unprecedented basis for large-scale research on medieval legal language. Moving beyond analyses limited to individual documents, regions, or archival collections requires the systematic application of automated methods capable of handling sources at a continental scale. Within the ERC-funded project DiDip (From Digital to Distant Diplomatics), traditional diplomatic scholarship is integrated with machine learning techniques drawn from natural language processing. Against this background, the present study investigates the reception of Justinianic Roman law in thirteenth-century Western Europe through the linguistic analysis of charter material. The analysis is based on corpus-wide encoding of lexical sequences and the systematic generation of partially reduced subsequences, allowing formal variation in legal terminology to be captured without sacrificing structural coherence. By representing multiple omission patterns in a unified, numerically encoded format, the method enables robust comparison across documents while remaining sensitive to both local lexical shifts and broader terminological conventions. A focused case study illustrates how Justinianic legal language was adapted, transformed, and regionally differentiated within Western European documentary practice of the thirteenth century.
Keywords: Roman Law Reception, Digital Diplomatics, Legal Terminology, Justinian I

ABSTRACT (ITALIANO)

Monasterium.net ospita un corpus digitale di oltre 600.000 carte europee, offrendo una base senza precedenti per la ricerca su larga scala sul linguaggio giuridico medievale. Superare le analisi limitate a singoli documenti, regioni o collezioni archivistiche richiede l'applicazione sistematica di metodi automatizzati in grado di gestire le fonti su scala continentale. All'interno del progetto DiDip (*From Digital to Distant Diplomatics*), finanziato dall'ERC, la borsa di studio diplomatica tradizionale è integrata con tecniche di apprendimento automatico derivate dal trattamento del linguaggio naturale. In questo contesto, il presente studio indaga la ricezione del diritto romano giustinianeo nell'Europa occidentale del XIII secolo attraverso l'analisi linguistica del materiale documentario. L'analisi si basa sulla codifica a livello di corpus di sequenze lessicali e sulla generazione sistematica di sottosequenze parzialmente ridotte, consentendo di catturare la variazione formale nella terminologia giuridica senza sacrificare la coerenza strutturale. Rappresentando molteplici modelli di omissione in un formato unificato e codificato numericamente, il metodo consente un confronto robusto tra i documenti, rimanendo sensibile sia ai cambiamenti lessicali locali che alle convenzioni terminologiche più ampie. Un caso di studio mirato illustra come il linguaggio giuridico giustinianeo sia stato adattato, trasformato e differenziato regionalmente all'interno della pratica documentaria dell'Europa occidentale del XIII secolo.
Parole chiave: Ricezione del Diritto Romano; Diplomatica Digitale; Terminologia Giuridica; Giustiniano I

1. INTRODUCTION

This study introduces a computationally driven framework for examining the integration of Roman, and specifically Justinianic, legal traditions into medieval documentary practice. While the impact of the *Corpus Iuris Civilis* on medieval jurisprudence has been the subject of sustained scholarly attention, the concrete ways in which its terminology and concepts were adopted in everyday legal documentation across time and space are still insufficiently understood. To address this issue, the study applies

[1] https://www.monasterium.net.

advanced text-analytical methods to a large body of medieval charters, concentrating on Central European material and, in particular, on Austrian documents dating from 1200 to 1400. This period coincides with a crucial phase in the reception of Roman law in Central Europe, offering a dense and well-preserved documentary record suitable for large-scale computational analysis.
Medieval charters were built upon a repertoire of highly conventionalized legal formulas, in which established textual patterns encoded specific juridical actions and effects. From an analytical standpoint, these formulas cannot be treated as fixed units, as they exhibit substantial variation across time and space. Such variability reflects shifts in legal procedure, regional drafting traditions, and the individual practices of notaries and scribes. Consequently, a single legal concept may be rendered through several formally distinct expressions within thirteenth-century Latin documentary sources. The notion of legal certainty or binding force, for example, can be articulated as *ex certa scientia*, *de certa scientia et matura deliberatione*, or *non ignorans sed scienter et consulte* in different charter contexts. This formal diversity poses a significant challenge for the systematic identification and large-scale analysis of formulaic language in extensive charter corpora.

2. STATE OF THE ART

Such intrinsic formal variability presents a substantial obstacle for conventional computational approaches. Techniques that rely primarily on frequency counts or on the exact repetition of word sequences struggle to detect formulas that are expressed through divergent linguistic realizations. Orthographic inconsistency, syntactic flexibility, and variation in word order further limit the effectiveness of standard sequence-based methods. Approaches that operate on fixed-length word sequences are particularly vulnerable to even minimal textual modifications (Manning et al., 2008). In medieval charter production, scribes frequently expanded established formulas with context-specific information, for example *super controversia mota inter abbatem monasterii sancti Martini et universitatem civium de Alba*, thereby interrupting otherwise stable formulaic structures. Because such interpolations alter the surface form of the text, conventional sequence-matching techniques fail to capture the underlying regularities (Koolen & Hoekstra, 2022). These insertions were not exceptional but constituted a routine aspect of documentary practice, reflecting the need to adapt inherited legal formulas to concrete procedural and social circumstances (Korkiakangas & Passarotti, 2011).
While approaches that permit non-adjacent word matching provide greater flexibility than strictly contiguous sequence models, their reliance on narrowly defined local contexts constrains their ability to capture the extended dependencies that are characteristic of legal formulaic language (Mikolov et al., 2013). Methods operating on gapped sequences can accommodate a certain degree of variation, yet their analytical reach remains bounded by the size of the contextual window. In medieval documentary practice, legal formulas frequently comprise complex conditional constructions whose logical components may be distributed across lengthy textual spans, exceeding the limits of such models. Extended formulations, for instance *quod si prefatus emptor infra terminum constitutum solutionem non fecerit, venditio presentis instrumenti robur obtineat*, often place the conditional premise and its juridical consequence at a considerable distance from one another. As a result, crucial structural relationships—such as those linking *quod si* to *robur obtineat*—risk remaining undetected within standard gapped-sequence frameworks.

3. METHOD

The methodological cornerstone of this study lies in the development of a specialized computational framework engineered for the systematic investigation of formulaic structures within medieval legal corpora. Central to this approach is an innovative n-gram modification technique that facilitates the identification of juridical formulas by transcending the inherent inconsistencies of the period's manuscript culture. Unlike traditional string-matching methods, this system is designed to detect semantic and functional equivalencies despite pervasive orthographic instability, lexical variations, or fluctuations in syntactic arrangement. By prioritizing the underlying conceptual consistency of legal expressions over rigid verbatim matching, the tool provides a robust solution to the challenges posed by scribal idiosyncrasies, thereby enabling a more nuanced mapping of the evolution of medieval legal discourse.
The operational architecture of this analytical framework relies on a sophisticated bipartite procedural sequence. In the initial phase, a corpus-wide token encoding protocol is employed, transforming lexical units into numerical sequences (Shannon, 1948). This discretization facilitates high-efficiency pattern

matching while preserving the granular contextual nuances of the source material. By mapping linguistic data into an integer-based format, the system effectively mitigates the interference of orthographic volatility, allowing for the identification of analogous juridical terminology through a more robust mechanism than traditional keyword-based queries. The secondary phase involves a stochastic subsequence generation technique, which systematically omits a predetermined number of elements within a given textual segment. By generating multiple variations of a sequence through the strategic removal of specific positions, the methodology isolates the core structural template of a legal formula. This approach is particularly effective in addressing the fluid nature of medieval documentation, as it recognizes the essential conceptual framework of a legal expression even when obscured by the stylistic adaptations, structural transpositions, or lexical substitutions typical of diverse scribal traditions.

To optimize computational performance, the framework utilizes an array-based encoding architecture that organizes integer-encoded sequences into structured matrices, where each row represents a distinct permutation of omitted elements. This unified data representation allows for the simultaneous evaluation of potential matches via the Intersection over Union (IoU) metric, facilitating an immediate assessment of both localized linguistic clusters and broader structural dependencies. By leveraging this multidimensional approach, the system achieves a comprehensive overview of pattern similarity that transcends simple linear comparisons. While the current implementation relies on the IoU metric for similarity detection, the system's modular architecture is designed to accommodate increasingly sophisticated evaluative criteria. This includes the potential for weighted analysis across various categories of correspondence, such as verbatim alignment, partial overlaps, syntactic reordering, and lexical substitutions. Such a versatile evaluative framework empowers researchers to calibrate the sensitivity of the analysis to align with specific philological objectives or the unique diplomatics of a given manuscript collection, ensuring that the methodology remains responsive to the nuanced demands of historical legal research.

The analytical utility of this framework is significantly augmented by its visualization suite, which serves as a critical interface for historical inquiry. By generating interactive HTML outputs, the system facilitates side-by-side textual comparisons and employs heatmaps to represent similarity densities across the entire corpus. This graphical representation allows scholars to identify "bridge words"—lexical pivots that link related juridical concepts—thereby uncovering hidden connections within the legal vernacular. Furthermore, the integration of diachronic analysis features enables the longitudinal study of linguistic evolution across diverse institutional contexts. By adopting a "distant reading" perspective, the system elucidates the dialectic between the static and fluid components of charter formulas, providing empirical evidence for the transformation of legal traditions. The multi-scalar nature of these visualizations allows for a seamless transition from macro-level institutional trends to micro-level philological scrutiny, ensuring that the granular details of the text remain accessible within a broad computational landscape.

The empirical foundation of this research is a specialized corpus of thirteenth-century Latin charters, systematically retrieved from Central European archives via the monasterium.net digital repository. While the digital availability of these primary sources offers an unprecedented opportunity for large-scale quantitative analysis, the computational processing of Medieval Latinity presents unique challenges, particularly regarding the tension between standardized legal formulas and regional scribal idiosyncrasies. By applying time-series clustering analysis, the study identifies synchronic patterns in the adoption and diffusion of specific legal terminologies across diverse ecclesiastical and secular institutions. Treating these 13th-century Latin expressions as longitudinal data points allows for a highly granular mapping of the reception of Roman law. This quantitative rigor reveals latent patterns of legal transmission that traditional qualitative diplomatic analysis might overlook, offering fresh insights into the professionalization of legal literacy during this pivotal century.

The systematic processing of the charter corpus is executed through a multi-stage methodological pipeline. Initially, documents are subjected to a stratified classification, categorized by their provenance, chronological data, and diplomatic typology. This taxonomy is essential for evaluating the diffusion of legal norms across various 13th-century institutional frameworks. In the second phase, the raw textual data undergoes rigorous preprocessing to mitigate the linguistic volatility inherent in High Medieval Latin. This involves the standardization of orthographic variations and the systematic expansion of scribal abbreviations—a necessary step to ensure the accuracy of subsequent algorithmic operations. Finally, our framework is deployed to detect and track recurring juridical patterns within the rectified texts. These results are subsequently subjected to a methodological triangulation, where computational outputs are cross-referenced with traditional historical heuristics. This dual-verification process ensures that the identified patterns possess both statistical significance and genuine historiographical validity.

The application of this analysis facilitates the identification of distinct chancery affiliations based on idiosyncratic patterns of juridical nomenclature (Härtel, 2015). Initial findings suggest that episcopal chanceries may have acted as early adopters of Roman legal concepts and terminology (Willoweit, 2000). This proactive reception is likely attributable to their profound engagement with classical pedagogy and a sophisticated command of Justinian's *Corpus Iuris Civilis* (Huschner, 2003). Incontrast, royal courts appear to have utilized a more calibrated and methodical integration of novel legal terms, synthesizing Roman principles with established customary frameworks (Bates, 1995). Such a deliberate approach suggests a strategic effort to modernize the administrative apparatus while ensuring institutional continuity. Conversely, the dissemination of Roman law within municipalities was frequently characterized by a temporal lag and a high degree of selectivity, prioritizing immediate pragmatic applications. Urban centers with robust mercantile networks likely adopted Roman legal terminology with greater alacrity than more isolated communities, thereby illustrating significant regional disparities in the diffusion of 13th-century legal literacy.

The systematic analysis of these clusters offers a transformative perspective on whether the dissemination of Roman legal nomenclature followed institutional hierarchies or geographical proximity, potentially contesting established paradigms regarding the transmission of legal innovation in High Medieval Europe. By conducting a diachronic investigation into the adoption rates of these terms, the study can identify specific chronological "peaks," which serve as indicators for the underlying drivers of change—be they systemic legal reforms, the influence of prominent jurists, or broader socio-political catalysts. Furthermore, the degree of alignment between these emergent patterns and documented legislative milestones or the establishment of new *studia generalia* provides a metric for the coordination of legal transformation. Correlating this timeline with major historical shifts—such as the proliferation of universities and the transregional migration of legal experts—facilitates a more rigorous determination of the primary vectors of influence. Ultimately, this multifaceted analysis seeks to isolate the predominant factors—whether institutional, spatial, or intellectual—that governed the reception and integration of Roman legal terminology during the 13th century.

While the presented framework demonstrates significant transformative potential for the analysis of medieval legal corpora, its current implementation is subject to certain methodological constraints that delineate clear trajectories for future refinement. A primary limitation lies in the system's current reliance on structural and syntactic analysis at the expense of semantic depth; this can occasionally result in the clustering of functionally divergent formulas that happen to share structural similarities. Addressing this requires the future integration of semantic analysis modules and domain-specific ontologies to better distinguish between the nuanced legal intentions of the scribes. Furthermore, the computational scalability of the system remains a consideration, as processing demands increase proportionally with corpus expansion, necessitating precise parameter calibration tailored to specific diplomatic categories. To mitigate these technical challenges, ongoing research is directed toward the development of optimized processing algorithms and more granular similarity metrics. Future iterations will focus on augmenting the framework's analytical precision by incorporating lexical semantics and specialized legal knowledge bases, thereby enhancing its capacity to differentiate between structurally analogous but functionally distinct juridical expressions.

Despite the aforementioned challenges, this methodology provides a substantial contribution to the fields of Digital Humanities and medieval legal history. Primarily, it demonstrates the efficacy of computational heuristics in navigating expansive corpora of medieval documentation, establishing a reproducible paradigm that can be transposed to other chronological or geographical contexts. The versatility of this framework makes it particularly suitable for any historical dataset characterized by formulaic syntax or standardized linguistic tropes. Furthermore, the research provides a transformative perspective on the reception of Roman law in Central Europe, revealing a complex, institutionally driven process that contradicts traditional, more linear narratives of legal dissemination. By highlighting the intricate interplay between administrative structures and normative transformations, the study proposes a new interpretive model for understanding the professionalization of medieval legal culture. Finally, the development of this specialized tool offers a robust analytical engine for the study of formulaic language that transcends the boundaries of legal history. Its adaptable pattern-matching capabilities and nuanced similarity metrics possess broad applicability for the computational analysis of various historical genres where repetitive structural templates are a defining characteristic.

ACKNOWLEDGEMENTS

The researchers associated with this study are members of the ERC-funded project DiDip, based within the Department of Digital Humanities at the University of Graz. The conceptualization and algorithmic development of the computational framework were spearheaded by Tamás Kovács, with specialized technical support provided by Anguelos Nicolaou during the code implementation phase. Johannes Laroche was responsible for the rigorous data evaluation and diplomatic source criticism, while Georg Vogeler provided overarching supervision and critical theoretical guidance throughout the research process. The authors express their profound gratitude to the European Research Council for the generous financial support that made this inquiry possible. Furthermore, we extend our thanks to the International Centre for Archival Research (ICARUS) and the various contributing archives for their invaluable cooperation in providing digital access to the primary source material via the Monasterium.net repository.

REFERENCES

Bates, D. (1995). Le rôle des évêques dans l'élaboration des actes ducaux et royaux entre 1066 et 1087. *Les Évêques Normands Du XIs. : Colloque de Cerisy-La-Salle, 30 Septembre-3 Octobre 1993*, 103–115.

Härtel, R. (2015). Urkundenlandschaften zwischen Donau, Rhein und Adria. In *Andreas Schwarcz/Katharina Kaska (Hgg.), Urkunden – Schriften – Lebensordnungen. Neue Beiträge zur Mediävistik. Vorträge der Jahrestagung des Instituts für Österreichische Geschichtsforschung aus Anlass des 100. Geburtstags von Heinrich Fichtenau (1912–2000) (Wien, 13.–15. Dezember 2012)* (Vol. 63, pp. 193–211). Beck.

Huschner, W. (2003). *Transalpine Kommunikation im Mittelalter: Diplomatische, kulturelle und politische Wechselwirkungen zwischen Italien und dem nordalpinen Reich (9.-11. Jahrhundert)* (Vol. 52). Hahnsche Buchhandlg.

Koolen, M., & Hoekstra, F. G. (2022). Detecting Formulaic Language Use in Historical Administrative Corpora. In F. Karsdorp, A. Lassche, & K. Nielbo (Eds.), *Proceedings of the Computational Humanities Research Conference 2022* (pp. 127–151).Willoweit

Korkiakangas, T. (2024). *A linguist's viewpoint: Formulaic language as a challenge for historical linguistics*. https://doi.org/10.5281/ZENODO.10461847

Korkiakangas, T., & Passarotti, M. (2011). Challenges in Annotating Medieval Latin Charters. *Journal for Language Technology and Computational Linguistics*, *26*(2), 105–116. https://doi.org/10.21248/jlcl.26.2011.150

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge university press.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [Cs]*. http://arxiv.org/abs/1301.3781

Shannon, C. E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal*, *27*, 379–423, 623–656.

Willoweit, D. (2000). Grundherrschaft und Territorienbildung. Landherren und Landesherren in deutschsprachigen Urkunden des 13. Jahrhundert. In *Strukturen und Wandlungen der ländlichen Herrschaftsformen vom 10. Zum 13. Jahrhundert. Deutschland und Italien im Vergleich, hg. V. Gerhard DILCHER and Cinzio VIOLANTE* (Vol. 14, pp. 215–233). Duncker & Humblot.

In Search of Lost Time: an Innovative Model for Representing Autograph Texts

GIULIA BALDELLI[1], DANIELE FUSI[2], MATTEO ZUPANCIC[3], FRANZ FISCHER[4], CLAUS ZITTEL[5]

[1]STUTTGART UNIVERSITY, GERMANY – GIULIA.BALDELLI@ILW.UNI-STUTTGART.DE
[2]STUTTGART UNIVERSITY, GERMANY – DANIELE.FUSI@TS.UNI-STUTTGART.DE
[3]UNIVERSITÀ DI PISA, ITALY – MATTEO.ZUPANCIC@UNIPI.IT
[4]UNIVERSITÀ CA' FOSCARI VENEZIA, ITALY, FRANZ.FISCHER@UNIVE.IT
[5]UNIVERSITÀ CA' FOSCARI VENEZIA, ITALY, CLAUS.ZITTEL@UNIVE.IT

ABSTRACT (ENGLISH)

This paper introduces the ongoing evolution of a novel digital model for representing and visualizing autograph texts in the context of genetic philology. Departing from traditional approaches that focus on comparing static versions extracted from an autograph sheet, this model shifts the focus from the effects (the texts) to the cause (the creative process itself). By leveraging a process-oriented, time-aware data structure, the model captures the dynamic flow of text creation and alteration, offering a more faithful and scalable representation of complex, annotated documents for scenarios like Goethe's *Venetian Epigrams*.
Keywords: digital scholarly editing; genetic editions; modelling; Goethe Venetian Epigrams.

ABSTRACT (ITALIANO)

Questo articolo presenta l'evoluzione di un nuovo modello digitale per la rappresentazione e la visualizzazione di testi autografi nel contesto della filologia genetica. Partendo dagli approcci tradizionali che si concentrano sul confronto di versioni statiche estratte da un documento autografo, il modello sposta l'attenzione dagli effetti (i testi) alla causa (il processo creativo stesso). Sfruttando una struttura dati orientata al processo e sensibile al tempo, il modello cattura il flusso dinamico della creazione e dell'alterazione del testo, offrendo una rappresentazione più fedele e scalabile di documenti complessi e annotati per scenari come gli Epigrammi Veneziani di Goethe.
Parole chiave: edizione scientifica digitale; filologia genetica; modellazione; Epigrammi veneziani di Goethe.

1. A CRITICAL DIGITAL EDITION OF GOETHE'S *VENETIAN EPIGRAMS*

In the spring of 1790, Johann Wolfgang von Goethe returned to Italy for the second time, arriving in the lagoon city of Venice. He lodged in a room overlooking the Canal Grande at the inn *All'insegna della Tromba*, where he was registered under the name Baron "Giovanni Gaeta" (also recorded as "Gassetta" or "Gajeta") from Saxony.[1] Going by this Italianized identity, Goethe spent nearly two months – from 31 March to 22 May – wandering the city's narrow streets, visiting its many artistic landmarks, observing its resonant and licentious urban life, and pursuing his emerging ambition to develop a morphology of animals. At the intersection of these impulses, and aided by the boredom induced by persistent rains that soaked the Venetian landscape, he began to jot down a series of brief poetic sketches. These would quickly grow into hundreds of compositions arranged in a series of handwritten notebooks which, chaotically yet progressively, took shape as the *Venetian Epigrams* (VE), conceived as a companion piece to the more famous *Roman Elegies*.
As Claus Zittel has observed in his 2023 article, the philological point of departure for the VE is remarkably unstable.[2] Despite an extensive and largely well-preserved manuscript tradition, the epigrams have not yet been edited in accordance with modern scholarly standards. This situation is primarily due to their complex history of composition and transmission, which is characterized by repeated interruptions and gaps in manuscript preservation. Further complicating matters is a dense interplay of self-censorship and external censorship, coupled with the fundamental impossibility of establishing a single authoritative version of the VE. The large number of textual witnesses, the fluctuating number of epigrams contained in each, and the abundance of variant readings render the reconstruction of their

[1] Oswald (2014), pp. 31-39.
[2] Zittel (2023), pp. 169.

genesis and transmission highly problematic. Ongoing processes of numbering and renumbering across manuscripts not only impede the systematic comparison of variants, but also generate unstable sequences that are difficult to reproduce adequately in print.
The manuscripts moreover exhibit multiple layers of revision: corrections, deletions, additions, marginal annotations, and alternative readings were introduced by several unknown hands at different times, using various writing instruments and inks. These interventions – ranging from punctuation changes and shifts in word order to parallel variants and distinct drafting stages – are often visible in the material record but cannot always be arranged into a clear chronological sequence.
The VE can therefore be understood not as a fixed text, but as a *textum mobile*. No single manuscript or print defines the work; rather, its textual identity emerges from the totality of its variants. The epigrams continuously reconfigure themselves across differing contexts and constellations. Their aesthetic identity is constituted precisely through variation, recombination, and contextual displacement, resulting in multiple, coexisting forms of textual identity. These identities are not the product of organic development, but of a combinatorial – and at times even aleatory – process. This necessitates a dynamic concept of the text and, concomitantly, a revision of the traditional notion of the "work," which the protean and fluid nature of the VE fundamentally resists.
Against this backdrop, the project develops a digital scholarly edition capable of representing the VE not merely as a sequence of textual states, but as a network of textual movements, both diachronic and synchronic. The platform will make visible, for each epigram, its transformations across different stages, the pretexts on which it draws, and the new contexts it enters. The primary aim is to model the *textum mobile* in its full complexity and thereby to render the entire corpus more accessible for philological research. Methodologically, the digital edition will trace the movements and transformations of individual epigrams across their various configurations while simultaneously situating them within each respective arrangement. It will virtually reconstruct Goethe's proposed organizational schemes at different stages and document, as far as possible, processes of revision. In addition, the edition will enable the formation of thematic groupings, the identification of recurring motifs, and the analysis of metrical patterns and correspondences.
In this way, the VEdition will provide readers, for the first time, with a platform to uncover the rich compositional history of a work that, despite its central position within Goethe's corpus, has been deeply shaped both by the vividness of its content and by its markedly open-ended nature. Additionally, it will serve as a means of exploring new ways of conceiving the editorial presentation of a complex and multi-layered textual development, by preserving the meaningful autonomy of each individual text carrier and each specific textual version, without subsuming them under the hierarchy of a presumed "final version" that cannot, in fact, be established. The fluid, almost metamorphic quality of the interplay between the base layer and its array of alterations can also be visually observed through animations, thereby introducing a new and heuristically powerful tool for both hermeneutic inquiry and didactic practice.

2. DIGITAL MODEL

The representation of autograph texts — original manuscripts with complex layers of annotations representing variants of its text — poses significant challenges for digital philology. In many traditional digital approaches, the autograph document (here named carrier) serves as the material evidence from which multiple virtual versions are derived. These versions are of course abstractions, defined by interpreting the carrier's content. Markup languages such as TEI are then used to connect them, marking deletions and additions relative to other versions — a process akin to "diffing" texts. This allows us to leverage existing critical edition practices to extract and compare multiple static versions from a document, using markup to annotate differences. We can then present text versions side to side, emphasizing their differences, and thus implicitly pointing at the creative process behind them. This is fit to methods and tools based on traditional text criticism, essentially based on comparing witnesses to build a reconstructed text, like comparing versions to hint at the transformation process which produced them.
While effective for many scenarios, this method might be less convenient with the high complexity and conceptual demands of deeply annotated documents like those of *Venetian Epigrams*, where literally dozens of alternative texts may coexist at the same level, reflecting an ongoing creative process rather than a finalized work. A systematic comparison and the consequent markup, connecting a big number of different versions where no one can claim to be the final one, would be rather difficult to scale. Moreover, the traditional approach still focuses on the *effects* of the underlying creative *process*, requiring the au-

dience to mentally reconstruct it from the differences surfaced on texts. The focus is on aligning versions and detailing differences to elucidate the process that generated them. Yet, the carrier with all its chaotic annotations is the witness of a creative process in the author's mind which progressively unfolded over time[3]; all the variants we derive and align are just its effects. The dynamic nature of this process is intrinsically difficult to fit into the presentation of multiple static texts.

In a digital edition, however, we can try to exploit time as our ally, instead of our enemy. The same time which accumulated all those changes on a text, making it a puzzle for scholarly interpretation, can be the solution to a more effective presentation of the creative process, reproducing a hypothesis about its reconstruction, unwinding versions one after another via animations.

So, the proposed model advocates a dramatic shift in focus: from the *texts* extracted from the carrier to the *process* that generated them. The carrier is seen not merely as a repository of texts, but as a witness to the creative process, with annotations serving as traces of the author's evolving intentions. A process being essentially a sequence of actions in time, a text making process is just represented as a sequence of editing operations (deleting, adding, replacing, moving, swapping) applied over time. Each operation produces a new version, and the accumulation of these operations constitutes the creative flow.

This implies that at this higher granularity level not all the outputs can be regarded as a self-contained text; they are rather steps along a transformation path. Only certain versions — those corresponding to significant alteration stages — are treated as waypoints along this path. This allows for the representation of multiple texts even in the absence of a final stage, with all stages coexisting as equally valid alternatives; and thus, it fits a scenario where a "constellation" of texts coexist on the same plane. There, the various text stages are like stars derived from the same nebula (the creative process), with the carrier acting as a telescope that captures the light (annotations) from the past, like in a photographic snapshot, thus allowing scholars to formulate a reconstructive hypothesis.

This shift from text outputs to the process generating them and reflected on the carrier requires us to define a new model, capable of composing many tensions: single vs. multiple documents (the carrier is materially one, but it virtually represents multiple versions of the text); objective vs. subjective layers (the carrier is the material evidence, while the reconstruction is subjective); and textual vs. graphical representations (as diplomatic data here is a foundational part of the reconstruction).

In this context, composing the first tension (single vs. multiple documents) was accomplished using a single data structure, termed "chain", capable of representing any number of text versions. The base text is modeled as a set of character nodes in a directed graph, linked one after another to build sequences. So, the chain includes a set of nodes, and a set of links, each tagged with the version it belongs to. With reference to the chain, editing operations just add new links (and possibly nodes, in case of insertions or replacements, which imply new characters), thus creating new versions without duplicating the entire text, while constantly increasing the items in those sets. For example, deleting a word involves creating a bypass link that skips the deleted portion, while adding or replacing text introduces new nodes and links. This approach ensures that all versions are efficiently represented within a single, compact structure defined via operations, which in turn are the essence of our process (Figure 1[4]).

[3] When speaking of time, we are not implying a historical interpretation which orders one operation after another as it were going to reflect the real sequence of events occurred in the text making process. There is a tendency towards a historically plausible reconstruction, but the order within each alteration stage is just conventional. Time here is the focus of the model in the sense of being an additional feature of it, where information *accumulates* both on the graphical and on the textual layer. Every editing operation adds information for the next state of the text; so, in a sense, earlier states are those with fewer information, and later states are those with more. This has some resemblance with information theory applied to time in physics viewed as the 'cumulative record' of what happened.

[4] Figure 1 shows 3 versions of a text: v0 "A red pen", v1 "A pen" (delete "red_"), v2 "A pencil" (insert "cil"). From left to right, just follow the version tag to compose the corresponding sequence of characters. Characters are nodes, linked by edges tagged with the corresponding version. Every version adds new edges, and possibly new nodes, to the chain, while nothing ever gets removed from the set. This accumulates more and more data as 'time' goes by, while preserving a compact model and representing each character with a single entity; just like on a manuscript the character once written stays there forever, and may become part of different sequences via annotations.

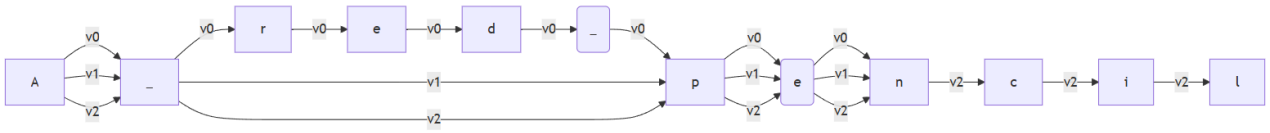


Figure 1. A chain sample

The same model also supports the annotation of both text and process through metadata linked to operations. So-called "features" — essentially name/value pairs[5] — can be attached to nodes or output versions, enabling the documentation of reasons for changes (e.g., metrical adjustments, ink color, different hands) and other relevant information. So, the chain not only generates text, but also injects annotations into it.

This solves the tension between a single data structure and the multiple texts computed from it. At the same time, this architecture maintains a clear separation between the objective layer (the material carrier and its signs, which get a fully diplomatic representation) and the subjective layer (the interpreted creative process). The chain structure, fed by editing operations, generates all possible text versions and their annotations, preserving both layers for scholarly analysis.

As for the tension between textual and graphical data, the model covers both textual data (the generated texts) and graphical data (the visual appearance of signs on the carrier). Here, the original approach for graphical representation has evolved, recognizing the relevance of a more abstract and economical strategy. The original, graphical approach faithfully reproduces the facsimile image, literally re-drawing hints on top of it (via SVG), and positioning added text according to diplomatic metadata. Animations are employed to represent the temporal flow of operations. This method, while accurate, can be complex and time-consuming, unless driven by a more automated data flow for imaging.

To enhance efficiency, a new, symbolic approach has been introduced, abstracting the visualization process.[6] Instead of redrawing each sign, operations are described using rendition features (like "diagonal line", "box", "cross", "below and to the left", "shifted right", etc.), while the visualization engine uses this visual grammar to do the math to properly size and position each element.

To this end, all signs on a carrier are grouped into three categories: base text (the text laid out as first on the sheet, representing the starting point for its transformation), added text (e.g. an interlinear word replacing a word in the base text), and so-called "hints", i.e. all the other signs whose purpose is to hint at the corresponding editing operation. This classification enables distinct rendering logic for each type, facilitating accurate digital visualization. For instance, say in a line of the base text the word "Wund:" (with colon) gets a diagonal stroke on it, while a smaller word "Wunder." (with dot) is written below and to the left of it. The stroke here is the hint, representing the visual counterpart of a deletion, while the added text is the visual counterpart of an addition. Taken together, these two represent a replacement operation. In the symbolic model, this operation might have features like (from a controlled vocabulary) "hints=diagonal stroke", "position=south-west", "font-size=smaller"; the rendering engine will take care of picking the hint for diagonal stroke from a catalog of visuals, resize it to fit the word being deleted, position it over that word, and draw the added text with a smaller size at the position specified.[7] The same catalog including the drawings representing hints, like lines or shapes, also links each hint to its designed animation for entering the visualization stage. This way, a single feature attached to an operation can pick the corresponding drawing from this catalog, and reuse it wherever needed, even across different projects, implicitly getting also the enter animation for it.

[5] According to the semantic relevance of time in this model, features also include additional metadata in the dynamic flow where they get injected by operation, defining their behavior and lifetime in the process.

[6] For reasons of economy here we can only provide an overview of this model. You can find more about it together with a demo of the hint designer at https://vedph.github.io/gve-doc/model/diplomatic.html.

[7] It should also be added that side by side with those lower-level rendition features, which directly specify position, color, font size, and the like, this approach also allows for the definition of higher-level features, acting as a shortcut towards multiple lower-level counterparts. For instance, a number added to epigrams could be encoded with a single higher-level feature, like epigram-number, and later automatically translated into lower-level features like smaller number, a brown ink, rotated by about -45 degrees, at the top-left of the epigram, rather offset from it. This results in an even more abstract and thus more economic annotation strategy.

This allows for consistent, scalable rendering across projects, a more human-friendly description, and also detailed metadata about the graphical typologies used, which could later be used as search criteria or diplomatic data to feed machine learning-based computations.
Also, it should be emphasized that animation here, whatever the approach (graphical or symbolic), far from being a fancy presentational feature is crucial in the integration of time as a semantic dimension. Animations replay the sequence of editing operations, allowing users to witness the transformation of the text step by step. This dynamic visualization makes the creative process explicit, moving beyond static comparisons to reveal the flow of textual genesis.
By focusing on the process rather than static stages, the model embraces the flux of time and aligns more closely with the reality of text creation. Both textual and graphical elements (hints) accumulate over time, mirroring the irreversible nature of ink marks on paper in real world. There, once a sign is added — whether text or graphical — it remains forever, with subsequent operations building upon it: for instance, we may represent the deletion of a word by drawing a stroke on it; but the deleted word is still there. The digital model here mimics this reality, both for text and visuals: just like in the textual model (the chain) we have two sets of nodes and links which get added once and stay there forever, so in the visual model we accumulate signs on the display surface operation after operation. At the end, we fully match what is found in the carrier, while time — the fact that we added these signs one at a time, relating it to a specific editing operation — is represented by animations progressively revealing it and thus vividly presenting our reconstructive hypothesis. One of the best demonstrations of the intrinsically digital nature of such a dynamic, process-based edition is precisely the fact that there is no other way of showing its result than having it playing over time under the audience's eyes.[8]
This allows us to face the challenges posed by the peculiar scenario of *Venetian Epigrams*, while also trying to provide a model and a set of tools designed from the ground up for reuse in other projects. In fact, the whole data model, with its chain capable of dynamically generating annotated texts, its nodes and links, and the operations acting on them and representing the core of the reconstructed creative process, can be created by just using a web-based editor UI. Its data gets stored in a standard document database and can be easily dumped into a full textual encoding (based on JSON), as well as exported into TEI, RDF, etc. for purposes different from the highly interactive visualization illustrated above.
Even the visualization software required to provide its truly digital and native presentation, focused on progressively unveiling the making of a text, step after step over time, thus promoting a sequence of stills into a sort of historical movie, is part of a set of software tools designed for reuse in any third-party environment, from a static HTML page to any higher-level JS framework. To this end, these core components are packaged into open-source libraries and based on standard W3C technologies: for instance, while the backend chain-based logic can be available via REST API (distributed in Docker-based images), the frontend visualization is implemented by W3C web components, ready for integration and reuse with minimal effort. Even graphical assets, like the visuals catalog including hints and animations, not only can be shared across projects, but also edited via another web component designed right to support in their creation, based on a mixture of SVG and very small JS fragments leveraging a popular animation framework.
This also implies yet another engagement option for the audience: just like scholars working in this edition define their own reconstruction of the process, so that it can later be played before the user's eyes, the community itself could be involved in critically evaluating this work and even propose different reconstructions by leveraging the same tools used by the authors themselves. They would then see the result playing before them in real time and may later be integrated into the edition itself. This kind of flexibility and extension is precisely the purpose of a set of digital models and tools born to face the challenge of a specific work, but designed to provide also a scholarly paradigm and practical publishing system, thus fully embracing the flux of time, not only for the protean nature of such texts, but also for the ever-changing assets of digital editions.

REFERENCES

Andrews, T. & Diehr, F. & Efer, Th. & Kuczera, A. & Zundert, J. van (eds.). (2022). Graph Technologies in the Humanities - Proceedings 2020 CEUR Workshop Proceedings.
Last accessed September 13, 2024 at https://ceur-ws.org/Vol-3110/.

[8] See e.g. the demo at https://gve-rendition.surge.sh/?sample=h5-48 (the play button starts the show).

Buzzoni, M. & Cucurnia, D. & Fenu, C. & Rosselli Del Turco, R. & Tancredi, G. (2024). Progetto di edizione genetica digitale del Canzoniere manoscritto di U. Saba (1919-20). In: Me.Te. Digitali. Mediterraneo in rete tra testi e contesti, Antonio Di Silvestro & Daria Spampinato (eds.), 215-220.
Last accessed September 13, 2024 at https://hdl.handle.net/10278/5078681.

Cummings, J. (2018). "A world of difference: Myths and misconceptions about the {TEI}." In: Digital Scholarship in the Humanities.
Last accessed September 13, 2024 at https://doi.org/10.1093/llc/fqy071.

Genovesi, C. & Isolani, A. & Lo Rito, C. & Marotta, D. & Matteoli, M. & Tozzinni, C. (2009). "Topic Maps and MVD for the Representation of Interpretative Variants." In: Digital Humanities. Conference Abstract, University of Maryland, College Park, June 22-25, 2009. The Maryland Institute for Technology in the Humanities (MITH).
Last accessed September 13, 2024 at https://dhabstracts.library.virginia.edu/works/948.

Haentjens, D. R. & David, J. B. (2017). "It's more than just overlap: Text As Graph." Presented at Balisage: The Markup Conference, Washington, DC, August 1-4, 2017. In: Proceedings of Balisage: The Markup Conference 2017. Balisage Series on Markup Technologies, vol. 19.
Last accessed September 13, 2024 at https://doi.org/10.4242/BalisageVol19.Dekker01.

Oswald, S. (2014). Früchte einer großen Stadt – Goethes Venezianische Epigramme. Heidelberg: Winter.
Pierazzo, E. & Rehbein, M. & Galley, A. "An Encoding Model for Genetic Editions."
Last accessed September 13, 2024 at https://tei-c.org/Vault/TC/tcw19.html.

Sahle, P. (2016). "What Is a Scholarly Digital Edition?" In: Digital Scholarly Editing: Theories & Practices, edited by Matthew James Driscoll & Elena Pierazzo, 19-40. Open Book Publishers.
Last accessed September 13, 2024 at https://doi.org/10.11647/OBP.0095.02.

Schmidt, D. & Colomb, R. (2009). "A data structure for representing multi-version texts online." In: International Journal of Human-Computer Studies 6 (67), 497-514.
Last accessed September 13, 2024 at https://doi.org/10.1016/J.IJHCS.2009.02.001.

Schmidt, D. (2009). "Merging Multi-Version Texts: a Generic Solution to the Overlap Problem." Presented at Balisage: The Markup Conference 2009, Montréal, Canada, August 11-14, 2009. In: Proceedings of Balisage. The Markup Conference 2009. Balisage Series on Markup Technologies, vol. 3 (2009).
Last accessed September 13, 2024 at https://doi.org/10.4242/BalisageVol3.Schmidt01.

Schmidt, D. & Fiormonte, D. (2010). "Multi-Version Documents: A Digitisation Solution For Textual Cultural Heritage Artefacts." In: Intelligenza Artificiale 1 (4), 56-61.

Schmidt, D. & Eggert, P. (2019). "The Charles Harpur Critical Archive." In: International Journal of Digital Humanities.
Last accessed September 13, 2024 at https://doi.org/10.1007/s42803-019-00021-9.

Turska, M. & Cummings, J. & Rahtz, S. (2016). "Challenging the myth of presentation in digital editions." In: Journal of the Text Encoding Initiative 9.
Last accessed September 13, 2024 at https://doi.org/10.4000/jtei.1453.

Zittel, C. (2023). "'…denn was ich berühre, / Wird mir unter der Hand gleich ein behendes Gedicht'. Textuelle Metamorphosen und poetisches Kalkül in Goethes Venezianischen Epigrammen." In: Verwandlung der Worte. Textuelle Metamorphosen in Goethes Schriften. Fassungen, Ausgaben, Übersetzungen. Catalano G., & Sampaolo G. (eds.), Roma: Studi Germanici, 167-191.

Naples Dante Project. Un ambiente integrato per l'analisi bibliografica, filologica e iconografica

GENNARO FERRANTE[1], ANDREA LA VEGLIA[2], DANIELE FUSI[3], STEFANO ANGELO RIZZO[4], ANGELO EUGENIO MECCA[5], DANIELE DURANTI[6], ANNA SVIRIDOVA[7]

[1] UNIVERSITÀ DEGLI STUDI DI NAPOLI "FEDERICO II", GENNARO.FERRANTE@UNINA.IT
[2] UNIVERSITÀ DEGLI STUDI DI SALERNO, ALAVEGLIA@UNISA.IT
[3] UNIVERSITÀ CA' FOSCARI VENEZIA, DANIELE.FUSI@UNIVE.IT
[4] UNIVERSITÀ DEGLI STUDI DI CATANIA STEFANO.ANGELO.RIZZO@GMAIL.COM
[5] UNIVERSITÀ DEGLI STUDI DI NAPOLI "FEDERICO II", ANGELO.MECCA@UNINA.IT
[6] GRUPPO META, DANIELE.DURANTI@DEDAGROUP.IT
[7] GRUPPO META, ANNA.SVIRIDOVA@DEDAGROUP.IT

ABSTRACT (ITALIANO)

Il paper presenta l'ecosistema del *Naples Dante Project* (NDP), sviluppato nel PRIN 2022 e potenziato in collaborazione con Changes, nell'ambito dello Spoke 3, illustrandone la pipeline digitale. L'attenzione è rivolta a: l'acquisizione e normalizzazione delle immagini (*IIIF*, imaging multimodale); la strutturazione modulare e astratta dei metadati; la classificazione semantica dei dati tramite un'ontologia dedicata e l'esposizione come *Linked Open Data*; l'elaborazione automatica (*HTR*, *Handwriting Identification*) per l'analisi paleografica e testuale; l'allineamento dei testi della tradizione manoscritta antica e della moderna prassi editoriale filologica. Modelli granulari e strumenti avanzati abilitano interrogazioni sui contenuti testuali, linguistici e iconografici, nonché sui testimoni materiali della *Commedia*, integrando standard interoperabili (TEI, IIIF, RDF) e tecnologie di imaging, trascrizione e collazione (*eScriptorium*/*Kraken*). Il paper riflette su come la svolta computazionale stia ridefinendo le pratiche filologiche, paleografiche e storico-artistiche, attraverso pipeline verificabili che connettono immagini, testi, metadati e conoscenza automatica, in un'ottica di sostenibilità editoriale a lungo termine.
Parole chiave: Semantic Web; Multi-Modal Imaging; Automatic Text Recognition; Handwriting Recognition; Natural Language Processing

ABSTRACT (ENGLISH)

Naples Dante Project*. An integrated environment for bibliographical, philological and iconographical analysis.*
The paper presents the ecosystem of the Naples Dante Project (NDP), developed in PRIN 2022 and enhanced in collaboration with Changes, within Spoke 3, illustrating its digital pipeline. The focus is on: image acquisition and normalisation (IIIF, multimodal imaging); modular structuring of more abstract metadata; semantic classification of data using a dedicated ontology and exposure as Linked Open Data; automatic processing (HTR, Handwriting Identification) for palaeographic and textual analysis; the alignment of texts from the ancient manuscript tradition and modern philological editorial practice. Granular models and advanced tools enable queries on textual, linguistic and iconographic content, as well as on the material witnesses of the Comedy, integrating interoperable standards (TEI, IIIF, RDF) and imaging, transcription and collation technologies (*eScriptorium*/*Kraken*). The paper reflects on how the computational revolution is redefining philological, palaeographic and art-historical practices through verifiable pipelines that connect images, texts, metadata and automatic knowledge, with a view to long-term editorial sustainability.
Keywords: Semantic Web; Multi-Modal Imaging; Automatic Text Recognition; Handwriting Identification; Natural Language Processing

1. INTRODUZIONE

Il *Naples Dante Project* (NDP) è un progetto di ricerca scientifica dell'Università degli Studi di Napoli Federico II, registrato e finanziato dal bando PRIN 2022. Il progetto ha partecipato al partenariato esteso P5 *Changes* finanziato dal PNRR, ed è incluso nella rete dei progetti strategici del *FeDHLab*, Laboratorio di *Digital Humanities* dell'Ateneo federiciano, fondato e sviluppato nell'ambito del Dipartimento di Eccellenza. NDP nasce dalla consapevolezza che la tradizione della *Commedia* di Dante costituisce un vero e proprio caso-limite per ogni approccio computazionale. Tra il XIV e il XVII secolo essa si articola in

una tradizione vastissima e stratificata, che comprende circa seicento manoscritti, oltre centotrenta frammenti e decine di edizioni a stampa, cui si affiancano apparati esegetici e iconografici complessi e una molteplicità di metadati eterogenei. La complessità non riguarda soltanto la quantità delle testimonianze, ma soprattutto la loro natura integrata: supporto materiale, testo, lingua, illustrazioni e apparati esegetici sono inseparabili e richiedono modelli capaci di tener conto simultaneamente di tutte queste dimensioni. La sfida è integrare la filologia testuale, orientata alla dimensione genealogica e diacronica, e la filologia delle testimonianze, centrata sulla materialità e sulla sincronia dei documenti, in un ambiente cognitivo multilivellare supportato dalle *Digital Humanities*. NDP affronta tale sfida attraverso l'integrazione di strumenti e metodologie avanzate per lo studio della *Commedia* di Dante e della sua tradizione manoscritta e figurativa (Ferrante, 2024a, 2024b; Autiero et al., 2025).
L'obiettivo principale del progetto è la modellazione e la gestione di dati bibliografici, testuali e iconografici relativi al testimoniale manoscritto e a stampa della *Commedia* di Dante in un *editing framework* modulare, le cui classi sono collegate e semanticamente relazionate in un'ontologia CIDOC. NDP integra inoltre nella sua pipeline: uno strumento di visualizzazione di riproduzioni digitali da *imaging* multi-modale (dalle scansioni multispettrali all'RTI); un ambiente per la trascrizione semi-automatica del testo manoscritto, a séguito di *training* di modelli specifici di *Handwritten Text Recognition*; uno strumento di comparazione delle proposte di edizione critica della *Commedia* dal 1862 a oggi; un'analisi comparativa delle istanze grafiche dei più antichi testimoni della *Commedia*, esito di un'indagine sistematica di *Handwriting Identification*. La *vision* di NDP si sviluppa attorno a tre punti focali: supporto, testo e immagine. Per quanto riguarda il supporto, il progetto ha lanciato *eCommedia*, un catalogo e archivio digitale dell'intera tradizione manoscritta del poema (*Dante_manuscripts*), partendo dal nucleo originario il corpus del progetto start-up IDP-*Illuminated Dante Project* (Ferrante, 2019; Perna, 2020). Attualmente il progetto include anche la tradizione frammentaria della *Commedia*, attraverso il sotto-progetto FraC-*Fragments of Commedia* (FraC). Quest'ultimo affronta il frammento come entità codicologica autonoma, definendone criteri, metodi e protocolli di ricerca. *eCommedia* ha inoltre recentemente integrato la catalogazione digitale e lo studio filologico della prima tradizione a stampa della *Commedia* (XV-XVII sec.) attraverso il sotto-progetto DIP-*DanteInPrint*, che ha sviluppato anch'esso un proprio specifico protocollo di analisi e di metadatazione. Allo sviluppo dell'analisi del supporto, contribuisce anche l'integrazione nella pipeline di dati delle scansioni multispettrali e multi-luminescenti, che consente di superare il concetto di facsimile digitale, trasformando l'immagine in un vero e proprio dato analitico, leggibile congiuntamente al testo e alla materialità del supporto. Su questa base si innesta l'apertura alla paleografia computazionale, che mira all'*handwriting identification* tramite reti neurali, con l'obiettivo di verificare attribuzioni discusse e di rintracciare nuove occorrenze di tali individualità grafiche all'interno della tradizione manoscritta. Al livello del testo *eCommedia* ha come obiettivo una vera e propria *recensio* digitale, che punta alla trascrizione semi-automatica della tradizione manoscritta e a stampa, seguendo un flusso di lavoro ben definito che parte dall'*Automatic Text Recognition* e arriva alla collazione automatizzata dei testimoni passando per il *text alignment*. Le tecniche di *Deep Learning* per la *collatio* verranno affinate da un sistema di annotazione filologico-linguistica manuale integrata nel framework di *editing* modulare utilizzato (cf. *infra*, par. 3). Sempre al livello di analisi del testo, oltre alla tradizione manoscritta più antica, NDP offre a studiosi e utenti comuni *Dante Critical Texts*, un tool comparativo delle proposte di testo "scientificamente ricostruito" della *Commedia* avanzate dalla critica nei tempi moderni, a partire dall'edizione di Karl Witte (1862) fino al lavoro del gruppo di Paolo Trovato, tuttora in corso (2021- ;vd. *infra*, par. 5). Di recente il progetto ha integrato nel tool l'edizione di Federico Sanguineti (2001).
Al livello dell'immagine, NDP ha voluto estendere l'approccio analitico della metadatazione iconografica di IDP a tutta la tradizione libraria antica della *Commedia*, includendo nell'indagine anche gli esemplari a stampa illustrati tratti dal corpus di *DanteInPrint* (28 edizioni illustrate tra incunaboli e cinquecentine). A questo protocollo esteso si è aggiunto *Dante Drawings*, un sotto-progetto che include disegni in serie della *Commedia* (configurabili come album, disegni preparatori per la stampa e altre tipologie) prodotti fino al XVI secolo. A questo scopo, il team ha sviluppato un modello di metadatazione iconografica che permette un livello di analisi al contempo trasversale (supporti diversi: manoscritto, libro a stampa, serie di disegni) e estremamente granulare. Per una sua maggiore interoperabilità con standard internazionali, il modello iconografico prevede l'implementazione, attraverso l'API *OpenRefine*, del *thesaurus* del progetto *Biblissima+* che fonde in sé i descrittori iconografici del database *Mandragore* e di *Initiale* per un totale di 27.000 lemmi. Si illustreranno di seguito i vari aspetti del progetto, a partire dall'ontologia fino ad arrivare alla pipeline progettata per portare avanti il complesso flusso di lavoro.

2. ALIGHIEROO: UN'ONTOLOGIA PER L'INTERPRETAZIONE SEMANTICA DELLA TRADIZIONE DANTESCA

ALIGHIERoo (*Advanced Literary and Iconographic Graph for Hermeneutics of Interconnected Editions and Resources – object-oriented*) è l'ontologia sviluppata dal *Naples Dante Project* per rappresentare semanticamente testi, immagini e testimoni della *Commedia*. Basata su CIDOC CRM e LRMoo, formalizza in OWL le relazioni tra entità testuali, codicologiche e iconografiche, descrivendo manoscritti, frammenti, edizioni e apparati figurativi (Cantone et al., 2025). Tale ontologia prende avvio dall'ontologia OWL OntoComedySources (De Domenico et al., 2020) e integra la logica descrittiva (DL) per garantire coerenza e rendere esplicite relazioni implicite (es. ricorrenze iconografiche, identità di personaggi, tipologie codicologiche), favorendo analisi verificabili. Il modello ontologico ha raggiunto un grado di maturità che consente di delineare con chiarezza le sue traiettorie di consolidamento e di espansione. Le prospettive di sviluppo individuate non rispondono a logiche di mera estensione semantica, ma intendono rafforzare il valore epistemologico e operativo del modello, assicurandone la durabilità, la trasparenza e l'utilità per la comunità scientifica. *ALIGHIERoo* costituisce dunque una infrastruttura semantica stabile e collaborativa, capace di connettere la tradizione di studi filologici con le metodologie digitali più avanzate, nonché di offrire alla comunità un modello aperto, rigoroso e continuamente perfezionabile in ottica bottom-up.
L'attuale versione del modello integra già solide relazioni tra entità testuali e materiali, ma l'esperienza d'uso ha mostrato l'opportunità di formalizzare ulteriormente i casi di paratesto quali commento, glosse o istruzioni ancillari (come le note al miniatore o quelle al copista, di carattere pressoché inedito). Tali fenomeni costituiscono una dimensione essenziale sia della tradizione dantesca che, più in generale, della cultura testuale medievale, poiché rivelano la presenza di una paratestualità operativa che incide direttamente sulla forma materiale e visiva dell'opera. A ciò si aggiunge la necessità di sviluppare strumenti di *mapping* e validazione automatica, capaci di verificare, oltre la consistenza semantica tra *ALIGHIERoo* e altri sistemi del NDP (come *Cadmus*), quelle di ontologie affini (CIDOC-CRM, LRMoo, LIDO). Tali strumenti, già in parte progettati, costituiranno un elemento essenziale per il controllo di qualità e per l'interoperabilità secondo i principi FAIR, permettendo una connessione semantica più ampia senza che si comprometta l'uso delle funzionalità di *reasoning*.

3. LA METADATAZIONE DI NDP IN CADMUS

La principale sfida di modellare le conoscenze rappresentate in questo progetto sta nell'accordare l'esigenza di alta granularità e strutturazione dei dati, peraltro estremamente eterogenei, con la facilità nella loro creazione, la sostenibilità nel tempo e la flessibilità nel riuso comune di modelli, metodi e strumenti integrati in un'unica architettura.
Al centro della soluzione adottata è un sistema (*Cadmus*) disegnato proprio per soddisfare simili requisiti, tramite un'architettura ad oggetti modulare, sia a livello di modellazione dei dati che del software utilizzato per crearli: la modularità è infatti la chiave per definire modelli dinamici per via di composizione, come una costruzione assemblata con gli stessi mattoni per edifici sempre diversi, consentendo un sistematico riuso a livello di modelli e di software di editing (Fusi, 2018; Bessi & Fusi, 2024). Questo assicura non solo economia e sostenibilità, ma anche un sistema dinamico e scalabile, capace di evolvere nel tempo senza pregiudizio per i dati già esistenti, integrando dati completamente eterogenei in un'architettura uniforme, che ne promuove una profonda integrazione; e per conseguenza induce a un approccio collaborativo, dove ogni progetto contribuisce alla creazione di un catalogo di modelli e moduli software per la loro creazione, di volta in volta riassemblati in nuove costruzioni. Di fatto, anche in questo progetto intere famiglie di modelli elaborate da altri progetti sono state riusate, e in taluni casi espanse e migliorate, come nel caso dei modelli codicologici definiti per il progetto *Itinera* (De Nisco et al., 2021; Gorla et al., 2022).
Di qui una modellazione dei dati libera e realizzata di concerto con gli esperti di dominio, non influenzata da specifici vincoli tecnologici imposti a priori, e il più possibile astratta e generalizzabile, a garanzia della sua estensibilità, ma nel contempo capace di comprendere un elevato livello di dettaglio, rappresentando la complessità per via di aggregazione di elementi più semplici.
Questa stessa architettura nell'intero sistema consente poi di aprirlo all'integrazione di risorse di terza parte, come ontologie, liste di autorità, servizi o componenti software esterni, o anche puri dati di esportazione o di recupero da fonti digitali *legacy* o pre-digitali, all'interno di un vero e proprio hub di editing dove convergono contenuti eterogenei, modellati in un'architettura aperta e tuttavia uniforme. I

dati creati o incorporati possono poi essere a loro volta oggetto di esportazione e integrazione in altri contesti, per via di sistemi di esportazione disegnati secondo lo stesso obiettivo di riuso, da regole di proiezione semantica per generare grafi RDF a moduli di esportazione in varie implementazioni TEI o in altri tipi di formati, come pure in database e servizi software utilizzabili per via di API o direttamente incorporabili in sistemi di terza parte. Tra le importanti sinergie che sono state raggiunte in questa prospettiva da NDP si contano quelle con *Biblissima+* che ha fornito l'accesso all'API del wikibase e *Manus Online* (MOL) e *Medieval Unicode Font Initiative* (MUFI) che hanno fornito gli *authority files* successivamente integrati localmente in *Cadmus.*

4. L'ARMONIZZAZIONE DELLE STRUTTURE DATI TRA ILLUMINATED DANTE PROJECT, CADMUS E MANUS ONLINE

Un passaggio fondamentale è quello dell'armonizzazione tra i modelli di *Cadmus* e l'ontologia, così da creare una coerenza semantica all'interno dell'ecosistema di NDP. Il sottosistema di mappatura semantica già presente in *Cadmus* consente di costruire un intero grafo per via di regole di proiezione che estraggono specifici dati dagli oggetti del database per rimodellarli in nodi e collegamenti di un grafo di destinazione, qui modellato secondo *ALIGHIERoo*. Per immettere nel nuovo sistema semantico i dati relativi alle schede di catalogo di *Illuminated Dante Project* (*IDP*) è stato necessario un sistematico confronto tra la struttura del database di IDP e la modellizzazione di *Cadmus*. La parte del framework predisposta ad importare il contenuto di IDP è *Codicology* che, come le altre parti di *Cadmus*, permette una descrizione molto granulare dell'oggetto da catalogare, e nel caso di specie i dati contenuti nei campi delle tabelle di IDP possono essere razionalizzati in moduli concettuali di dimensione molto inferiore. Poiché dunque non c'è corrispondenza biunivoca tra i campi del database di IDP e i moduli di *Cadmus* si è proceduto alla definizione di procedure standardizzate per reinterpretare i dati ed è stata predisposta una mappatura concettuale tra IDP e *Cadmus*. A valle di questa mappatura si è verificata la compatibilità di Codicology con il modulo di descrizione codicologica della *Text Encoding Initiative* (TEI), ossia <msDesc>. In tal modo sarà possibile mantenere l'interoperabilità con *Manus Online* dove sono già presenti le schede dei manoscritti di IDP.

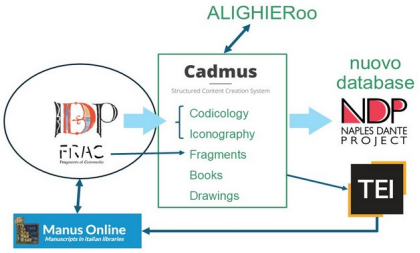


Figura 1. Pipeline di armonizzazione e mappatura IDP - NDP (Cadmus) – MOL (TEI).

5. DANTE CRITICAL TEXTS: UN TOOL DI COMPARAZIONE DELLE EDIZIONI CRITICHE DELLA COMMEDIA DI DANTE

Parallelamente al lavoro di catalogazione e metadatazione dei manoscritti, è in corso il progetto *Dante Critical Texts* (DCT), che prevede la digitalizzazione di tutte le edizioni critiche elaborate per la Commedia dantesca, a partire dall'edizione di Karl Witte (1862) fino a quelle più recenti pubblicate per il settecentenario dantesco del 2021 (Inglese, Lanza, Malato, Tonello-Trovato). Tali digitalizzazioni seguono il modello dell'electronic mirror image, ossia dell'immagine speculare elettronica del testo a stampa (Brunner, 1991, p. 63; Tissoni, 2015, §§ 29–33), costituendo di fatto delle edizioni critiche "digitalizzate", che potranno in futuro essere integrate nell'ambito di un nuovo lavorio ecdotico nativamente digitale (Magnani, 2018, pp. 86–90; La Veglia, 2025, p. 59). Rispetto ad altri grandi progetti di digitalizzazione come Perseus e il ThLG che forniscono l'immagine speculare elettronica di una e una sola edizione a stampa, in DCT si mettono in dialogo diverse proiezioni testuali, che vanno a costituire una "edizione di edizioni" che dà il senso della profondità filologica del testo. Già in questa prima fase sono state digitalizzate le edizioni di Vandelli (1921), Petrocchi (1966), Sanguineti (2001), Malato (2021), Trovato-Tonello (2022) e dopo esser stato predisposto l'allineamento dei testi delle edizioni è stata avviata una

collazione automatica tramite un software diff[1]. Le diverse edizioni allineate sono visualizzabili sul sito dedicato al progetto [2]. La digitalizzazione delle edizioni è stata eseguita con parziale supporto dell'OCR, in un meccanismo semiautomatico, ma in futuro sarà implementato un *workflow* quasi completamente automatico per la digitalizzazione che integrerà OCR e *text alignment*. Due sono i punti di forza di tale progetto: il primo è quello di rendere liberamente disponibili online i prodotti dei lavori filologici più recenti, di norma rinchiusi tra le pagine di costose edizioni a stampa dall'intoccabile diritto d'editore (La Veglia, 2025) e il secondo consiste nel conservare il valore scientifico della variante testuale, che i primissimi lavori di digitalizzazione avevano trascurato, creando delle "nuove vulgate informatiche" (Varvaro, 1999, p. 57; Palumbo, 2015, p. 147).

6. LA NECESSITÀ DI UNA PIPELINE INTEGRATA DELLE RISORSE: L'APPLICAZIONE DI MAPP A NDP

A tale stato di avanzamento del progetto, si è resa necessaria l'elaborazione di una pipeline digitale integrata in grado di coprire tutte le fasi del ciclo di trattamento digitale delle risorse manoscritte.
Nell'ambito del partenariato esteso CHANGES, NDP ha approntato in collaborazione con DedaNext, la piattaforma *Manuscripts Philological Pipeline* (MAPP). Tale piattaforma realizza una pipeline modulare e interoperabile che copre l'intero ciclo di vita dei documenti – dall'acquisizione digitale e dalla metadatazione descrittiva alla trascrizione automatica, all'integrazione di immagini multispettrali e alla pubblicazione online – offrendo strumenti di ricerca sia tradizionale sia semantica. Fondata su tecnologie open source e conformi ai principi FAIR, la piattaforma integra *Cadmus* per la metadatazione, *eScriptorium* per la segmentazione del testo, l'allineamento testo-immagine e la trascrizione automatica e *Virtuoso* per la gestione delle triple RDF dell'ontologia in un unico sistema che ha come cuore pulsante il framework *MetaFAD*. A tali tasselli si aggiungono dei visualizzatori avanzati che supportano anche immagini multispettrali.
MaPP mette a disposizione una funzionalità di interrogazione SPARQL che consente agli utenti di formulare query volte a interrogare le relazioni semantiche tra le entità del grafo, permettendo così interrogazioni complesse su più livelli. Per agevolarne l'utilizzo anche da parte di un pubblico non tecnico, sono inoltre disponibili alcune query predefinite, che l'utente può eseguire direttamente.
Virtuoso consente di selezionare diversi formati di output per la restituzione dei risultati, adattandosi alle esigenze dei diversi utenti. Sono quindi disponibili sia formati di visualizzazione orientati alla navigazione esplorativa, sia output più tabellari e strutturati, idonei ad analisi più formali.
Lo strumento Esplora Triple, inoltre, offre la possibilità di navigare liberamente e in maniera trasversale l'intero patrimonio codicologico, iconografico e testuale presente nell'Infrastruttura, permettendo agli studiosi di esplorare i collegamenti tra le diverse entità e di fornire supporto alle loro ricerche scientifiche.
Il *Naples Dante Project* costituisce la prima applicazione concreta degli esiti del progetto MAPP, adottandone la pipeline per la gestione e l'esplorazione integrata dei dati testuali, iconografici e materiali relativi alla Commedia dantesca.

7. CONSIDERAZIONI CONCLUSIVE: L'ORIZZONTE EPISTEMOLOGICO DI NDP

Le strategie di analisi computazionale del *Naples Dante Project* partono dal presupposto di una concezione integrata della filologia: supporto, testo, lingua e immagine sono inseparabili. L'ambiente digitale permette uno studio filologico multidisciplinare nel quale possono dialogare critica testuale, filologia materiale, paleografia, codicologia, linguistica storica e iconografia, superando l'impostazione delle filologie "a compartimenti stagni". Nel complesso ecosistema che si sta costituendo nell'ambito del progetto è possibile realizzare una rappresentazione multidimensionale del testo e dei testimoni, attraverso tutti i cosiddetti "livelli di rappresentazione" (Tarte, 2011; Monella, 2018; Reggiani, 2025, pp. 10–18, 40–41). In questa prospettiva lo spazio digitale diventa un ambiente cognitivo multilivellare che permette di integrare in modo virtuoso la filologia del testo e la filologia delle testimonianze e favorire un dialogo con la comunità scientifica tramite l'interoperabilità dei dati e la connessione con le infrastrutture di ricerca. Gli strumenti di analisi computazionale dei dati rappresentano non solo un supporto pratico alla ricerca ma diventano anche un'occasione di riflessione teorica e metodologica sui paradigmi interpretativi

[1] Cf. "diff" in *The Open Group Base Specifications Issue 8*, IEEE Std 1003.1-2024 https://pubs.opengroup.org/onlinepubs/9799919799/ (cons. 08/02/2026).
[2] https://www.dante.unina.it/dct/public/pagine/progetto (cons. 08/02/2026).

dell'oggetto di studio filologico. Inoltre, gli sforzi di modellizzazione profusi nei diversi ambiti del progetto hanno portato a risultati scientifici che possono essere esportabili anche all'infuori degli studi danteschi.

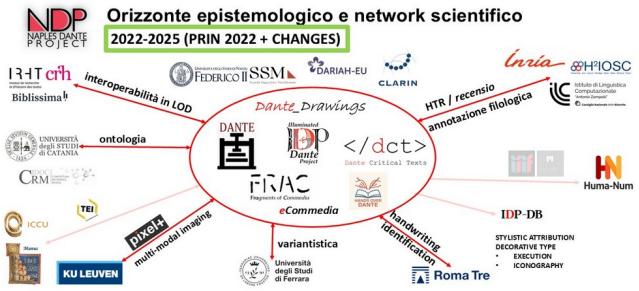


Figura 2. Gli obiettivi prossimi e futuri di NDP e le sinergie in campo

RINGRAZIAMENTI

Il progetto è finanziato da: PRIN 2022 - NAPLES DANTE PROJECT; Spoke 3 del PNRR PE 5 Changes - Digital Libraries, Archives and Philology; Dipartimento di Eccellenza 2023-2027, Dipartimento di Studi Umanistici, Università degli Studi di Napoli Federico II.

BIBLIOGRAFIA

Autiero, F., Celotto, V., Ferrante, G., Fusco, C., Gorla, S., Liberti, G. A., Palomba, M., Picarelli, S., Rizzo, S. A., & Tripodi, S. (2025). Verso la svolta computazionale della critica dantesca. In S. Rebora, M. Rospocher, & S. Bazzaco (A c. Di), Diversità, Equità e Inclusione: Sfide e Opportunità per l'Informatica Umanistica nell'Era dell'Intelligenza Artificiale, Proceedings del XIV Convegno Annuale AIUCD2025. AIUCD. https://doi.org/10.6092/unibo/amsacta/8380

Bessi, B., & Fusi, D. (2024). Modelling the Archipelago: Corfu as a Case Study for a Digital Edition of Cristoforo Buondelmonti's Liber Insularum. Open Research Europe, 4, 11. https://doi.org/10.12688/openreseurope.16712.1

Brunner, T. F. (1991). The Thesaurus Linguae Graecae: Classics and the computer. Library Hi Tech. https://www.semanticscholar.org/paper/The-Thesaurus-Linguae-Graecae%3A-classics-and-the-Brunner/9e310f3ad722a09b85fe3229da8d46e73fcf6ff4#citing-papers

Cantone, D., Nicolosi Asmundo, M., & Rizzo, S. A. (2025). ALIGHIERoo: Advanced Literary and Iconographic Graph for Hermeneutics of Interconnected Editions and Resources - Object Oriented. In D. Guidotti, L. Pandolfo, & L. Pulina (A c. Di), Proceedings of the 40th Italian Conference on Computational Logic, Alghero, Italy, June 25-27, 2025 (Vol. 4003). CEUR-WS.org. https://ceur-ws.org/Vol-4003/short23.pdf

De Domenico, A., Cantone, D., & Nicolosi Asmundo, M. (2020). An Ontology for the Iconographic Sources of La Divina Commedia. In F. Calimeri, S. Perri, & E. Zumpano (A c. Di), Proceedings of the 35th Italian Conference on Computational Logic—CILC 2020, Rende, Italy, October 13-15, 2020 (Vol. 2710, pp. 157–168). CEUR-WS.org. https://ceur-ws.org/Vol-2710/paper10.pdf

De Nisco, N., Gorla, S., & Valenti, A. (2021). Una banca dati per Petrarca e il suo tempo: Criteri, modelli e obiettivi. DigItalia, 16(2), Articolo 2. https://doi.org/10.36181/digitalia-00037

Ferrante, G. (2019). "Illuminated Dante Project". Per un archivio digitale delle piú antiche illustrazioni della 'Commedia'. Parte i. Un "case study" quattrocentesco (mss. Italien 74, Riccardiano 1004, Guarneriano 200). In M. Ciccuto & L. M. G. Livraghi, Dante visualizzato: Carte ridenti, II : XV secolo, prima parte. Franco Cesati editore.

Ferrante, G. (A c. Di). (2024a). La Commedia di Dante nell'antica tradizione manoscritta: Per uno studio integrato e computazionale dei codici, dei testi e delle immagini. Editrice Antenore.

Ferrante, G. (2024b). Verso il Naples Dante Project. Recensio digitale integrale dei codici della Commedia di Dante. In R. Bardi & L. Canova (A c. Di), «A guisa d'uom che 'n dubbio si raccerta». Vecchie e nuove prospettive per la biografia e l'opera dantesca (pp. 145–160). Cesati. https://www.researchgate.net/publication/386506607_Verso_il_Naples_Dante_Project

Ferretti Cuomo, L., Tonello, E., & Trovato, P. (2022). Dante Alighieri. Commedia. Inferno. Libreriauniversitaria.it.

Fusi, D. (2018). Sailing for a Second Navigation: Paradigms in Producing Digital Content. Seminari Romani Di Cultura Greca : N.s. VII, 2018, 213–276.

Giuseppe Vandelli (1865-1937) (A c. Di). (1921). La Divina Commedia. R. Bemporad e figlio editori. http://archive.org/details/LaDivinaCommediaTestoCriticoACuraDiGiuseppeVandelli

Gorla, S., Picarelli, S., Ferrante, G. (Rino), & Autiero, F. (2022). Dante e Petrarca allo (stesso) scrittoio. Per lo sviluppo di un'ontologia di IDP a partire dall'istanza manoscritti di Itinera. AIUCD. https://www.academia.edu/69019066/Dante_e_Petrarca_allo_stesso_scrittoio_Per_lo_sviluppo_di_un_ontologia_di_IDP_a_partire_dall_istanza_manoscritti_di_Itinera

Karl, W. (1862). Dante Alighieri. La Divina commedia. R. Decker.

La Veglia, A. (2025). Being a Classicist in the Digital Age: New Challenges and Cultural Paradigms Between Copyright Issues and Open Access. In N. Reggiani (A c. Di), Digital Papyrology III: The Digital Critical Edition of Greek Papyri: Issues, Projects, and Perspectives (pp. 49–70). De Gruyter. https://doi.org/10.1515/9783111070162-004

Magnani, M. (2018). The Other Side of the River Digital Editions of Ancient Greek Texts Involving Papyrus Witnesses. In N. Reggiani (A c. Di), Digital Papyrology II (pp. 87–102). De Gruyter. https://doi.org/10.1515/9783110547450-005

Malato, E. (A c. Di). (2021). La divina commedia: 2. Inferno. Salerno editrice.

Monella, P. (2018). Livelli di rappresentazione del testo nell'edizione del De nomine di Orso Beneventano. Umanistica Digitale, 2, Articolo 2. https://doi.org/10.6092/issn.2532-8816/7286

Palumbo, G. (2015). Alberto Varvaro e l'ecdotica: Per un glossario antologico. Ecdotica, (1/2015). https://doi.org/10.7385/99229

Perna, C. (2020). IDP - Illuminated Dante Project: Un archivio e database per la più antica iconografia dantesca (secc. XIV-XV). DigItalia, 15(2), 150–158. https://doi.org/10.36181/digitalia-00022

Petrocchi, G. (A c. Di). (1966). Dante Alighieri. La Commedia, secondo l'antica vulgata (Vols. 1–4). Mondadori.

Reggiani, N. (A c. Di). (2025). The digital critical edition of Greek papyri: Issues, projects, and perspectives. De Gruyter.

Sanguineti, F. (A c. Di). (2001). Dantis Alagherii Comedia. SISMEL-Ed. del Galluzzo.

Tarte, S. M. (2011). Digital Visual Representations in Papyrology: Implications on the Nature of Digital Artefacts. https://www.academia.edu/776645/Digital_Visual_Representations_in_Papyrology_Implications_on_the_Nature_of_Digital_Artefacts.

Tissoni, F. (2015). Pour un corpus numérique comparatiste des traductions d'Homère. Corpus Eve, (2). https://doi.org/10.4000/eve.1273

Varvaro, A. (1999). The «New Philology» from an Italian Perspective (M. Cherchi, Trad.). Text, 12, 49–58.

La curatela biblioteconomica delle raccolte digitali MAB: linee gestionali, requisiti, workflow

NICOLA BARBUTI

DIPARTIMENTO DI RICERCA E INNOVAZIONE UMANISTICA DIRIUM – UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO, ITALIA – NICOLA.BARBUTI@UNIBA.IT

ABSTRACT (ITALIANO)

Le collezioni digitali sono una componente crescente del patrimonio culturale e scientifico. Nei domini MAB (Musei, Archivi, Biblioteche), richiedono il ripensamento di categorie descrittive e di metodi e pratiche di gestione, ancora oggi modellate sull'analogico o sulla digitizzazione intesa come semplice "travaso di un oggetto fisico in formato digitale". Nelle raccolte che derivano da beni fisici, la qualità documentaria dipende dalla tenuta del legame con l'originale, con la catena di acquisizione e con le condizioni di riuso; nelle risorse born-digital collegate alla tutela e alla valorizzazione del patrimonio, identità e autenticità dipendono da formati, profili di codifica, dipendenze software/hardware, versioni, provenienza e livelli di accesso. Il paper propone una riflessione sulla curatela di collezioni digitali MAB, delineandone un workflow misurabile che include selezione, acquisizione, ingest, descrizione, accesso, preservazione attiva, e prevede metadati minimi per tracciabilità e verificabilità, nonché la separazione tra master preservati e derivati di fruizione e riuso. Discute, inoltre, sinteticamente un modello di uso dell'IA basato su versionamento degli output e controllo human-in-the-loop, con indicatori essenziali di sostenibilità e responsabilità.
Parole chiave: born-digital; MAB; curatela digitale; preservazione; intelligenza artificiale.

ABSTRACT (ENGLISH)

Paper Title for AIUCD2026: Curation of LAM Digital Collections; Management, Requirements, and Workflows

Digital collections are an increasing share of cultural and scholarly heritage. Within the GLAM domain, they require a reconsideration of descriptive and management practices still rooted in analogue holdings, or in digitisation understood as a simple transfer. In collections of digitized artefacts, documentary quality encloses the persistent link with the original item, the capture chain, and the conditions of reuse; in born-digital resources connected with the care and enhancement of heritage, identity and authenticity rely on formats, encoding profiles, software and hardware dependencies, versions, provenance, and access levels. This paper proposes a reflection on the curation of GLAM digital collections, grounded in measurable workflow that includes appraisal, acquisition, ingest, description, access, active preservation, a basic metadata set for traceability and verifiability, and the separation between preserved masters and access derivatives. Also an AI model based on output versioning, human-in-the-loop control, and essential indicators of sustainability and accountability is briefly discussed.
Keywords: born-digital; GLAM; digital curation; preservation; artificial intelligence.

1. STATO DELL'ARTE

La riflessione sulla curatela delle risorse e delle collezioni digitali nei domini MAB è concentrata su una distinzione che non è soltanto terminologica, ma metodologica e operativa. Le risorse digitizzate mantengono un rapporto costitutivo con originali materiali e ne incorporano le proprietà documentarie: provenienza, ordinamento, responsabilità descrittiva; le risorse born-digital, invece, possono coincidere con schede, inventari, apparati critici, manifest, dataset di lavoro, ambienti di pubblicazione, software. In entrambi i casi la curatela non riguarda soltanto il file, ma il sistema di relazioni che ne rende comprensibile l'identità. Una riproduzione digitale di un manoscritto, ad esempio, resta scientificamente utilizzabile se conserva il nesso con gli elementi di classificazione, anche digitale, che ne definiscono l'esistenza (segnatura, sequenza, parametri di acquisizione e condizioni di accesso); allo stesso modo, una scheda di catalogo nativa digitale o un manifest IIIF hanno valore se riconducibili alla storia dei loro aggiornamenti, alle responsabilità redazionali e al contesto istituzionale che li ha prodotti. Per questo, la riflessione scientifica su curatela e preservazione digitale sta progressivamente spostando l'attenzione

dalla semplice custodia dei file a un insieme di azioni orientate al rischio, alla fissità e alla documentazione delle trasformazioni, rendendo esplicita l'esigenza di progettare la preservazione come requisito "by design" lungo l'intero ciclo di vita di risorse e collezioni digitali (Digital Preservation Coalition, 2008; Kowalczyk, 2008). Nei contesti MAB, ciò significa assumere sin dall'inizio che acquisizione, descrizione, pubblicazione e preservazione non siano fasi isolate, ma parti di una medesima architettura documentaria. Gli studi dedicati alla gestione e alla preservazione di collezioni digitali mostrano che la scalabilità si ottiene solo integrando selezione, descrizione, preservazione e accesso in un workflow coerente, superando la separazione tra trattamento descrittivo e gestione tecnica e riconoscendo che l'*unità di cura* può coincidere con un insieme relazionale di oggetti digitali, metadati, ambienti, dipendenze, non necessariamente con il singolo file (Carroll et al., 2011; Pacheco et al., 2023). Questo approccio è cruciale tanto per materiali digitizzati, nei quali la catena di produzione (acquisizione, controlli qualità, derivazione dei file, pubblicazione) e le proprietà significative dell'originale rappresentato devono restare ricostruibili ed essere preservate, quanto per materiali born-digital "di contesto", quali software, opere interattive, videogiochi, o per record di progettazione digitale in cui repository di codici, librerie e servizi esterni concorrono a comporre e a definire le informazioni che, nel tempo, devono restare intellegibili e verificabili (Leventhal & Thompson, 2021). Sul versante organizzativo, manuali e best practice hanno stabilizzato la centralità di policy, piani di preservazione, ruoli e processi di controllo qualità, così che autenticità e affidabilità risultino sostenute da evidenze e non da mere dichiarazioni. Nei progetti di digitizzazione, ciò si traduce nella necessità di definire criteri minimi condivisi (ad esempio profili di acquisizione, convenzioni per master e derivati, livelli di accesso, tracciabilità degli interventi) e di adottare modelli "a pacchetti", in cui le risorse digitali circolino insieme a metadati tecnici, diritti e registri di eventi che descrivono ingest, trasformazioni e verifiche. In questa cornice si collocano i profili applicativi che governano la pluralità degli standard, rendendo le mappature esplicite e validabili e facilitando la cooperazione tra biblioteche, archivi e musei, soprattutto quando i sistemi sono federati o distribuiti su più istituzioni. Un fronte importante di rinnovamento della gestione biblioteconomica riguarda l'impiego dell'Intelligenza Artificiale (IA) a supporto della curatela delle collezioni digitali: sistemi di riconoscimento digitale (OCR/HTR) e di Natural Language Processing (NLP) migliorano trascrizione, estrazione, indicizzazione e riconciliazione di grandi volumi di pagine e schede, ma allo stesso tempo richiedono protocolli di valutazione replicabili e la distinzione netta tra dato osservato e output inferito (Karatzas et al., 2011). I modelli generativi, inoltre, spostano l'attenzione sulla governance dell'accesso e dell'interpretazione, perché l'interfaccia conversazionale può produrre sintesi non immediatamente verificabili; ne consegue la necessità di log, citazione delle fonti, tracciabilità degli output e controllo umano, insieme a una gestione esplicita dei rischi che incidono sulla fiducia pubblica (Wach et al., 2023). A livello europeo, recenti agende di ricerca per il digital heritage sottolineano la necessità di coniugare sperimentazione, sostenibilità e responsabilità, includendo qualità dei dati, trasparenza e valutazione degli impatti come condizioni per un'innovazione pubblica e riusabile (Münster et al., 2024). Nel complesso, le agende convergono su una prospettiva dei MAB che li identifica in infrastrutture di affidabilità: istituti capaci di produrre evidenze tecniche e descrittive che consentano, nel lungo periodo, di controllare le trasformazioni, rendere riproducibili le procedure e sostenere l'intellegibilità e la verificabilità delle interpretazioni, anche quando strumenti e interfacce di accesso cambiano.

2. PROBLEMI, OBIETTIVI E SPECIFICITÀ DEI DOMINI MAB

La complessità della curatela digitale nei domini MAB emerge con particolare evidenza quando le risorse digitali sono relative a beni analogici "non convenzionali" – manoscritti, libri antichi e moderni, fotografie, opere d'arte, oggetti museali, documenti d'archivio – e ne sono la controparte virtuale per preservazione, studio, accesso e interazione (Barbuti, 2025). In questo quadro si sovrappongono tre vincoli, che nell'analogico risultano più facilmente distinguibili: il vincolo tecnico di leggibilità, il vincolo giuridico di accesso e il vincolo epistemico di contestualizzazione. Sul piano tecnico, anche quando le risorse derivano da processi di digitizzazione – a es., immagini ad alta risoluzione, manifest IIIF, file audio/video – l'intellegibilità dipende da formati, profili di codifica e software di visualizzazione, nonché dalla disponibilità di metadati tecnici e di controllo qualità; la perdita di informazioni di cattura su parametri di calibrazione, profili colore, risoluzione, compressione, o di riferimenti strutturali su sequenza, paginazione, unità documentali, può produrre un degrado informativo pur in presenza di un bitstream integro. Per esempio, un'immagine ad alta risoluzione di un codice miniato può risultare integra sul piano del bitstream, ma perdere valore documentario se non conserva parametri di acquisizione, sequenza e

collocazione; analogamente, le riproduzioni di fotografie d'autore o le scheda digitali di restauro possono richiedere livelli di accesso differenziati per ragioni di diritti, riservatezza o tutela. Sul piano giuridico, la connessione con l'analogico non elimina, anzi, spesso accentua la necessità di distinguere tra proprietà materiale e diritti di riproduzione e riuso, nonché di gestire eventuali dati personali presenti nei documenti o nelle descrizioni; ne conseguono livelli di accesso differenziati (open, mediato, ristretto) e policy esplicite per embargo, riuso e citazione. Sul piano epistemico, infine, la risorsa digitale non può essere interpretata come "entità granulare autonoma": essa deve essere ancorata a provenienza, contesto di produzione e, ove esistenti, legami con l'oggetto fisico e con altre entità (collocazione, segnature, serie, descrizioni di dominio), perché senza tali relazioni resterebbe muta, poco interrogabile e difficilmente verificabile (Tammaro et al., 2012; Tammaro, 2016). Il problema operativo, per istituti spesso limitati in risorse, consiste nel trasformare questa complessità in scelte replicabili e verificabili, evitando la formazione di depositi di file non governati e di collezioni digitali che, pur esistendo, non risultano realmente curabili né fruibili in modo scientificamente fondato. Una tassonomia orientata alle decisioni che sia utile alla curatela di risorse connesse ai patrimoni analogici dei MAB può essere ricondotta a cinque classi ricorrenti: 1) collezioni generate da originali fisici che includono sia master di conservazione, che derivati per l'accesso e la fruizione: risorse digitali riproducenti manoscritti, stampe, fondi fotografici, oggetti museali, opere d'arte, ecc.; 2) documentazione born-digital prodotta dalle attività di tutela e valorizzazione del patrimonio analogico, che costituisce parte integrante dell'evidenza istituzionale e della ricostruibilità delle interpretazioni – a es., schede di catalogo, inventari, relazioni conservative, apparati critici, apparati di mostra, didascalie e percorsi digitali –; 3) oggetti born-digital generati dai workflow di digitizzazione e di pubblicazione, da trattare come derivati versionati e controllabili: pacchetti METS/MODS, metadati PREMIS, manifest IIIF, trascrizioni OCR/HTR, annotazioni e segmentazioni; 4) contenuti digitali di mediazione culturale collegati alle collezioni fisiche, che richiedono regole di attribuzione, qualità e riuso coerenti con le fonti e con i livelli di accesso: a es., mostre e percorsi online, materiali educativi, campagne e contributi partecipativi connessi a oggetti specifici; 5) archivi virtuali personali e comunitari ad alta sensibilità, quali email, messaggistica, web/social, ecc. La definizione di un profilo applicativo per le risorse riconducibili ai domini MAB consentirebbe di rendere comparabili le decisioni di curatela attraverso invarianti e criteri espliciti, evitando che differenze organizzative si traducano in difformità non motivate di qualità o di accessibilità. In questa prospettiva, la fase di selezione può essere modellata esplicitando le finalità, quali valore storico, scientifico, probatorio e di valorizzazione; livello di servizio atteso: a es., open, mediato, ristretto, consultazione in sede, da remoto autenticato; vincoli: diritti, dati personali, condizioni di riproduzione; strategia tecnica, focalizzandola su normalizzazione dei formati, politiche di derivazione, migrazioni pianificate. La separazione tra master di preservazione e derivati di fruizione rappresenta qui un principio strutturale, perché permette di mantenere invariate le evidenze fondamentali e, al contempo, di aggiornare nel tempo strumenti di accesso, formati di rilascio e interfacce, quali a esempio visualizzatori IIIF, portali, API, senza riscrivere "silenziosamente" l'oggetto preservato. In tale cornice, autenticità, integrità e verificabilità assumono la funzione di qualità progettuali che orientano l'architettura del repository, le procedure di ingest, la descrizione e le scelte di accesso, rendendo praticabile l'adozione di automazioni, inclusa l'IA, come supporto alla discovery e alla descrizione, senza indebolire responsabilità e ricostruibilità nel lungo periodo.

3. METODOLOGIA DI CURATELA E WORKFLOW

Dal punto di vista organizzativo, la curatela delle risorse digitali evidenzia i limiti di una gestione "a compartimenti". La presa in carico richiede un raccordo stabile tra competenze biblioteconomiche e archivistiche per selezione, descrizione, autorità, policy di accesso, competenze tecnico-informatiche per formati, qualità della riproduzione, controlli di integrità, sicurezza dei trasferimenti, ambienti di pubblicazione, e competenze legali per licenze, privacy, riuso. Nei contesti di MAB, tali competenze risultano spesso distribuite tra istituti diversi; ne consegue l'esigenza di cooperazione e standardizzazione che riguarda non solo i metadati, ma anche i processi, i livelli di servizio e le soglie di qualità accettabili, quali, a esempio, i requisiti minimi di acquisizione e i criteri di pubblicazione. Un ulteriore vincolo, frequente nei progetti di digitizzazione e valorizzazione, è la presenza di ecosistemi applicativi e piattaforme di pubblicazione: in questi casi, la selezione deve considerare la portabilità effettiva, distinguendo tra acquisizione dei master, acquisizione della struttura di fruizione e acquisizione delle evidenze di produzione e trattamento. Queste esigenze e vincoli rendono necessario un workflow

leggibile, documentabile e comparabile. Una metodologia di curatela applicabile a collezioni digitizzate e a oggetti born-digital collegati al patrimonio analogico può essere descritta come un ciclo di vita misurabile, nel quale ogni fase produce evidenze, decisioni e controlli suscettibili di verifica. La logica di questo workflow è rappresentata nella Figura 1, che mostra la continuità tra selezione, acquisizione, ingest, descrizione, accesso e preservazione attiva, insieme ai principali punti di controllo.



Figura 1. Schema sintetico del ciclo di vita della curatela per collezioni digitali MAB: fasi, evidenze e controlli.

La selezione è concepita come fase documentaria: criteri di rilevanza, completezza e provenienza vengono registrati insieme a un profilo di sensibilità – diritti e dati personali presenti nei contenuti o nelle descrizioni – e a un profilo di complessità tecnica – varietà di formati, requisiti di qualità della riproduzione, necessità di OCR/HTR, dipendenze da viewer e protocolli di accesso – così da poter prevedere costi e rischi di trattamento. In una collezione di carteggi, per esempio, questa fase può distinguere tra immagini master, metadati descrittivi, trascrizioni e allegati sensibili, assegnando fin dall'inizio livelli di servizio differenti. A fini comparativi, questi insiemi possono essere espressi come classi bassa, media, alta, e associati a un razionale motivato. L'acquisizione è definita come trasferimento controllato dai laboratori di digitizzazione o dai sistemi gestionali di cataloghi, inventari e repository: si registrano fonte, modalità di esportazione o conferimento, intervallo temporale, strumenti e responsabilità; si calcolano checksum iniziali e si attivano controlli di sicurezza prima dell'ingest, preservando evidenze su eventuali anomalie. L'ingest nel repository è governato dall'organizzazione logica (packaging): identificazione dei formati e verifica di conformità ai profili adottati, estrazione di metadati tecnici e strutturali, controlli di corrispondenza tra oggetto fisico, unità descrittiva e file digitali, produzione di pacchetti che distinguano master preservati e derivati di accesso; le trasformazioni, quali normalizzazione, generazione di derivati, OCR/HTR, creazione di manifest e visualizzatori, sono registrate come eventi con strumenti, versioni e parametri. In un progetto di digitalizzazione di registri parrocchiali o di un fondo fotografico, la possibilità di ricondurre ogni derivato ai file master, ai controlli qualità e alla sequenza dell'unità descrittiva impedisce che la pubblicazione web produca una perdita silenziosa di contesto. La descrizione, dunque, va impostata come contestualizzazione relazionale. Nei beni MAB la leggibilità dipende da collocazioni, segnature, serie e legami semantici; perciò, si integrano dati di autorità e vocabolari controllati con note di produzione, diritti e restrizioni, riconoscendo i metadati descrittivi come parte della preservazione e non solo come strumento di identificazione di dati. L'accesso è progettato per livelli – aperto, mediato o ristretto – e con soluzioni coerenti con la natura delle risorse. La fruizione avviene tramite derivati ottimizzati, quali immagini compresse, stream audio/video, pagine web, e, quando opportuno, tramite visualizzatori standardizzati, a esempio IIIF, mantenendo sempre esplicito il rapporto con il master: un percorso espositivo digitale che combina immagini di opere, didascalie nate digitalmente e documentazione di restauro resta affidabile solo se il lettore può distinguere fonte, derivazione, responsabilità redazionale e condizioni d'uso. La preservazione attiva è intesa come ciclo periodico di controlli di fissità, aggiornamento o sostituzione dei supporti, monitoraggio dell'obsolescenza dei formati e migrazioni pianificate. La scelta tra migrare o rigenerare derivati è motivata in funzione delle proprietà significative da mantenere – fedeltà, leggibilità e riproducibilità – e va registrata come evento. Il miglioramento continuo è sostenuto da indicatori relativamente semplici ma decisivi: copertura dei metadati minimi, esiti dei controlli di fissità, conformità ai profili di qualità, tempi medi di ingest, tasso di richieste evase per livello di restrizione e quota di derivati rigenerati senza

impatto sui master. Questi indicatori rendono la curatela replicabile e comparabile anche in contesti a risorse limitate, aumentando la qualità complessiva delle evidenze disponibili. La metodologia, inoltre, include una dimensione di sicurezza spesso sottovalutata nei MAB, che prevede la gestione di malware e macro con quarantena, isolamento degli ambienti di analisi e conservazione delle evidenze delle azioni di sanificazione. Sul lato della preservazione attiva, la scelta tra migrazione ed emulazione è trattata come decisione basata su proprietà significative. Per documenti strutturati si privilegia la normalizzazione verso formati aperti quando possibile; per software e oggetti interattivi, l'emulazione o la virtualizzazione possono essere necessarie documentando dipendenze e configurazioni. Quando strettamente necessario per la narrazione, si esplicitano i controlli minimi che rendono il workflow ripetibile:

- controllo di completezza tra unità analogica descritta, file digitali, struttura di fruizione – sequenze, paginazione, legami – e tracciamento delle lacune;
- controllo di fissità programmato sui master e verifica periodica della rigenerabilità dei derivati;
- controllo di coerenza tra diritti/embargo e modalità di accesso/pubblicazione, inclusa la corretta separazione tra master e derivati;
- controllo di riproducibilità delle trasformazioni per derivati, OCR/HTR e manifest/punti di accesso.

4. METADATI, INTEROPERABILITÀ E IA: MODELLO OPERATIVO E GOVERNANCE

Nelle collezioni digitali, i metadati costituiscono l'infrastruttura che connette descrizione, preservazione e accesso. Nei contesti digitali MAB, i metadati assumono un ruolo anche epistemico, in quanto sostengono la verificabilità delle interpretazioni e i legami dell'oggetto digitale con l'originale fisico. Il modello operativo che si propone integra identificativi persistenti: metadati descrittivi che devono includere responsabilità, datazione, contesto, collocazione e relazioni archivistiche/bibliografiche; metadati tecnici relativi a formato, profilo, dimensione, checksum, parametri di acquisizione; metadati di preservazione che devono registrare eventi, agenti, strumenti e versioni, diritti e condizioni di accesso, nonché relazioni del tipo parte-di, versione-di, derivato-di. La distinzione tra metadati "osservati", vale a dire prodotti da acquisizione, controlli e registri di workflow, e metadati "inferiti", relativi ad attribuzioni, riconciliazioni e classificazioni, è mantenuta come requisito, così da evitare che arricchimenti automatici sostituiscano l'evidenza. Un nodo specifico è la gestione dell'incertezza: attribuzioni, datazioni e relazioni possono essere probabilistiche o conflittuali, mentre riconciliazioni automatiche e normalizzazioni su authority file producono inferenze che devono essere distinguibili dal dato osservato. Per questo risulta utile rendere sistematici elementi quali fonte dell'asserzione, data di acquisizione o riconciliazione, grado di confidenza e strumento usato, così da consentire revisioni e controlli a posteriori, anche in ottica di interoperabilità semantica e di conoscenza connessa tra descrizioni di biblioteca, archivio e museo. L'interoperabilità, infatti, va considerata come traducibilità controllata tra pratiche descrittive e requisiti tecnico-amministrativi: la pluralità degli standard non verrebbe eliminata, ma sarebbe governata attraverso profili applicativi, mappature esplicite e regole di validazione dei campi minimi, così che i dati possano circolare e riusarsi senza perdere provenienza e qualità. Per rendere operativa la traducibilità controllata si ritiene utile fare riferimento al recente profilo applicativo METS ECO-MiC, creato dal Ministero della Cultura (MiC) nell'ambito del *Piano Nazionale di Digitalizzazione del patrimonio culturale* (PND)[1] per la metadatazione dei beni digitizzati e per la circolazione dei pacchetti informativi. In una logica di sistema MAB, l'interesse di questo nuovo profilo applicativo non risiede solo nella struttura tecnica, ma nella capacità di rendere esplicite alcune decisioni che nei progetti di digitizzazione restano spesso implicite: quali metadati descrittivi siano assunti come minimi, quali metadati tecnici e di provenienza siano necessari per audit periodici di verificabilità, come si governi il rapporto tra master e derivati e come si colleghino diritti e modalità di fruizione. Dal punto di vista strutturale, il METS ECO-MiC sfrutta l'articolazione canonica del METS della Library of Congress (LoC)[2], ponendo però vincoli più stringenti e introducendo convenzioni coerenti per l'uso di schemi di estensione. Nella sezione descrittiva, l'impiego del profilo MODS consente di rappresentare responsabilità, identificativi, livelli di completezza e collegamenti con sistemi sorgente; nella sezione amministrativa diventano centrali la co-registrazione di metadati tecnici, diritti e provenienza digitale. L'adozione di NISO MIX per le immagini consente di esprimere in modo standardizzato parametri tecnici, caratteristiche dei file e controlli di integrità, mentre

[1] Attualmente in corso. https://digitallibrary.cultura.gov.it/il-piano/
[2] https://www.loc.gov/standards/mets/

METSRights e il collegamento a dcterms per le licenze rafforzano la propagazione dei vincoli d'uso lungo catene di esportazione e riuso. A esempio, in progetti che pubblicheranno collezioni fotografiche con visualizzatore IIIF tale struttura rende più verificabile il nesso tra file master, derivati, diritti di pubblicazione e sequenze di fruizione. Tuttavia, emergono alcune criticità che un'adozione consapevole deve considerare fin dall'impostazione del workflow. La verbosità e la complessità del METS, accentuate da un profilo vincolante, richiedono infatti competenze specialistiche e toolchain robuste; inoltre, senza automazioni affidabili, la produzione dei pacchetti risulta un collo di bottiglia. Anche l'interoperabilità reale dipende dalla coerenza delle mappature verso gli standard descrittivi di dominio e dalla capacità di mantenere allineati, nel tempo, vocabolari, versioni di schema e regole di validazione. In architetture di questo genere, funzionalità di IA potrebbero entrare in gioco validamente come tecnologie di potenziamento, a condizione che siano sottoposte a vincoli di verificabilità. Sistemi di OCR/HTR e di NLP possono supportare trascrizione e indicizzazione di manoscritti e stampati digitalizzati, estrazione di entità e normalizzazione di nomi/luoghi, clustering tematico e rilevazione di duplicati concettuali tra descrizioni e apparati. Tuttavia, i dati prodotti devono essere versionati, etichettati come derivati e sottoposti a controllo umano, anche mediante campionamento, mentre la riproducibilità richiede che sequenze e modelli siano descritti con parametri, versioni e metriche (Karatzas et al., 2011). Per i modelli generativi il rischio principale è l'opacità dell'inferenza e la produzione di risposte non immediatamente verificabili: un impiego coerente con il mandato MAB richiederebbe interfacce ancorate a recupero controllato su collezioni curate, citazione delle fonti, log delle interrogazioni e separazione tra master preservati e derivati generati, evitando riscritture "silenziose" dei metadati. Inoltre, in presenza di risposte conversazionali, è necessario registrare modello, versione, prompt e contesto documentale utilizzato, così da poter riprodurre o contestare l'esito. Sul piano della compliance, la governance include minimizzazione dei dati, accesso differenziato e anonimizzazione controllata quando necessaria, oltre a strategie contro le dipendenze da piattaforme proprietarie. L'Agenda Europea sul digital heritage insiste su sostenibilità e responsabilità come condizioni per l'innovazione (Münster et al., 2024), mentre analisi critiche sui modelli generativi richiamano l'attenzione su rischi e controversie che impattano direttamente sulla fiducia pubblica (Wach et al., 2023). In termini architetturali, il profilo applicativo cui ci si riferisce presuppone che descrizione e preservazione siano "co-registrate": gli eventi di preservazione non sono solo log tecnici, ma parte della documentazione che sostiene autenticità e responsabilità. Ne deriva l'utilità di distinguere metadati di evidenza – checksum, formati, strumenti, eventi – da metadati di interpretazione – soggetti, enti, luoghi, relazioni – e da metadati di servizio – indicizzazione, embedding, classificazioni –, trattando queste ultime due tipologie come derivati rigenerabili e versionabili. Questa distinzione è cruciale in caso di adozione dell'IA, poiché output come embedding e classificazioni automatiche tendono a essere percepiti come descrizione "stabile", mentre sono sensibili a modello, corpora e parametri. Per questo, politiche di utilizzo dell'IA, procedure di validazione e meccanismi di sostituzione controllata degli output costituiscono condizioni operative per aggiornare i derivati senza alterare i master; allo stesso modo, livelli di accesso e vincoli d'uso devono propagarsi lungo trasformazioni ed esportazioni.

5. CONCLUSIONI

Nei domini MAB la curatela di risorse e collezioni digitali non coincide con la mera conservazione di file, ma identifica la costruzione di un sistema di evidenze che renda leggibili nel tempo origine, trasformazioni, responsabilità e condizioni di riuso delle risorse. Quando la collezione è digitizzata, la qualità dipende dalla capacità della risorsa digitale di mantenere il legame sia con l'originale materiale che con la propra catena di produzione; quando la documentazione è born-digital, la qualità diventa parte dell'evidenza istituzionale e deve restare verificabile, versionabile e contestualizzata. Una riproduzione digitale di un manoscritto, una scheda di catalogo nativa digitale e una mostra online di un fondo fotografico appartengono a forme diverse della stessa responsabilità curatoriale. Il workflow di curatela delineato mira proprio a rendere leggibile questa responsabilità lungo il ciclo di vita: selezione documentata, acquisizione controllata, ingest con packaging, descrizione relazionale, accesso per livelli, preservazione attiva e indicatori minimi di monitoraggio. Il diagramma non semplifica artificialmente la complessità, ma mostra che le decisioni diventano governabili quando ogni fase lascia tracce verificabili e quando master, derivati ed output di automazione restano chiaramente distinti. In tale quadro, l'IA può entrare in gioco come estensione responsabile della curatela, migliorando scala e discovery quando gli output sono identificati come derivati, versionati, confrontabili con le fonti e sottoposti a controllo umano.

Ne deriva una conseguenza metodologica e istituzionale: per evitare la formazione di depositi digitali opachi o di servizi culturalmente efficaci ma deboli sul piano documentario, i MAB hanno bisogno di policy condivise, ruoli espliciti, metadati minimi qualificati, controlli di fissità e profili applicativi capaci di trasformare decisioni narrative in vincoli verificabili sul dato. Su questa base può svilupparsi una sperimentazione più matura dell'IA e, più in generale, una governance delle collezioni digitali che non opponga innovazione e affidabilità, ma le renda reciprocamente dipendenti. Questa impostazione avrebbe anche un effetto pratico sul lavoro quotidiano degli istituti. In una biblioteca che pubblica manoscritti digitizzati, in un archivio che rende accessibili serie documentarie con descrizioni dinamicamente aggiornate nel tempo, o in un museo che collega oggetti fisici, immagini, apparati critici e percorsi web, la qualità del servizio non dipende soltanto dalla ricchezza dell'interfaccia, ma dalla capacità di documentare come i contenuti siano stati selezionati, trasformati e resi disponibili. Un workflow più narrativamente leggibile come quello proposto aiuterebbe a evitare che la complessità tecnica resti invisibile: mostra dove si prendono le decisioni, quali evidenze vengono prodotte e in quali punti i controlli impediscono che errori, perdite di contesto o inferenze automatiche non verificate si sedimentino nei sistemi. Ne consegue che la maturità digitale dei MAB non può essere misurata soltanto dal numero di risorse digitali pubblicate o dal livello di automazione raggiunto. Conta piuttosto la capacità di mantenere allineati standard, infrastrutture, responsabilità redazionali e politiche di accesso, così che una collezione resti verificabile anche quando cambiano software, piattaforme, modelli di IA o modalità di interazione. In questo senso, l'innovazione più utile non è quella che sostituisce il giudizio curatoriale, ma quella che lo rende più esplicito, confrontabile e riproducibile. La prospettiva delineata suggerisce, quindi, una linea operativa per la ricerca e per le pratiche istituzionali: integrare profili applicativi, indicatori di qualità, strumenti di audit e automazioni trasparenti in una governance unitaria delle collezioni digitali MAB.

BIBLIOGRAFIA

Barbuti, N. (2025). La gestione del nuovo sapere digitale contemporaneo: Scenari, criticità, sfide, prospettive. In S. Rebora, M. Rospocher, & S. Bazzaco (Eds.), Diversità, equità e inclusione: Sfide e opportunità per l'informatica umanistica nell'era dell'intelligenza artificiale – Atti del XVI Convegno Annuale AIUCD (Verona, 11–13 giugno 2025), 134–140.

Carroll, L., Farr, E., Hornsby, P., & Ranker, B. (2011). A comprehensive approach to born-digital archives. Archivaria, 72, 61–92.

Digital Preservation Coalition. (2008). Preservation management of digital materials: The handbook (November 2008). https://www.dpconline.org/graphics/handbook/

Karatzas, D., Robles Mestre, S., Mas, J., Nourbakhsh, F., & Roy, P. P. (2011). Robust reading competition —Challenge 1: Reading text in born-digital images (web and email). In Proceedings of the 2011 International Conference on Document Analysis and Recognition, 1485–1489.

Kowalczyk, S. (2008). Digital preservation by design. In M. S. Raisinghani (Ed.), Handbook of research on global information technology management and service. IGI Global.

Leventhal, A., Thompson, J. (2021). Preserving born-digital design and construction records (Technology Watch Report). Digital Preservation Coalition. https://www.dpconline.org/graphics/handbook/

Münster, S., Maiwald, F., di Lenardo, I., Henriksson, J., Isaac, A., Graf, M. M., Beck, C., & Oomen, J. (2024). Artificial intelligence for digital heritage innovation: Setting up a R&D agenda for Europe. Heritage, 7(2), 794–816. https://doi.org/10.3390/heritage7020038

Pacheco, A., Guardado Da Silva, C., & Vieira De Freitas, M. C. (2023). A metadata model for authenticity in digital archival descriptions. Archival Science, 23, 629–673.

Tammaro, A. M. (2016). Heritage curation in the digital age: Professional challenges and opportunities. The International Information & Library Review, 48(2), 122–128. https://doi.org/10.1080/10572317.2016.1176454

Tammaro, A. M., Madrid, M., Casarosa, V. (2012). Digital curators' education: Professional identity vs. convergence of LAM (libraries, archives, museums). In M. Agosti, F. Esposito, S. Ferilli, N. Ferro (Eds.), Digital libraries and archives (IRCDL 2012), 184–194.

Wach, K., Duong, C. D., Ejdys, J., Kazlauskaitė, R., Korzynski, P., Mazurek, G., Paliszkiewicz, J., & Ziemba, E. (2023). The dark side of generative artificial intelligence: A critical analysis of controversies and risks of ChatGPT. Entrepreneurial Business and Economics Review, 11(2), 7–30.

[8/A] Grafi, linked data e mapping

DOG: an open-source ecosystem for mapping, narrating, and exploring humanities data

ALESSANDRO LARUFFA[1], ALESSANDRO CAPRA[2]
[1]ISTITUTO STORICO ITALIANO PER L'ETÀ MODERNA E CONTEMPORANEA
[2]DH UNICA

ABSTRACT (ENGLISH)

This paper introduces DOG, a Drupal 10 installation designed to integrate data from Omeka-S and GeoNode—more generally, from standard WMS (Web Map Service) servers—implemented by the DH UNICA - Digital Humanities Lab of Cagliari University. DOG enables the creation of websites for geographic data, digital objects, and temporal datasets. A suite of modules facilitates the visual composition of pages, including geographic data from a GeoNode instance and georeferenced objects from Omeka. These elements can also be arranged along an integrated timeline synchronised with the map. DOG integrates diverse functionalities, allowing users to explore and analyse information across multiple scales.
Leveraging established open-source platforms such as Drupal, Omeka-S, and GeoNode provides a strategic approach to software development. Access to active communities and extensive documentation enhances interoperability, problem-solving, collaboration, and sustainability. The flexibility of open-source solutions allows tailored adaptations to evolving requirements and emerging technologies. This "standing on the shoulders of giants" approach fosters the creation of innovative, scalable, and robust solutions that can be applied across diverse contexts.
DOG was inspired by research issues. The paper presents two case studies at the roots of the tool. "Colonizzazioni interne e migrazioni" is a digital history project that explores the relationship between diasporas and internal colonisation in early modern Europe. The platform collects documents, images, artworks, and cartographies related to settlement projects promoted by European authorities between the 16th and 18th centuries. "DigitalIstStor" (DIS) integrates Italian Institute for Modern and Contemporary Age (IHIMCA)'s archival materials with bibliographic references and data from national and international archival investigations. Both projects inspired DOG as a tool to harmonise heterogeneous data into interactive, coherent digital platforms that enhance public engagement.
Keywords: digital history; open source; digital archives; data visualisation.

ABSTRACT (ITALIANO)

DOG: un ecosistema open source per la mappatura, la narrazione e l'esplorazione dei dati relativi alle discipline umanistiche
Il contributo presenta DOG, un'installazione di Drupal 10 progettata per integrare dati provenienti da Omeka-S e GeoNode – più in generale da server WMS (Web Map Service) standard – sviluppata dal DH UNICA – Digital Humanities Lab dell'Università di Cagliari. DOG consente di realizzare siti web per la presentazione di dati geografici, oggetti digitali e dataset temporali. Un insieme di moduli consente la composizione visiva delle pagine, integrando dati geografici provenienti da un'istanza di GeoNode e oggetti georeferenziati in Omeka. Tali elementi possono inoltre essere organizzati lungo una timeline integrata, sincronizzata con la mappa. DOG combina funzionalità eterogenee, permettendo agli utenti di esplorare e analizzare le informazioni a diverse scale.
L'adozione di piattaforme open source consolidate, come Drupal, Omeka-S e GeoNode, rappresenta una scelta strategica nello sviluppo software. L'accesso a comunità attive e a un'ampia documentazione favorisce l'interoperabilità, la risoluzione dei problemi, la collaborazione e la sostenibilità nel tempo. La flessibilità delle soluzioni open source consente inoltre adattamenti mirati a requisiti in evoluzione e a tecnologie emergenti. Questo approccio, riconducibile al principio del "standing on the shoulders of giants", favorisce la realizzazione di soluzioni innovative, scalabili e robuste, in grado di adattarsi a diversi contesti.

DOG è nato da specifiche esigenze di ricerca. L'articolo presenta due casi studio che costituiscono le radici dello strumento. “Colonizzazioni interne e migrazioni” è un progetto di digital history che esplora il rapporto tra diaspore e colonizzazione interna nell’Europa della prima età moderna. La piattaforma raccoglie documenti, immagini, opere d'arte e cartografie relative ai progetti di insediamento promossi dalle autorità europee tra il XVI e il XVIII secolo. “DigitalIstStor” (DIS) integra i materiali d'archivio dell’Istituto Storico Italiano per l’Età Moderna e Contemporanea (ISIME) con riferimenti bibliografici e dati provenienti da indagini archivistiche nazionali e internazionali. Entrambi i progetti hanno ispirato DOG come strumento per armonizzare dati eterogenei in piattaforme digitali interattive e coerenti, capaci di favorire il coinvolgimento del pubblico
Parole chiave: storia digitale; open source; archivi digitali; visualizzazione di dati.

1. INTRODUCTION

In recent years, the Digital Humanities (DH) landscape has increasingly emphasised the integration of spatial and temporal data as essential dimensions of humanities research, both methodologically and technologically (Spadi, Degl’Innocenti, & Di Meo, 2024). This shift has contributed to the development of the spatial humanities, a field where geospatial methods, digital mapping, and narrative visualisation intersect to support analysis, storytelling, and public dissemination (Alvanides, Paulino, & Bartzokas-Tsiompras, 2025).

The adoption of Geographic Information Systems (GIS) in humanities research has enabled the exploration of spatial patterns, relationships, and processes in cultural heritage and historical inquiry (Gregory & Healey, 2007; Kabadayi et al., 2022). Web-based GIS platforms further support public engagement and interactive exploration, leveraging standards such as Web Map Service (WMS) to integrate heterogeneous datasets across platforms (Goodchild, 2007). Despite these advances, GIS tools often lack native support for integrating rich narrative structures, temporal datasets, and digital objects within a modular, user-friendly framework.

Temporal visualisation has similarly emerged as a critical dimension for structuring historical narratives. Tools such as TimelineJS, Neatline, and other timeline frameworks allow the organisation of events, sources, and objects chronologically, often with multimedia support (Nowviskie, 2013; Harvard Digital Methods Resources, 2025). While these solutions excel at narrative presentation, their interoperability with complex geospatial infrastructures remains limited, often requiring technical expertise for integration.

Several platforms attempt to combine narrative, spatial, and temporal elements. Open-source frameworks such as Omeka and extensions such as Neatline enable mapping and timeline visualisation alongside digital collections (Nowviskie, 2013). Proprietary or lightweight tools such as StoryMapJS and ArcGIS StoryMaps provide rich storytelling capabilities but often operate within closed ecosystems or simplified data models, limiting long-term sustainability and extensibility (Esri, 2021; Knight, 2017). Conversely, GeoNode provides robust GIS management and sharing capabilities but does not natively support temporal storytelling or integration with digital objects (Corti, Bartolini, Tzotsos, & Delipetrev, 2019).

Recent scholarship emphasises the importance of modular, interoperable, and community-driven infrastructures in DH (Ciotti, 2021; Ciotti & Roncaglia, 2021; Quill, 2024). Platforms that build on widely adopted open-source tools, rather than bespoke systems, offer greater sustainability, adaptability, and scalability, enabling multi-scale exploration of heterogeneous data and integration of new methods and technologies (Luzietti et al., 2025).

Despite a growing ecosystem of spatial, temporal, and narrative tools, the landscape remains fragmented: mapping platforms, timeline tools, and content management systems frequently exist in isolation. There remains a notable gap in frameworks capable of coherently integrating geographic data, digital objects, and temporal datasets within a modular, web-oriented environment while maintaining interoperability with established open-source infrastructures (Spadi, Degl’Innocenti, & Di Meo, 2024; Rodrigues & Correia, 2025).

To address this gap, this paper examines the specific research and methodological needs of two distinct digital history projects. By analysing their challenges with data heterogeneity, changing borders, and spatiotemporal synchronisation, we then introduce DOG, a unified web-oriented framework designed to address these research objectives directly. A unified approach that integrates spatial, temporal, and

narrative capabilities with open standards and active communities, thereby contributing to ongoing discussions on sustainable, flexible, and interoperable DH platforms.

2. COLONIES AND DIASPORAS IN EARLY MODERN EUROPE: "COLONIZZAZIONI INTERNE E MIGRAZIONI"

"Colonizzazioni interne e migrazioni" is a digital history project that explores the relationship between diasporas and internal colonisation in early modern Europe.[1] In the early modern period, European states developed internal colonisation policies to promote demographic, agricultural, and commercial growth. Between the sixteenth and eighteenth centuries, foreign settlers played a central role in these strategies of territorial reorganisation. While numerous studies have examined individual cases of internal colonisation (Ciasca, 1926; Aymard, 1974; Bartlett, 1979; Alcoberro i Pericay, 2011; Santus, 2013; Moya García & Cuesta Aguilar, 2014), a comprehensive and comparative synthesis of these phenomena remains lacking. The research project "Migration and Internal Colonisation in the Early Modern Mediterranean", which lay the groundwork for "Colonizzazioni interne e migrazioni", seeks to address this historiographical gap (Salice, 2017).

In addition to providing concise narrative syntheses of selected case studies, the platform collects, enriches with metadata, and geolocates documents, images, artworks, and cartographic materials related to settlement projects promoted by European authorities between the sixteenth and eighteenth centuries. Each narrative entry is developed through the systematic analysis of available sources, including scholarly literature, archival records, iconographic and architectural evidence, historical maps, and, where possible, oral memories. In this context, "Colonizzazioni interne e migrazioni" represents a dynamic process for the systematic collection, analysis, and dissemination of historical data, enhanced through computational tools and methods. Archival, bibliographic, and field-collected materials—including photographs, oral testimonies, and other multimedia—are structured in standardised spreadsheets and imported into Omeka-S. Metadata are aligned with controlled vocabularies such as Dublin Core, SAN, and OAD. All materials are transformed into digital objects and spatially contextualised through interactive cartography embedded within the narrative framework.

The portal is structured around two thematic pathways. The first, Populating with Foreigners, provides an overview of internal colonisation policies pursued by various European states from the Renaissance to the Enlightenment. During this period, public authorities implemented settlement strategies aimed at attracting migrants to depopulated and unproductive areas in order to secure, reclaim, cultivate, and connect marginal territories to major urban centres and ports, thereby integrating them into a sovereign-controlled space of production and exchange. The second pathway, From Migrants to Colonists, reconstructs the trajectories of different migratory flows and their incorporation into settlement and territorial reorganisation policies across the Euro-Mediterranean region. These processes unfolded along complex spatial and temporal trajectories, often at religious and political frontiers, and involved port cities, courts, and governmental institutions. By presenting accessible syntheses of Mediterranean migratory experiences that resulted in settlement projects abroad, the platform highlights the entangled dynamics of migration, governance, and territorial transformation in the early modern period.

Spatial layers are generated in QGIS, in which each point combines location with attributes derived from documentary, bibliographic, and multimedia sources. Historical and contemporary base maps are integrated, including georeferenced charts and portolans, allowing diachronic analyses of maritime routes and coastal occupation. Political borders for different periods are represented across multiple layers, styled with graduated symbology, and uploaded to GeoNode. GeoNode supports continuous updating, collaborative editing, and filtering by thematic or chronological criteria. Crucially, it enables the creation of multi-layered maps and geostories that provide a dynamic spatial and temporal perspective on internal colonisations. Consequently, integration and public dissemination require a system that combines ? metadata-rich items and geospatial layers into interactive web pages, allowing multiple WMS layers to be displayed simultaneously on maps and synchronised with interactive timelines. In "Colonizzazioni interne e migrazioni", the core methodological challenge lies in creating a tightly integrated workflow among digital object repositories, GIS tools, and web-publishing layers.

[1] See https://storia.dh.unica.it/colonizzazioninterne/it/home (accessed 7 February 2026)

3. FROM ROME TO EUROPE. DIS (DIGITALISTSTOR) AND DIGITAL CULTURAL HERITAGE

The DIS (DigitalIstStor) is a web portal that establishes a comprehensive knowledge framework for the Italian Historical Institute for the Modern and Contemporary Age (IHIMCA)'s heritage, with a specific focus on its archival materials.[2] The DIS integrates the reference bibliography with documentation held in IHIMCA's archive and data generated through archival investigations in local, national, and international sites. The pilot of DIS is the Association of European Historians (AEH) archival group: recent studies have analysed unpublished documents about the Association, its genesis, its political and cultural ties, as well as its evolution over time (Guerra, 2016, 2018, 2023; Laruffa, 2022; Valente, 2023). These unexplored materials raise several questions concerning the connections and interactions among the AEH, the European Community (EC) institutions, other scholarly networks, and the historiographical strands of the time.

In 1985, the Association began collaborating with historians from the Soviet bloc. Within this framework, the AEH sought to promote a "new" European narrative by reinterpreting the continent's history through a collective lens—free from nationalism and political subordination. The goal was to highlight the shared roots and the richness of Europe's internal diversity. Accordingly, alongside scholars from EC member states, the Association engaged researchers from the Soviet bloc to approach the historical evolution of European society from a broad perspective. These contacts offer new perspectives on the interactions between East and West during the Cold War in the 1980s. Most Cold War studies focus on high politics and international relations among state actors, and a fundamental assumption is that the Iron Curtain created an insurmountable barrier between two rigid blocs (Westad, 2013; Levering, 2016; McMahon, 2021). Shifting the focus beyond superpower dynamics, we can observe other layers of interaction that, although different, played equally important roles in shaping individuals' everyday lives during the Cold War (Auti-Sarasmo, 2010, 2011; Romano, 2022). Accordingly, the newly discovered archival documents could reveal the Association of European Historians as an international scholarly network that extended beyond the Iron Curtain.

By this means, spatialising the AEH data can reveal the size and evolution of these interactions before and during the fall of the Iron Curtain. To satisfy these research objectives, the historical workflow requires the progressive transformation of heterogeneous historical data into interoperable digital objects, spatial layers, and narrative visualisations.

Building upon this structured corpus of metadata-digital objects, the second phase focuses on the spatialisation of data using GIS technologies. The historical scope of the case study—characterised by rapidly changing political borders—required a more complex cartographic solution than a single static base map. To address this issue, multiple historical boundary layers corresponding to different chronological phases were created and uploaded to GeoNode. GeoNode supports collaborative workflows, thematic and chronological filtering, and the integration of GIS data into external applications, thus defining the requirements for a multi-layer geographic infrastructure.

The final stage of the workflow concerns web-based synthesis and storytelling. This requires an extended layout builder capable of creating interactive visual components that synchronise maps and timelines, allowing digital objects to be explored chronologically. At the same time, their geographic distribution is simultaneously visualised across multiple selectable WMS layers. Within this framework, spatial, temporal, and documentary dimensions converge into a coherent system that supports both scholarly analysis and public-facing communication.

4. THE DOG FRAMEWORK

"Colonizzazioni interne" and DIS research challenges demand a platform capable of synchronising rich archiving, multi-layered cartography, and temporal progression. The DOG framework was built specifically to answer these needs. It relies on a set of open-source platforms and modules that together enable seamless integration, visualisation, and interaction with digital objects, geospatial data, and temporal datasets. By combining well-established content management and geospatial systems with custom-developed components, DOG provides a modular environment that supports interactive, multi-scale exploration of cultural heritage resources.

[2]See https://www.digitaliststor.it/ (accessed 7 February 2026)

At its core, DOG relies on Drupal, a mature open-source content management system that has been under continuous development since 2001. Drupal offers a flexible and scalable foundation for a wide range of websites, from news portals and community platforms to complex institutional sites. Its modular architecture enables the integration of additional functionality through contributed or custom modules, making it an ideal backbone for a system that must orchestrate diverse content types and external services.

Complementing Drupal, Omeka-S serves as the primary repository for digital objects. First released in 2008, Omeka-S is a CMS specifically designed for managing and presenting digital collections. Within DOG, Omeka-S objects can be queried and displayed dynamically on Drupal pages, providing editors with a streamlined way to integrate rich cultural content without duplicating data.

The geospatial dimension of DOG is handled through GeoNode, a web-based platform for publishing and managing geographic data that has been in development since 2010. GeoNode, together with any geoserver that exposes data via the Web Map Service (WMS) standard, allows DOG to visualise georeferenced layers directly on the website. This integration enables a coherent presentation of maps alongside digital objects, allowing users to explore both spatial and temporal aspects of the data simultaneously.

A key feature that demonstrates the power of this integrated architecture is the timeline map block, which links geospatial data to temporal visualisation. Configurable through Drupal's Layout Builder, this component allows editors to present georeferenced and time-stamped objects, multimedia content, and external WMS layers in a single, interactive interface. Each WMS layer can be assigned a temporal range, defining when it appears relative to other data points on the map or along the timeline. This feature facilitates complex narrative and exploratory workflows, enabling users to visualise historical and cultural datasets across both space and time.

The connection between Drupal and Omeka-S is managed by a dedicated Drupal–Omeka-S connector that interfaces directly with the Omeka-S API. Accessible through the administration panel, this connector allows configuring the Omeka-S resource URL and, when required, API keys for restricted resources. Once configured, the system validates the connection and enables real-time browsing and querying of published items, ensuring that the website always reflects the most up-to-date content.

By combining these elements—Drupal's content management capabilities, Omeka-S's digital object repository, GeoNode's geospatial services, and dynamic visualisation modules such as the timeline map block—DOG creates a coherent ecosystem in which heterogeneous datasets can be integrated, explored, and presented interactively. Its architecture exemplifies a modular and interoperable approach to Digital Humanities infrastructure, facilitating not only the display of content but also the construction of complex, multi-layered narratives spanning space and time

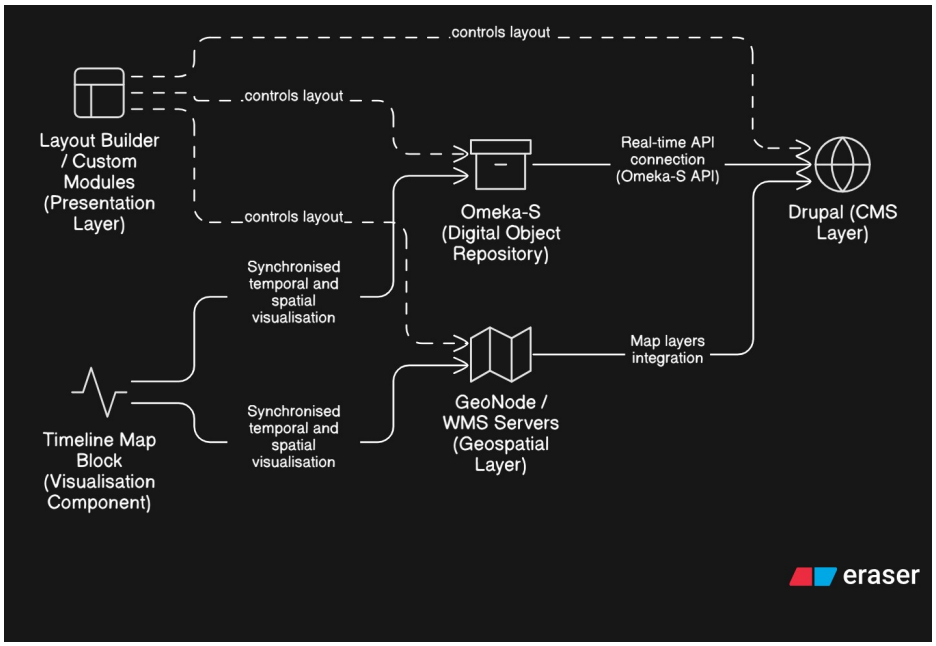


Figure 1. DOG system architecture schematic.

As a result, the DOG can adapt to a wide range of contexts. In the next paragraphs, we will address two user stories. By validating this architecture against the previously discussed requirements of "Colonizzazioni interne" and DIS, the DOG tool proves to be an effective technical answer to complex research challenges.

5. CONCLUSIONS

The DOG exhibits versatility and adaptability, making it suitable for a wide range of institutional and research contexts. Its deployment is not limited to university environments; it can be effectively applied in cultural institutions, archives, museums, and research centres that require the integration, visualisation, and dissemination of heterogeneous data. The modular architecture of DOG allows components to be customised according to the specific needs of a project, whether for small-scale case studies or for extensive, multi-layered digital infrastructures. By separating the management of digital objects, geospatial data, and interactive layouts into distinct yet interoperable layers, DOG enables projects to scale incrementally while maintaining structural coherence.

The framework's open-source foundation enhances its flexibility and sustainability. Leveraging platforms such as Drupal, Omeka-S, and GeoNode ensures access to active and experienced communities, fostering collaboration, knowledge sharing, and rapid problem-solving. This community-oriented context also promotes interoperability, as all components conform to widely recognised standards, including OGC geospatial data protocols and Linked Open Vocabularies for metadata. Such alignment facilitates the integration with external datasets, ensuring that outputs are both reusable and extendable.

Importantly, DOG should not be understood merely as a static software product; it is conceived as a processual framework for the production, organisation, and dissemination of knowledge. Its implementation encompasses workflows spanning data acquisition, geospatial representation, digital object management, and interactive web publication, supporting both the analytical and narrative dimensions. By releasing DOG as open-source software, the project invites public participation in its co-creation, encouraging contributions that expand its functionality and refine its design in response to the evolving needs of the digital humanities community. This approach positions DOG as a living platform in which collaborative development and iterative improvement are integral to its conceptual and operational framework. Accordingly, the adoption of a headless architecture implemented using an API-first pattern has yielded significant advantages. The headless approach has enabled the rapid development of new tools using JavaScript-based programming languages that interact directly with system APIs. This has significantly streamlined the creation and deployment of customised applications (Salice, 2023), and more will follow.[3]

Beyond its analytical and infrastructural capabilities, DOG also plays a significant role in fostering public engagement by constructing user-centred narrative pathways. The user stories presented in this paper illustrate how the platform supports forms of digital storytelling that make complex historical data accessible, explorable, and meaningful to non-specialist audiences. By synchronising maps, timelines, and digital objects within interactive interfaces, DOG enables users to actively navigate historical narratives, aligning with digital public history approaches that emphasise participation, interpretation, and the circulation of knowledge beyond academic boundaries (Noiret, 2022). Accordingly, user stories mediate devices between scholarly research and wider publics, encouraging exploration, comparison, and contextual understanding across spatial and temporal scales. Such practices resonate with the concept of shared authority, which redefines historical knowledge production as a dialogical process involving historians, institutions, and communities, particularly within digital environments (Cauvin, 2021).

Moreover, DOG's modular architecture and reliance on open standards create favourable conditions for future participatory extensions, including crowdsourcing, user-generated content, and collaborative annotation workflows. As such, DOG contributes to a broader reconfiguration of historical practice, in which public engagement becomes an embedded component of digital humanities infrastructures.

To summarise, DOG exemplifies a scalable, modular, and interoperable infrastructure that bridges archival, spatial, and narrative practices within a single coherent system. Its open-source ethos and community-driven design provide both flexibility and robustness, enabling researchers and institutions to

[3]DOG is going to be adopted in the next future by The Atlas of Language and Texts of Italian Gastronomic Culture from the Middle Ages to Unification (AtLiTeG - see https://www.atliteg.org/ accessed 7 February 2026) and Historical Glossary of Medieval Italian Commercial and Accounting Terminology (see https://glossari.dh.unica.it/ accessed 7 February 2026).

develop custom solutions while benefiting from shared expertise and collective problem-solving. Thus, DOG represents not only a practical tool for digital scholarship but also a methodological model for sustainable, collaborative, and open-ended digital humanities practice, adaptable to diverse domains and continuously evolving through the contributions of its users.

REFERENCES

Alcoberro i Pericay, A. (2011). La "nova Barcelona" del Danubi (1735–1738): La ciutat dels exiliats de la guerra de successió. Rafael Dalmau.

Alvanides, S., Paulino, J. V., & Bartzokas-Tsiompras, A. (Eds.). (2025). Spatial humanities and contemporary geographical approaches [Special issue]. European Journal of Geography.

Aymard, M. (1974). La Sicile, terre d'immigration. Cahiers de la Méditerranée, 2(1), 134–157. https://doi.org/10.3406/camed.1974.1791

Bartlett, R. P. (1979). Human capital: The settlement of foreigners in Russia, 1762–1804. Cambridge University Press.

Cauvin, T. (2021). Shared authority in public history. Bloomsbury Academic. https://doi.org/10.5040/9781350927926

Ciasca, R. (1926). Momenti della colonizzazione in Sardegna nel secolo XVIII. Annali della Facoltà di Lettere e Filosofia della Regia Università di Cagliari, 1, 93–174.

Ciotti, F. (2021). Digital humanities and the challenge of sustainability. Umanistica Digitale, 5(10), 11–25. https://doi.org/10.6092/issn.2532-8816/11254

Ciotti, F., & Roncaglia, G. (2021). Digital humanities: Metodi, strumenti, saperi. Carocci.

Corti, P., Bartolini, A., Tzotsos, A., & Delipetrev, B. (2019). GeoNode: An open source framework to build spatial data infrastructures. ISPRS International Journal of Geo-Information, 8(7), 294. https://doi.org/10.3390/ijgi8070294

Day, J. (1987). Uomini e terre nella Sardegna coloniale: XII–XVIII secolo. Celid.

Esri. (2021). ArcGIS StoryMaps: Storytelling with maps. Esri Press.

Goodchild, M. F. (2007). Citizens as sensors: The world of volunteered geography. GeoJournal, 69(4), 211–221. https://doi.org/10.1007/s10708-007-9111-y

Gregory, I. N., & Healey, R. G. (2007). Historical GIS: Structuring, mapping and analysing geographies of the past. Progress in Human Geography, 31(5), 638–653. https://doi.org/10.1177/0309132507081495

Guerra, A. (2016). Fra via Caetani e l'Europa. Armando Saitta e l'Istituto Storico Italiano per l'età Moderna e Contemporanea. Annali di storia moderna e contemporanea, (4), 77–106.

Guerra, A. (2018). Armando Saitta. Due storie d'Europa. Sapienza Università di Roma.

Guerra, A. (2023). "Svegliare le menti e le coscienze a un sentire europeo, e non più semplicemente nazionale". La storia europea di Armando Saitta. In A. Laruffa & M. Verga (Eds.), Armando Saitta e la ricerca storica nell'Italia del secondo Novecento (pp. 255–270). Viella.

Harlaftis, G., & Vassallo, C. (2004). New directions in Mediterranean maritime history. International Maritime Economic History Association.

Ivetic, E. (2022). Il grande racconto del Mediterraneo. Il Mulino.

Knight, L. (2017). TimelineJS and StoryMapJS: Tools for digital storytelling. Northwestern University Knight Lab.

Laruffa, A. (2022). Historians' affiliation with the European integration process in the 1980s. Journal of European Integration History, 28(2), 269–292.

Levering, R. B. (2016). The Cold War: A post–Cold War history. John Wiley & Sons.

Luzietti, R. B., Spadi, A., Giampietro, N., Mancuso, G., Caravale, A., D'Eredità, A., … Quochi, V. (2025). Digital humanities and heritage science: Moving from landscaping to a dynamic research observatory in an Open Science Cloud. Umanistica Digitale. https://doi.org/10.6092/issn.2532-8816/21226

Manconi, F. (2010). La Sardegna al tempo degli Asburgo: Secoli XVI–XVII (Vol. 5). Il Maestrale.

McMahon, R. J. (2021). The Cold War: A very short introduction. Oxford University Press.

Moya García, E., & Cuesta Aguilar, M. J. (2014). Las colonizaciones de la sierra sur de Jaén y Carolina de Sierra Morena: Procesos y resistencias de las localidades matrices. In M. A. López Arandia (Ed.), Ciudades y fronteras: Una mirada interdisciplinar al mundo urbano (ss. XII–XXI) (pp. 119–134). Universidad de Extremadura.

Noiret, S. (2022). Sharing authority in online collaborative public history practices. In S. Noiret, M. Tebeau, & G. Zaagsma (Eds.), Handbook of digital public history (pp. 49–60). De Gruyter Oldenbourg. https://doi.org/10.1515/9783110430295-004

Nowviskie, B. (2013). Neatline: Designing space and time for the humanities. Journal of Digital Humanities, 2(1).

Quill, T. (2024). Spatial humanities: Perspectives from the library. Journal of Library and Information Science. https://doi.org/10.1080/15420353.2024.2356495

Rodrigues, A., & Correia, N. (2025). Using technology in digital humanities for learning and knowledge dissemination. arXiv. http://arxiv.org/abs/2501.10275

Romano, A. (2022). The European Community and Eastern Europe in the long 1970s: Challenging the Cold War order in Europe. Routledge.

Salice, G. (2017). Migrazioni e colonizzazione interna nel Mediterraneo d'età moderna: Un progetto di umanistica digitale. RiMe – Rivista dell'Istituto di Storia dell'Europa Mediterranea, 1(I n.s.), 95–117.

Salice, G. (2023). ASMSA. Atlante digitale per la storia marittima della Sardegna. Umanistica Digitale, (15), 117–132.

Santus, C. (2013). Moreschi in Toscana. Progetti e tentativi di insediamento tra Livorno e la Maremma (1610–1614). Quaderni Storici, (3), 745–778. https://doi.org/10.1408/75614

Spadi, A., Degl'Innocenti, E., & Di Meo, C. (2024). DARIAH.it: Data integration strategies and solutions for digital resources management and research in the arts and humanities. Mimesis Journal, 13(2), 119–134. https://doi.org/10.13135/2389-6086/9920

Valente, M. (2023). Un'internazionale degli storici? Armando Saitta e le reti storiografiche europee. In A. Laruffa & M. Verga (Eds.), Cultura e organizzazione. Armando Saitta e la storiografia italiana del secondo Novecento (pp. 181–204). Viella.

Westad, O. A. (Ed.). (2013). Reviewing the Cold War: Approaches, interpretations, theory. Routledge. https://doi.org/10.4324/9781315039824

Beyond Spatial Data: Exploring Gender and Property in the Catasto Generale Toscano through the Florentia Illustrata Semantic Platform

ERICA ANDREOSE[1] , REMO GRILLO[1,2] , GIANMARCO SPINACI[1,2]

[1] PATRIMONIO CULTURALE NELL'ECOSISTEMA DIGITALE, UNIVERSITÀ DI BOLOGNA, BOLOGNA, ITALY, ERICA.ANDREOSE2@UNIBO.IT

[2] I TATTI - THE HARVARD UNIVERSITY CENTER FOR ITALIAN RENAISSANCE STUDIES, FIRENZE, ITALY, RGRILLO@ITATTI.HARVARD.EDU, GSPINACI@ITATTI.HARVARD.EDU

ABSTRACT (ENGLISH)

This contribution presents the results of a case study conducted on *Florentia Illustrata*, a research infrastructure based on the *ResearchSpace* system that integrates georeferenced data from the *Catasto Generale Toscano* into a Knowledge Graph and visualizes them through an interactive map. The study evaluates the extent to which a semantic architecture grounded in CIDOC-CRM can support socio-economic and intersectional inquiry starting from a dataset that was originally alphanumeric and static. The assessment is conducted through an exploratory analysis of female landownership in nineteenth-century Florence, using SPARQL queries to extract and compare ownership patterns by gender. The results show that women are clearly represented as property holders in the cadastral records, although at a much smaller scale than men. The combination of semantic querying and interactive spatial visualization enables these patterns to be explored at multiple scales, from city-wide distributions to individual parcels and owners. At the same time, the study highlights limitations related to gender inference and identity resolution in historical sources, underscoring the need for explicit handling of uncertainty. Overall, the contribution demonstrates how a semantic, event-based Knowledge Graph can extend the analytical potential of cadastral data beyond spatial description, enabling flexible and user-driven historical inquiry.

Keywords: gender history; parcel-based cadastre; Knowledge Graphs; nineteenth-century Florence; Linked Open Data.

ABSTRACT (ITALIANO)

Il contributo illustra i risultati di un caso di studio condotto su *Florentia Illustrata*, un'infrastruttura di ricerca basata su *ResearchSpace* che integra dati georeferenziati del *Catasto Generale Toscano* in un grafo di conoscenza e li visualizza attraverso una mappa interattiva. Lo studio valuta in che misura un'architettura semantica fondata su CIDOC-CRM sia in grado di supportare interrogazioni di natura socio-economica e intersezionale a partire da un dataset originariamente alfanumerico e statico. La valutazione è condotta attraverso l'analisi delle proprietà femminili nella Firenze dell'Ottocento, utilizzando query SPARQL per estrarre e confrontare i pattern in base al genere. I risultati mostrano che le donne sono presenti come proprietarie nei registri catastali, sebbene su una scala nettamente inferiore rispetto agli uomini. La combinazione tra ricerca semantica e visualizzazione spaziale interattiva consente di esplorare tali pattern su scale diverse, dalla distribuzione sull'intero tessuto urbano fino al livello dei singoli appezzamenti e dei proprietari. Lo studio mette anche in luce le limitazioni legate all'inferenza del genere e alla risoluzione delle singole identità nelle fonti storiche, sottolineando la necessità di una gestione esplicita dell'incertezza. Nel complesso, il contributo dimostra come un Knowledge Graph semantico ed event-based possa estendere il potenziale analitico dei dati catastali oltre la sola descrizione spaziale, permettendo forme di indagine storica flessibili e guidate dall'utente.

Parole chiave: storia di genere; catasto particellare; Knowledge Graphs; Firenze XIX secolo; Linked Open Data.

1. INTRODUCTION

Florentia Illustrata is a digital platform based on the ResearchSpace system (Oldman & Tanase, 2018). The project aims to make nineteenth-century Florentine cadastral data accessible by transforming them

into a semantically enriched digital resource. At the core of *Florentia Illustrata* lies the cadastral dataset[1] (Belli et al. 2022), an extensive collection of cadastral georeferenced maps and detailed descriptive tables of land parcels, providing information on ownership, toponymy, land use, and other key attributes. *Florentia Illustrata* integrates these heterogeneous sources into a Knowledge Graph that supports both exploratory analysis and complex querying. This article is conceived as a case study and a stress-test of the flexibility of the technological infrastructure and of the underlying Knowledge Graph[2]. ResearchSpace, the research infrastructure underlying the application, is designed as a research environment that enables users to act as active designers of their own knowledge systems (Oldman, 2021). From this perspective, *Florentia Illustrata* is employed as a testbed to assess whether, and to what extent, the platform allows users to extract information that goes beyond the purely urban or spatial dimension, engaging instead with social and economic questions.

As a case study, this contribution focuses on a specific and often underrepresented segment of the dataset: women landowners in nineteenth-century Florence. Despite the legal and social constraints affecting women during this period, the dataset reveals a significant female presence among property holders, which invites further investigation. The research is guided by two sets of questions addressing different but complementary aspects. The first concerns the structure and capabilities of the platform itself: (1) *Is the dataset, together with the dynamic visualization and semantic modeling proposed by Florentia Illustrata, sufficiently rich and flexible to support socio-economic research on the historical urban context? What improvements to the Knowledge Graph could further enhance this potential? (2) Are female landowners represented in the dataset? What types of urban spaces did they control, and how do these patterns compare with those associated with male ownership?*

2. METHODOLOGY

To address these questions, the study adopts the workflow and interaction paradigm provided by *Metapolis*, a *ResearchSpace* map-centred research suite conceived to support spatio-temporal historical inquiry by bringing together georeferenced sources, interpretative work, and semantic data in a single environment. In this setting, the role of geospatial visualization is not a final output, it's rather the main point of access to the underlying Knowledge Graph and a practical way to formulate and test hypotheses through exploration. *Florentia Illustrata* in particular is implemented within *Metapolis* using its advanced semantic mapping interface (Galeazzo et al., 2024) to translate custom SPARQL queries into interactive layers on the cadastral map[3]. This makes it possible to isolate specific subsets of parcels and actors (e.g., women property holders) and immediately inspect their spatial distribution, while iteratively refining the query and the visual criteria. The interface supports the separate management of historical map layers and enables dynamic coloring and labeling driven by query results (for example, ownership specialization, property categories, or actor types). Beyond the two-dimensional view used in this case study, the *Metapolis* mapping environment is designed for richer spatio-temporal interaction, including time-based filtering and the possibility (conceptually and at the level of the UI design) of extending analysis toward representations of change across a broader space–time volume. Overall, the methodological choice is aligned with the goal of making complex research questions answerable through an interface that encourages engagement, direct and intuitive spatial exploration rather than relying only on tabular outputs.

The dataset (Belli et al., 2022) derives from the first geometric-parcel cadastre of Florence (c. 1735–1835) and consists of 17,681 records, each corresponding to a unique land parcel (*appezzamento*). Within the cadastral system, an *appezzamento* refers to a discrete unit of land represented on the maps as a polygonal shape defining its physical boundaries and associated with attributes such as size, function, and ownership. The RDF Knowledge Graph of *Florentia Illustrata* is structured according to the

[1] Firenze University Press, Redazione (2023). Mappe. figshare. Dataset. https://doi.org/10.6084/m9.figshare.23500524.v1 (cons. 03/02/2026).
Firenze University Press, Redazione (2023). Banche dati geografiche. figshare. Dataset. https://doi.org/10.6084/m9.figshare.23500503.v1 (cons. 03/02/2026)
Firenze University Press, Redazione (2023). Banche dati alfanumeriche. figshare. Dataset. https://doi.org/10.6084/m9.figshare.23500473.v1 (cons. 03/02/2026)

[2] Andreose, E. (2025). Florentia Illustrata Knowledge Graph. Università di Bologna. https://amsacta.unibo.it/id/eprint/8236/ (cons. 20/04/2026)

[3] A list of the queries is available in the repository online: https://github.com/EricaAndreose/florentiaillustrata_gender_analysis. (cons. 20/04/2026)

CIDOC Conceptual Reference Model (CIDOC-CRM), selected for its ability to represent historical entities and processes as temporally situated events rather than static attributes (Doerr et al., 2007). Land parcels are modeled as instances of E24_Physical_Human_Made_Thing, while ownership relations are represented through E8_Acquisition, which captures events of legal transfer and introduces a temporal and contextual dimension to property relations. Individuals involved in these transactions are modeled as E21_Person and linked to acquisition events via the inverse property P22i_acquired_title_through. Personal attributes preserved in the cadastral sources, such as given name, surname, patronymic, and relational descriptors (e.g. "wife of", "widow"), are retained as identifiers associated with person entities, enabling socially oriented queries.

The Knowledge Graph was queried using SPARQL to address research questions related to gender and land ownership. Since the original cadastral data do not include an explicit gender attribute, the analysis initially relied on relational identifiers preserved in the sources. In particular, individuals associated with the identifier *"moglie di"* ("wife of") were operationally treated as female landowners. This criterion enabled the extraction and aggregation of ownership data, allowing the calculation of both the number of parcels and the total surface area associated with each identified woman through acquisition events. During this process, it became evident that this operational definition risked excluding women not recorded through marital relationships, such as unmarried women or individuals identified through alternative social or legal roles. This issue directly informed the limitations discussed in the following section and motivated additional disambiguation procedures.

2.1 LIMITATIONS

The limitations discussed here derive directly from the structure of the original cadastral sources and their semantic encoding. The Catasto Generale Toscano does not record biological sex or gender as explicit attributes, reflecting the administrative priorities of the period. Therefore, in this exploration, gender attribution is indirectly inferred[4]. While this does not affect conventional cadastral analysis, it introduces uncertainty when addressing gender-oriented research questions.

To mitigate this limitation, an additional disambiguation strategy was applied outside the ResearchSpace environment, working directly on the RDF serialization of the Knowledge Graph. The goal was to identify potential female landowners not captured by the initial query criteria, while minimizing false positives. The pipeline is resumed as follows. **(1)** All individuals modeled as instances of *E21_Person* were extracted from the Knowledge Graph. From this set, persons explicitly associated with a *"moglie di"* identifier were isolated and temporarily excluded. **(2)** Personal given names of the remaining individuals were compared against an external reference dataset of Italian given names. This dataset, derived from an open GitHub repository[5] published in July 2025, includes names annotated with gender attribution and statistical confidence values. **(3)** Names were retained only if they appeared exclusively under gender = "f" across all entries in the dataset. To avoid false assumptions based on names statistics, any name carrying at least one masculine entry was discarded, regardless of the relative frequency of its feminine occurrences. **(4)** Through this process, 79 individuals were identified as likely unmarried female landowners. Among them, three were explicitly associated with the title *"suora"* (nun), indicating religious affiliation rather than marital status. These individuals were subsequently reintegrated into the main research dataset and treated as a distinct and analytically relevant subgroup It must be noted that neither strategy — relying on a relational descriptor tied to male property holders, nor on probabilistic name-gender attribution — provides certainty in identifying female landowners. This epistemological limit is itself historically meaningful: the degree of noise surrounding female identities in the dataset reflects the extent to which these sources were not designed to make women visible as autonomous actors in the cadaster.

A further limitation concerns identity resolution and homonymy. Neither the original cadastral sources nor the Knowledge Graph provide sufficient biographical detail to reliably distinguish between individuals sharing the same name. Although person entities are modeled using combinations of given name, surname, patronymic, and relational descriptors, these identifiers are neither consistently present nor systematically unique. As a result, distinct historical actors may be conflated, while references to the

[4] For the technical implementation, see the repository online https://github.com/EricaAndreose/florentiaillustrata_gender_analysis. (cons. 20/04/2026)

[5] The external list of Italian names is available at: https://gist.github.com/tezzutezzu/8f025345cadc5f92b9b311bf032b264d. (cons. 20/04/2026)

same individual may remain fragmented. Addressing this issue would require the integration of external prosopographical resources or authority control mechanisms, which lies beyond the scope of the present study.

3. RESULTS

This section presents the findings, focusing on gender-based patterns in property ownership. The results are organized along three complementary dimensions: the distribution of property types associated with female landowners, a comparative analysis of female and male ownership patterns, and an aggregation of these patterns into macro-categories of urban space. The most frequent property type owned by women is *casa* (622 parcels) followed by *casa e corte* (194). The two categories are the majority of female-owned properties indicating a strong concentration of (Figure 1, Col A). Non-residential spaces, such as *bottega* (shop) and *casa e bottega* (house and shop), appear only sporadically in this subset, with fewer than 15 occurrences each. This distribution suggests that, while women were present as urban property holders, their ownership was predominantly concentrated in private domestic architecture rather than in spaces explicitly associated with commercial or productive activity, as also shown in the zoom-in for the *attività* (busyness) macro category. The structure of property ownership further differentiates women's presence in the cadastral system, with a clear predominance of exclusive ownership over co-ownership (Figure 1, Col B). A total of 309 property types appear exclusively under female ownership, compared to 181 appearing only in co-ownership arrangements, with a small overlap of 21 property types shared between the two categories. While the dataset does not allow for a full reconstruction of the legal mechanisms underlying these arrangements, the prevalence of exclusive ownership challenges simplistic assumptions that women's property was primarily mediated through male relatives or shared legal structures.

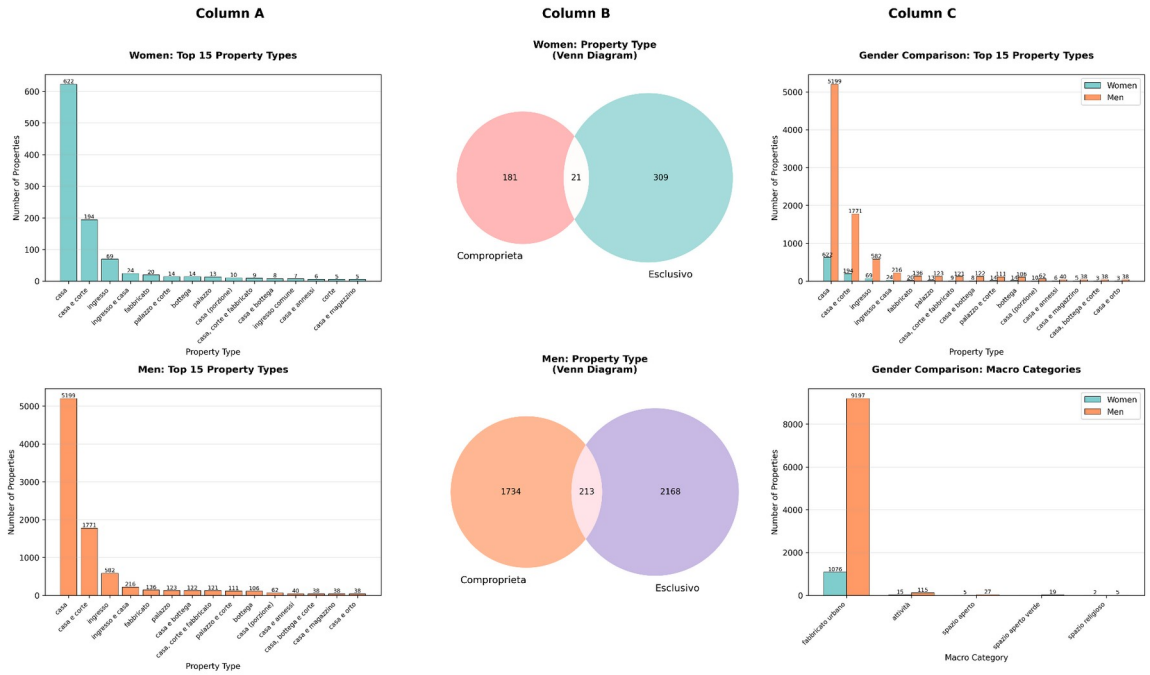


Figure 1. The charts describe the visualization results on the Florentia Illustrata Knowledge Graph. Column A, shows the distribution of the top 15 properties per woman (top) and men (bottom). Column B displays Venn Diagrams for the two specializations of property holding: Comproprietà and Esclusivo for women (top) and men (bottom). Column C shows the comparison of distribution for the top 15 categories (top) and the full set of properties aggregated by macro category (bottom).

To abstract from individual property labels and enable an aggregated comparison, property types were grouped into macro-categories of urban space (Figure 1, Col C). The result reveals that both women and men mainly owned properties classified as *fabbricato urbano* (urban building). However, while women are associated with approximately 1100 such parcels, men are associated with more than 9000, reinforcing the asymmetry observed at the level of individual property types. Other macro-categories (such as

attività, *spazio aperto*, *spazio aperto verde*, and *spazio religioso*) are almost exclusively owned by men, with female presence either minimal or absent.

4. CONCLUSIONS

The results demonstrate that women owned property in 19th-century Florence, as evidenced by their clear representation as property holders in the cadastral dataset. These properties are specifically clustered around typologies belonging to residential properties. A similar pattern can be seen for properties owned by men, but with substantially higher absolute numbers.
By linking semantic querying to interactive spatial visualization, *Florentia Illustrata* makes it possible to turn competency questions into inspectable views: a query becomes a layer, a category becomes a dynamic legend, and an interpretative hypothesis can be explored by changing filters, labels, and colouring criteria. At the scale of the whole city, this supports rapid visual comparison across owner types and the identification of patterns and exceptions (Figure 2). At the parcel level, it enables close reading of single cases, combining cadastral metadata with recorded family relations and the ability to follow an individual's holdings across the map (Figure 3). This interplay between aggregated views and detailed inspection is what makes the results not only viewable but also interactively discussable.
At the same time, the case study makes clear that interpretative strength depends on data quality and identity resolution. Gender attribution inferred from indirect markers, and the persistence of homonymy, introduce uncertainty that should be handled carefully in historical claims. Future work could therefore focus on making such uncertainty more explicit within the graph and interface, and on integrating stronger identity management strategies. Even in its current form, *Florentia Illustrata* shows how a semantic architecture can support meaningful, user-driven interaction with cadastral sources, helping researchers move between questions, evidence, and spatialized results without reducing interpretation to a purely tabular exercise.

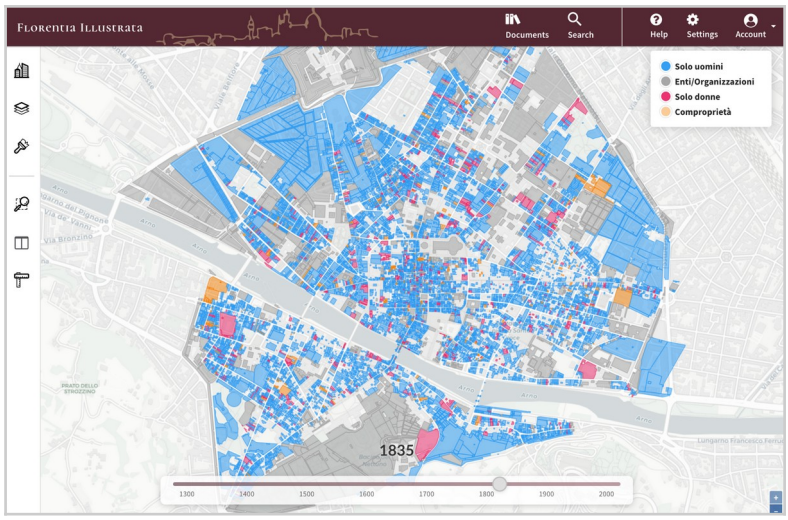


Figure 2. Overall view of 1835 Florence using the Semantic Map Advanced component, with a dynamic map legend coding that divides properties by their owners gender. In some cases - especially in the case of siblings - there are multiple proprietaries including both. Institutions are in grey.

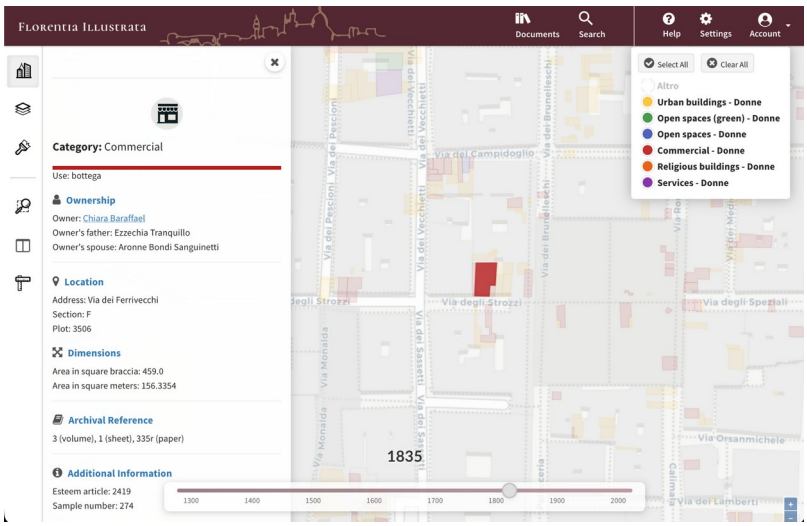


Figure 3. Detail of a commercial woman-owned property in the center of Florence, in the context of all property types of other women in the same area. Florentia Illustrata allows to inspect cadastral metadata and highlight people relations, as well as filtering all properties of the same person.

ACKNOWLEDGEMENTS

The project Florentia Illustrata has been possible with the supervision of Lukas Klic, head of Digital Humanities Research at I Tatti, the Harvard University Center for Italian Renaissance Studies. We wish to thank Gianluca Belli, Fabio Lucchesi, and Paola Raggi for making the underlying cadastral dataset openly available (Belli et al., 2022).

REFERENCES

Belli, G., Lucchesi, F., & Raggi, P. (2022). Firenze nella prima metà dell'Ottocento: La città nei documenti del Catasto Generale Toscano. Firenze University Press. https://doi.org/10.36253/979-12-215-0002-8

Doerr, M., Ore, C.-E., & Stead, S. (2007). The CIDOC Conceptual Reference Model: A new standard for knowledge sharing. In Proceedings of the International Conference on Dublin Core and Metadata Applications (pp. 51–56). Dublin Core Metadata Initiative. https://doi.org/10.13140/2.1.1420.6400

Galeazzo, L., Grillo, R., Spinaci, G., *et al.* (2024). A geospatial and time-based reconstruction of the Venetian Lagoon in a 3D web semantic infrastructure. In CEUR Workshop Proceedings (Vol. 3643, pp. 212–225). CEUR-WS.

Oldman, D. (2021). Digital research, the legacy of form and structure, and the ResearchSpace system. In K. Golub & Y.-H. Liu (Eds.), Information and knowledge organisation in digital humanities: Global perspectives (pp. 131–153). Routledge.

Oldman, D., & Tanase, D. (2018). Reshaping the Knowledge Graph by Connecting Researchers, Data and Practices in ResearchSpace. In D. Vrandečić, K. Bontcheva, M. C. Suárez-Figueroa, V. Presutti, I. Celino, M. Sabou, L.-A. Kaffee, & E. Simperl (Eds.), The Semantic Web – ISWC 2018 (Vol. 11137, pp. 325–340). Springer International Publishing. https://doi.org/10.1007/978-3-030-00668-6_20

Per la storia del mercato dell'arte tra otto e novecento: raccolta, modellazione e valorizzazione dei dati sugli antiquari italiani

VALENTINA ROSSETTI[1], MARILENA DAQUINO[1], FRANCESCA MAMBELLI[2], LUDOVICA PANNITTO[1], FRANCESCA TOMASI[1]

[1]UNIVERSITÀ DI BOLOGNA, ITALIA - VALENTINA.ROSSETTI@UNIBO.IT, MARILENA.DAQUINO2@UNIBO.IT, LUDOVICA.PANNITTO@UNIBO.IT, FRANCESCA.TOMASI@UNIBO.IT
[2]FONDAZIONE FEDERICO ZERI, UNIVERSITÀ DI BOLOGNA, ITALIA - FRANCESCA.MAMBELLI6@UNIBO.IT

ABSTRACT (ITALIANO)

Il mercato antiquario italiano tra Otto e Novecento è un settore della storia economica e culturale divenuto di grande interesse nell'ultimo decennio, seppur ancora poco studiato. Questo contributo affronta le sfide metodologiche della ricostruzione prosopografica (resa complessa da alcune caratteristiche intrinseche di questo ambito: dispersione ed eterogeneità delle fonti, natura informale delle pratiche commerciali, complessità delle reti professionali e familiari), e quelle di progettazione di un flusso di gestione e pubblicazione dei dati sostenibile nel tempo. La metodologia implementata si articola in quattro fasi: 1. raccolta e normalizzazione delle fonti narrative in linguaggio naturale; 2. modellazione dei dati; 3. sviluppo di un'architettura tecnologica sostenibile; 4. progettazione di una piattaforma informativa multi-vista che implementa i principi della "rich-prospect browsing", bilanciando panoramica e dettaglio attraverso sei interfacce navigabili.
Parole chiave: Storia dell'arte digitale; Gestione dati; Narrazione digitale; Visualizzazione dell'informazione; Mercato dell'arte

ABSTRACT (ENGLISH)

For the art market history between the nineteenth and twentieth centuries: collection, modeling, and valorization of data on italian antiquaries. The Italian antiques market between the 19th and 20th centuries is an area of economic and cultural history that has become of great interest in the last decade, although it remains relatively unexplored. This contribution addresses the methodological challenges of prosopographic reconstruction (made complex by certain intrinsic characteristics of this field: dispersion and heterogeneity of sources, informal nature of commercial practices, complexity of professional and family networks), and those of designing a sustainable data management and publication flow over time. The methodology implemented consists of four phases: 1. collection and normalization of narrative sources in natural language; 2. data modeling; 3. development of a sustainable technological architecture; 4. design of a multi-view information platform that implements the principles of "rich-prospect browsing," balancing overview and detail through six navigable interfaces.
Keywords: Digital art history; Data management; Digital storytelling; Data visualization; Art market

1. INTRODUZIONE

La ricostruzione della storia del mercato antiquario italiano tra Otto e Novecento presenta sfide metodologiche significative. Sono numerosi gli studi che hanno approfondito le figure dei professionisti, delle gallerie più famose o che si sono concentrati sulle grandi vendite all'asta (De Benedictis, 1998; Morandotti, 2008; Rebora, 2020), spesso sottovalutando il ruolo svolto dagli antiquari 'minori' nella circolazione delle opere d'arte, nella definizione del canone e del gusto, e nella formazione di collezioni pubbliche e private. La complessità che caratterizza il settore è dovuta principalmente a: 1) meccanismi di funzionamento economico e peculiarità delle pratiche commerciali - come ad esempio la natura riservata delle transizioni, l'irreperibilità della documentazione contabile, l'ambiguità nelle regole del commercio transnazionale; 2) complessità delle entità e delle reti professionali, con presenza di dinastie familiari, di forme di associazionismo atipico o temporaneo, di collaborazioni verticali tra professionalità diverse (antiquari-restauratori-periti-fotografi); 3) dispersione e frammentarietà delle fonti costituite da materiali documentari eterogenei, come cataloghi d'asta, fotografie, articoli e pubblicità su quotidiani o

periodici, documentazione commerciale, autobiografie 'romanzate' degli antiquari stessi (Jandolo, 1935), carte private, testimonianze orali non sempre attendibili.
Il progetto presentato in questo contributo nasce all'interno della Fondazione Federico Zeri, il cui archivio fotografico comprende fondi di particolare rilevanza per lo studio del mercato antiquario nell'area e nell'arco cronologico individuato. Con il supporto dell'Associazione Amici di Federico Zeri, la Fondazione ha promosso, a partire dal 2022, una serie di attività di ricerca svolte da borsisti e da personale interno, i cui esiti sono stati la base di questo lavoro. L'indagine ha preso avvio dai materiali presenti nel fondo Zeri (libri, cataloghi d'asta, fotografie spesso annotate da Zeri o da studiosi, collezionisti, mercanti) allargandosi poi alle riviste periodiche specializzate, ad altre fonti bibliografiche e a testimonianze orali raccolte per mezzo di interviste a eredi di antiquari e a professionisti ancora attivi. La natura narrativa e frammentaria dei dati raccolti ha posto immediatamente il problema della loro strutturazione e accessibilità. Le informazioni, contenenti dati biografici, geografici, cronologici e relazionali, sono state collezionate in linguaggio naturale - con parametri non standardizzati - all'interno di file di testo.
Trasformare questo corpus in un dataset interrogabile, visualizzabile e analizzabile, con lo scopo di realizzare percorsi narrativi di interesse sia per esperti di dominio sia per il pubblico generico, ha richiesto un intervento di co-progettazione da parte di storici dell'arte e digital humanists secondo una metodologia esposta di seguito.
Questo contributo ha quindi due obiettivi principali: 1) presentare un flusso di lavoro sostenibile per la raccolta, gestione e pubblicazione di dati che continuano ad emergere dalla ricerca, proponendo un approccio flessibile, scalabile e replicabile per progetti di digital humanities che affrontano sfide simili; 2) ideare un percorso narrativo che permetta di visualizzare e analizzare le reti professionali, i passaggi generazionali, l'andamento cronologico e geografico del mercato antiquario italiano

2. STATO DELL'ARTE

Esistono diversi progetti di prosopografia digitale - ovvero lo studio sistematico di biografie collettive in un contesto storico definito utilizzando metodologie computazionali - che hanno trovato applicazioni significative nella storia del patrimonio culturale, permettendo di ricostruire reti professionali, sociali e intellettuali attraverso l'analisi di dataset strutturati. Il Bloomsbury Art Market[1] (Nathan, 2023) rappresenta uno dei più ampi sforzi in questa direzione, permettendo l'accesso a oltre 4.000 articoli su gallerie, case d'asta e mercanti attivi in Europa e Nord America dal 1900 ad oggi. Il progetto, precedentemente noto come "Art Market Dictionary", ha coinvolto un team internazionale di oltre 550 autori che hanno sviluppato un database che integra dati fattuali (anni, indirizzi, specializzazioni) con narrazioni storiche, informazioni su mostre, clienti, artisti trattati e bibliografie. L'interfaccia permette ricerche avanzate per contenuto, luogo, periodo e movimenti artistici, ma privilegia un approccio enciclopedico-consultivo per utenti esperti, senza percorsi narrativi guidati o visualizzazioni interattive per pubblici non specialistici. Il modello commerciale (accesso istituzionale a pagamento) lo posiziona come risorsa accademica piuttosto che strumento di public engagement aperto.
Di grande interesse per la storia del mercato è anche il Getty Provenance Index[2] (Davis, 2019; Getty Research Institute, 2024a) che include oltre due milioni di record relativi a transazioni estratte da registri commerciali di gallerie storiche, inventari di collezioni private, cataloghi d'asta e dossier su collezionisti, case d'asta e istituzioni. Il database permette di tracciare la provenienza di opere d'arte e di identificare autori e collezionisti attraverso ricerche per nome, luogo e periodo. Il progetto rimane comunque incentrato sulle opere commercializzate piuttosto che sulle reti professionali del mercato antiquariale, con informazioni biografiche sui mercanti che rimangono frammentate e incidentali.
Il RKDartists (Netherlands Institute for Art History)[3] ha sviluppato uno dei più importanti database prosopografici in ambito artistico, con oltre 350.000 record su artisti, collezionisti e altre figure legate all'arte olandese e fiamminga, ma toccando solo marginalmente la figura dell'antiquario (RKD, 2024). I dati sono pubblicati come Linked Open Data, ma attorno ad essi non è costruita nessuna narrazione guidata che faciliti la comprensione del fenomeno del mercato dell'arte per un pubblico più ampio. Anche PHAROS - The International Association of Photo Archives[4] (Caraffa et al., 2020; PHAROS, 2024), che

[1] https://www.bloomsbury.com/us/ (cons. 07/02/2026)
[2] https://www.getty.edu/databases-tools-and-technologies/provenance/ (cons. 07/02/2026)
[3] https://research.rkd.nl/nl (cons. 07/02/2026)
[4] https://artresearch.net/resource/start (cons. 07/02/2026)

aggrega metadati su opere d'arte e fotografie nella piattaforma artresearch.net, documenta l'attività di fotografi e artisti, ma non struttura sistematicamente informazioni biografiche su mercanti e antiquari. Esistono ulteriori iniziative che rendono disponibili materiali utili per lo studio del mercato dell'arte, anche se non strutturati prosopograficamente, come la Archives Directory for the History of Collecting in America[5], incentrata sui collezionisti americani (Frick Art Reference Library, 2024) o il database del Department of Image Collections[6] della National Gallery of Art, che conserva gli archivi fotografici di importanti mercanti dell'arte e gallerie storiche (National Gallery of Art, 2024).
L'analisi della letteratura rivela quindi la sostanziale assenza di progetti prosopografici dedicati agli antiquari italiani in epoca storica. I progetti esistenti si concentrano più in generale sulla provenance delle opere e sul collezionismo, prediligendo forme di pubblicazione dei dati pensate per utenti esperti (catalogo con filtri di ricerca avanzati, query sparql), rinunciando a percorsi interpretativi in grado di contestualizzare storicamente il fenomeno e utilizzare tecniche di data visualization per veicolare informazioni complesse ad un pubblico più ampio.
La progettazione di interfacce per collezioni digitali simili rappresenta una sfida specifica per le Digital Humanities. La letteratura identifica approcci diversi alla presentazione dell'informazione, con implicazioni sia cognitive che interpretative. In particolare, sono state individuate due strategie contrapposte per la presentazione di collezioni digitali, ciascuna con pro e contro (Daquino, 2021; Moretti et al., 2024; Renda, 2025). Il primo approccio riguarda il paradigma della riduzione del carico cognitivo, che si basa sul principio di frammentare l'informazione in chunks piccoli e gestibili (Krug. 2014). Secondo questo approccio le interfacce dovrebbero essere minimaliste, con una singola barra di ricerca, risultati paginati e dettagli mostrati su richiesta esplicita. Se da un lato questo metodo riduce il sovraccarico cognitivo inziale favorendo l'attenzione su singole risorse, dall'altro l'utente non ha una rappresentazione d'insieme, con conseguente disorientamento e difficoltà nella navigazione (Whitelaw, 2015). Il secondo approccio si basa invece su un'interfaccia generosa, che mostra l'intera collezione – o un subset significativo - fin dall'inizio, offrendo una visualizzazione d'insieme navigabile con filtri di ricerca. Questa soluzione ha il vantaggio di supportare la scoperta casuale della collezione, per cui l'utente ne comprende immediatamente dimensione e natura, ma il rischio è quello di sopraffarlo con grandi quantità di dati.
Tuttavia, i visitatori di siti culturali non sempre hanno un obiettivo specifico, ma praticano una ricerca esplorativa, per cui nascondere il quadro d'insieme può frustrare anziché aiutare. Un terzo approccio, il Rich-prospect browsing, sintetizza i due procedenti strutturando l'esplorazione in 4 fasi principali (Ruecker et al., 2011; Shneiderman, 1996): (1) Overview first – panoramica dell'intera collezione che fornisce orientamento spaziale, cronologico o tematico; (2) Zoom and filter – l'esplorazione viene raffinata progressivamente per mezzo di filtri; (3) Details on demand – si può accedere a schede di dettaglio solo dopo una selezione consapevole dell'oggetto di interesse; (4) Multiple entry points – offre diversi percorsi di esplorazione (cronologico, geografico, tematico, alfabetico) che si adattano a diverse domande di ricerca. Questo approccio supporta la scoperta e la creazione di relazioni inedite tra le risorse culturali, evidenzia pattern nascosti e favorisce l'esplorazione libera dei dati.

3. METODOLOGIA

Trasformare fonti discorsive frammentarie in un dataset strutturato e interrogabile ha richiesto l'applicazione di un processo metodologico end-to-end, che si è articolato in molteplici fasi. Di seguito evidenziamo le sfide progettuali che hanno richiesto decisioni di modellazione e strategie per la definizione di flusso di lavoro sostenibile, che riguarda raccolta e pubblicazione dei dati in autonomia senza l'intervento del personale tecnico, nonché per l'esplorazione e l'analisi dei dati che risponda alle esigenze di ricerca di esperti e del pubblico generalista.

- Dispersione fonti. Integrazione di fonti eterogenee in un dataset coerente e interrogabile.
- Informalità del settore. Modello dati flessibile per gestire la diversa granularità delle informazioni.
- Complessità delle reti. Modello relazionale esplicito, visualizzazioni specifiche (albero, network).
- Memoria orale a rischio. Public engagement con interviste dirette.
- Fonti narrative. Normalizzazione manuale e dialogo tra esperti tecnici e umanistici.
- Sostenibilità. Workflow sostenibile, sito web statico, autonomia operativa per gli storici dell'arte.

[5] https://research.frick.org/directory (cons. 07/02/2026)
[6] https://www.nga.gov/research/library/image-collections/art-dealers-archives (cons. 07/02/2026)

Il lavoro qui presentato ha preso in esame le fonti raccolte dagli esperti della Fondazione Federico Zeri, che hanno consultato: a) gli archivi commerciali e le fotografie della fototeca Zeri, con informazioni implicite su opere, attribuzioni, prezzi, corrispondenze tra antiquari, annotazioni varie; b) lo spoglio completo del periodico "L'Antiquario" (1908-1932), prima rivista italiana dedicata al settore, contente annunci commerciali, resoconti di vendite, necrologici, articoli su collezioni e mostre, pubblicità di gallerie; c) bibliografia d'interesse dispersa in più luoghi, tra cui cataloghi d'asta, monografie, atti di convegni o articoli in riviste specialistiche; d) decine di eredi di famiglie antiquarie storiche e antiquari attivi nel secondo dopoguerra, attivando un processo di public engagement con interviste dirette.
La seconda fase ha riguardato la normalizzazione e modellazione dei dati, in stretta collaborazione tra storici dell'arte e digital humanist. Il processo, partendo dal lavoro ontologico ZAMO (Zeri Art Market Ontology) sviluppato presso la Fondazione Zeri (Veggi & Mambelli, 2023; Veggi, 2024), ha permesso di identificare le principali entità e le loro relazioni e di strutturare i dati in tabelle secondo un sistema di annotazione in grado di gestire la diversa granularità delle informazioni. Le tabelle, che includono i metadati fondamentali necessari a identificare la figura degli antiquari, il contesto in cui hanno operato e le relazioni intrattenute durante la loro carriera, sono state poi trasformate in un database relazionale. Infine, sono state semplificate in un template di lavoro riutilizzabile dagli storici dell'arte per proseguire l'attività in autonomia nelle successive fasi di arricchimento dei dati.
La terza fase operativa è stata incentrata sulla definizione dell'ambiente web, che ha richiesto scelte architetturali ponderate per garantire non solo qualità scientifica, ma anche sostenibilità a lungo termine e autonomia operativa per gli storici dell'arte. Sulla scia di altri casi di successo presenti in letteratura [Renda & Daquino, 2025], si è scelto di progettare un sito web statico sulla piattaforma GitHub, uno degli ambienti più noti per depositare e versionare codice open source, prediligendo una soluzione low cost, con tecnologia standard e portabile. Il repository GitHub del progetto è disponibile online[7]. Per massimizzare l'autonomia degli storici dell'arte, è stata pensata una pipeline di trasformazione dei dati TSV → JSON → HTML. Il flusso di lavoro inizia con l'esportazione dei documenti TSV e il caricamento sul repository GitHub, insieme ai file di testo e alle immagini. Il sistema rileva automaticamente i *commit* con etichetta "data update" e attiva le GitHub Actions. Gli utenti non esperti possono caricare i dati utilizzando l'interfaccia drag-and-drop user-friendly di GitHub. Uno script Python (build_data.py) legge i file TSV, validando l'integrità dei dati verificando che gli ID siano univoci e i riferimenti corretti, generando dei file JSON necessari per ogni vista. Vengono prodotti file specifici per ciascuna interfaccia (e.g. antiquari.json, luoghi.json). GitHub Pages pubblica il sito aggiornato senza ulteriori interventi manuali. Il vantaggio principale di questo sistema è che, una volta configurato, il processo è completamente automatico e gli storici dell'arte non devono mai interagire direttamente con il codice.
Il sito web, statico, non richiede elaborazione server-side e può quindi essere ospitato gratuitamente su GitHub Pages. Per il front-end, invece, è stato scelto JavaScript vanilla, minimizzando le dipendenze da framework esterni e garantendo longevità del codice. Per realizzare viste narrative specifiche, sono state invece selezionate librerie JavaScript specificamente ottimizzate per ciascun tipo di visualizzazione. La vista Mappa utilizza Leaflet.js con il plugin MarkerCluster per gestire efficacemente la densità dei punti geografici. La Timeline è stata implementata con HTML e CSS personalizzati per garantire massima flessibilità nella presentazione cronologica degli eventi. Gli alberi genealogici utilizzano amCharts 5 ForceDirectedTree, una libreria specializzata per visualizzazioni gerarchiche interattive. Per il Network è stato scelto D3.js con algoritmi force-directed per il layout automatico del grafo. La funzionalità di ricerca combina jQuery per il filtering e JavaScript vanilla per il rendering dei risultati.
La quarta fase operativa ha riguardato il processo di pubblicazione dei dati, iniziato con la progettazione dall'architettura dell'informazione dell'applicazione web con l'obiettivo di rispondere alle principali domande di competenza[8]: 1) Chi erano gli antiquari attivi in Italia tra Otto e Novecento? 2) Dove operavano gli antiquari e quali erano i principali luoghi del mercato dell'arte? 3) Come si è evoluta la geografia del mercato nel tempo? 4) Quali erano le reti familiari e professionali, chi collaborava con chi? 5) Quali sono stati i momenti chiave nella storia del mercato antiquario? 6) Quali sono le fonti che hanno permesso la ricostruzione storica proposta?
Da questa riflessione sono state isolate sei entità principali da cui è derivata la progettazione di un'architettura dell'informazione multi-vista, che riflette consapevolmente i principi della rich-prospect bowsing (Ruecker, 2011), bilanciando le esigenze di panoramica e dettaglio per rispondere ai bisogni di

[7] https://github.com/FondazioneFedericoZeri/Mercato_dell_arte/tree/main (cons. 07/02/2026)
[8] https://github.com/FondazioneFedericoZeri/Mercato_dell_arte (cons. 07/02/2026)

un target sia esperto, sia generalista. È stato quindi pensato un percorso narrativo guidato che inizia da: una sezione introduttiva (Homepage), una vista dedicata a fornire il framework interpretativo (Progetto), una pagina di ricerca degli antiquari che rimanda a schede di dettaglio (Antiquari), una mappa geografica con filtri temporali (Luoghi), una sezione dedicata agli eventi principali del periodo di riferimento (Eventi), una pagina dedicata alla rete antiquariale (Persone) e una sezione dedicata alle fonti (Bibliografia). Il sito web è pubblicato online[9].

4. DISCUSSIONE E CONCLUSIONI

Il progetto ha prodotto risultati significativi su diversi livelli interconnessi: storiografico e di preservazione della memoria, metodologico, legato all'accessibilità. Dal punto di vista storiografico, la piattaforma fornisce la prima mappatura prosopografica sistematica del mercato antiquario italiano, documentando 71 entità antiquariali, più di 250 persone, 153 luoghi geo-referenziati, 33 eventi significativi e centinaia relazioni professionali esplicite, colmando una lacuna nella letteratura.
Sul piano della preservazione della memoria, è stata attivata una campagna di public engagement con impatto tangibile. L'approccio ha privilegiato il dialogo diretto con decine di eredi di famiglie antiquarie e professionisti del secondo dopoguerra, la cui memoria copre quasi un secolo di storia del settore ed è a rischio imminente di scomparsa. Questa azione ha generato un senso di fiducia in alcuni stakeholder, che ha portato alla donazione nel 2024 dell'archivio fotografico Sestieri a cui è seguita la donazione del fondo Zabert. Il caso dimostra come la ricerca accademica, quando condotta con sensibilità verso le persone e le loro storie familiari, possa attivare processi di salvaguardia del patrimonio culturale.
Dal punto di vista metodologico, la necessità di progettare un flusso di lavoro sostenibile ha portato a valutare diverse alternative tecnologiche. I CMS tradizionali come Omeka e WordPress sono stati scartati perché richiedono server dedicati, database SQL, manutenzione continua e comportano costi di hosting ricorrenti che potrebbero risultare insostenibili nel lungo periodo per un'istituzione culturale di media dimensione (Wikle & Williamson, 2024; Wikle & Williamson, 2025). Analogamente, i framework JavaScript moderni come React, Vue o Angular presentano processi di build complessi, numerose dipendenze da gestire e rischiano l'obsolescenza rapida, rendendo problematica la manutenzione futura del progetto. Anche l'utilizzo di un database SQL è stato escluso perché avrebbe richiesto un server dedicato, procedure di backup e costi aggiuntivi (Williamson et al., 2021). La scelta, invece, di tecnologie standard, gratuite e ampiamente diffuse (GitHub Pages, file TSV/JSON, HTML/CSS/JavaScript vanilla) risponde alla necessità di gestire piccoli dataset e creare progetti DH sostenibili nel lungo periodo per istituzioni culturali che non dispongono di team tecnici permanenti o alti budget. La progettazione di template user-friendly in cui inserire nuovi dati e il drag-and-drop delle tabelle in GitHub ha permesso agli storici dell'arte della Fondazione Zeri di operare in autonomia senza la necessità di essere supervisionati.
La piattaforma web ha infine permesso di rendere accessibile a un pubblico ampio (studiosi, studenti, collezionisti, mercato dell'arte, pubblico curioso) un corpus di informazioni precedentemente disperse e di difficile consultazione. Le 6 viste navigabili permettono diversi livelli di approfondimento: dalla scoperta casuale (navigazione geografica su mappa) all'analisi specialistica (scheda di dettaglio), dalla contestualizzazione storica (timeline eventi) alla ricerca bibliografica.
In conclusione, questo contributo mostra un modello replicabile per progetti simili: fonti narrative frammentate, dataset medio-piccoli, budget limitati, personale non tecnico. L'applicazione web risultante, strutturata come un percorso esplorativo con elementi di storytelling digitale, bilancia con successo panoramica ed esplorazione di dettagliato. Il progetto è concepito come infrastruttura aperta e dinamica: dataset, codice e documentazione sono pubblicamente accessibili su GitHub, permettendo a studiosi, studenti e appassionati di contribuire, riutilizzare e adattare il lavoro per altri contesti. Tale apertura è coerente con i principi dell'Open Science e massimizza l'impatto a lungo termine della ricerca. I lavori futuri riguarderanno la modellazione ontologica dei dati riutilizzando CIDOC-CRM e Linked Art, oltre alla riconciliazione delle persone con authority files esistenti. È altresì prevista un'estensione della ricerca d'archivio, per cui il set di dati verrà arricchito e reso interoperabile con un'applicazione web specifica per i cataloghi d'asta, in corso di realizzazione sempre nell'ambito delle attività di ricerca della Fondazione Federico Zeri.

[9] https://fondazionefedericozeri.github.io/Mercato_dell_arte/index.html (cons. 07/02/2026)

RINGRAZIAMENTI

Fondazione Federico Zeri, Associazione Amici di Federico Zeri.

BIBLIOGRAFIA

Caraffa, C., Pugh, E., Stuber, T., & Ruby, L. W. (2020). PHAROS: A digital research space for photo archives. Art Libraries Journal, 45(1), 2–11. https://doi.org/10.1017/alj.2019.34

Daquino, M. (2021). Linked Open Data native cataloguing and archival description. JLIS.it, 12(3), 91–104. https://doi.org/10.4403/jlis.it-12703

Davis, K. (2019). Old metadata in a new world: Standardizing the Getty Provenance Index for linked data. Art Libraries Journal, 44(4), 162–166. https://doi.org/10.1017/alj.2019.24

De Benedictis, C. (Eds.). (1998). Per la storia del collezionismo italiano: Fonti e documenti (2a ed.). Milano: Ponte alle Grazie. ISBN: 9788868332860

Frick Art Reference Library. (2024). Archives Directory for the History of Collecting in America [Database]. https://research.frick.org/directory

Getty Research Institute. (2024a). Getty Provenance Index Initiative. https://www.getty.edu/projects/getty-provenance-index-initiative/

Jandolo, A. (1935). Le memorie di un antiquario. Milano: Ceschina.

Krug, S. (Eds.). (2014). Don't make me think, revisited: A common sense approach to web usability (3rd ed.). New Riders. ISBN: 139780321965516

Morandotti, A. (Eds.). (2008). Il collezionismo in Lombardia: Studi e ricerche tra '600 e '800. Officina Libraria. ISBN: 9788889854037

Moretti, A., Heibi, I., & Peroni, S. (2024). A workflow for GLAM metadata crosswalk. arXiv. https://doi.org/10.48550/arXiv.2405.02113

Nathan, J. (Eds.). (2023). Bloomsbury Art Markets: Protagonists, Networks, Provenances. Bloomsbury Digital Resources. https://www.bloomsburyvisualarts.com/bloomsbury-art-markets

National Gallery of Art. (2024). Department of Image Collections: Art dealers' archives [Archival collection]. https://www.nga.gov/research/library/image-collections/art-dealers-archives

PHAROS: The International Association of Photo Archives. (2024). Artresearch.net [Digital research platform]. https://artresearch.net

Rebora, S. (Eds.). (2020). Nel salotto del collezionista: Arte e mecenatismo tra Otto e Novecento [Catalogo della mostra]. Silvana Editoriale. ISBN: 9788836646128

Renda, G., & Daquino, M. (2025). Ragù: Preserving Italian cookbooks in a sustainable crowdsourced digital collection. Umanistica Digitale, 9(20), 405–418. https://doi.org/10.6092/issn.2532-8816/21113

RKD – Netherlands Institute for Art History. (2024). RKDartists [Database]. https://rkd.nl/en/explore/artists

Ruecker, S., Radzikowska, M., & Sinclair, S. (Eds.). (2011). Visual interface design for digital cultural heritage: A guide to rich-prospect browsing. Routledge. ISBN: 9781138250307

Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. Proceedings of the IEEE Symposium on Visual Languages, 336–343. https://doi.org/10.1109/VL.1996.545307

Veggi, M. (2024). A first ontological model for the description of the art market in the semantic web. arXiv. https://doi.org/10.48550/arXiv.2404.00395

Veggi, M., & Mambelli, F. (2023). Modelling the art market in the semantic web: A preliminary analysis. Umanistica Digitale, 7(16), 141–166. https://doi.org/10.6092/issn.2532-8816/17208

Whitelaw, M. (2015). Generous interfaces for digital cultural collections. Digital Humanities Quarterly, 9(1). http://www.digitalhumanities.org/dhq/vol/9/1/000205/000205.html

Wikle, O. M., & Williamson, E. P. (2024). Exploring static web in the Digital Humanities classroom: The Learn-Static initiative. IDEAH, 4(2). https://doi.org/10.21428/f1f23564.f88a989c

Wikle, O., & Williamson, E. P. (2025). Static web methodology as a sustainable approach to Digital Humanities projects. The Code4Lib Journal, 57. https://journal.code4lib.org/articles/18372

Williamson, E., Wikle, O., Becker, D., Seiferle-Valencia, M., Doney, J., & Martinez, J. (2021). Using static web technologies and Git-based workflows to re-design and maintain a library website (quickly) with non-technical staff. College & Undergraduate Libraries, 28(3), 216–239. https://doi.org/10.1080/10691316.2021.1960973

Un grafo per la ricerca storico-filologica in archivio: il caso del fondo Giuseppe Albini

LUCIA GIAGNOLINI[1], MOHAMED IHEB OUERGHI[2], VALENTINA PASQUAL[3], CECILIA TAMAGNINI[4], FRANCESCA TOMASI[5]

[1] DIPARTIMENTO DI FILOLOGIA CLASSICA E ITALIANISTICA, ALMA MATER STUDIORUM - UNIVERSITÀ DI BOLOGNA; ARCHIVIO STORICO, ALMA MATER STUDIORUM - UNIVERSITÀ DI BOLOGNA, LUCIA.GIAGNOLINI2@UNIBO.IT
[2] DIPARTIMENTO DI FILOLOGIA CLASSICA E ITALIANISTICA, ALMA MATER STUDIORUM - UNIVERSITÀ DI BOLOGNA, MOHAMEDIHEB.OUERGHI@STUDIO.UNIBO.IT
[3] DIPARTIMENTO DI FILOLOGIA CLASSICA E ITALIANISTICA, ALMA MATER STUDIORUM - UNIVERSITÀ DI BOLOGNA, VALENTINA.PASQUAL2@UNIBO.IT
[4] DIPARTIMENTO DI FILOLOGIA CLASSICA E ITALIANISTICA, ALMA MATER STUDIORUM - UNIVERSITÀ DI BOLOGNA, CECILIA.TAMAGNINI7@UNIBO.IT
[5] DIPARTIMENTO DI FILOLOGIA CLASSICA E ITALIANISTICA, ALMA MATER STUDIORUM - UNIVERSITÀ DI BOLOGNA, FRANCESCA.TOMASI@UNIBO.IT

ABSTRACT (INGLESE)

A Graph for Historical-Philological Archival Research: The Case of the Giuseppe Albini Collection. This paper presents a project dedicated to the Giuseppe Albini Archive, conceived to investigate the interplay between humanistic culture, academic institutions, and political power in early twentieth-century Italy. Albini (1863–1933) - classical philologist, professor, Rector of the University of Bologna, and Senator of the Italian Kingdom - offers a privileged point of view to observe processes of cultural and institutional legitimation within the academic world between the liberal era and the Fascist regime. The paper focuses on the description of the first phase of the project, devoted to the semantic modelling of the archival inventory and its transformation into an interoperable *knowledge graph*. This phase has been implemented through the adoption of Semantic Web ontologies (RiC-O and CIDOC CRM) and the development of a Python-based Extract, Transform, Load (ETL) pipeline aimed at migrating the archival data into Linked Data. The contribution outlines the ontological choices, mapping strategies, and pipeline architecture, showing how this initial phase provides the foundation for subsequent stages of the project, oriented toward collaborative semantic annotation, computational analysis of correspondence networks, and the visualization of the spatio-temporal dynamics of intellectual networks.
Keywords: Linked Open Data; Network Analysis; Archival Studies; Ontologies

ABSTRACT (ITALIANO)

Il contributo presenta un progetto dedicato al Fondo Giuseppe Albini, concepito per analizzare l'intreccio tra cultura umanistica, istituzioni accademiche e potere politico nell'Italia del primo Novecento. La figura di Albini (1863–1933), filologo classico, docente, Rettore dell'Università di Bologna e Senatore del Regno, costituisce un punto di vista privilegiato per osservare i processi di legittimazione culturale e istituzionale che caratterizzarono il mondo accademico tra età liberale e regime. Il paper si concentra sulla descrizione della prima fase del progetto, dedicata alla modellazione dell'inventario archivistico del fondo e alla sua trasformazione in *knowledge graph*. Tale fase è stata realizzata mediante l'adozione di modelli ontologici canonici (RiC-O e CIDOC CRM) e lo sviluppo di una pipeline *Extract, Transform, Load* (ETL) in Python, finalizzata alla migrazione in Linked Data. Il contributo illustra le scelte ontologiche, le strategie di mapping e l'architettura della pipeline, azioni necessarie per sviluppi successivi del progetto, orientati all'annotazione semantica collaborativa, all'analisi computazionale delle reti di corrispondenza e alla visualizzazione delle dinamiche spazio-temporali delle reti intellettuali.
Parole chiave: Linked Open Data; Network Analysis; Archivistica; Ontologie

1. IL PROGETTO ALBINI

Come si sono intrecciati cultura classica, autorità accademica e potere politico nell'Italia del primo Novecento? In che modo le reti intellettuali europee si sono configurate e riconfigurate in questo periodo, e quali forme hanno assunto i rapporti tra sapere umanistico e contesto politico-istituzionale?

Queste domande richiedono un approccio che integri storia culturale, storia politica e storia delle reti intellettuali, capace di svelare non solo le produzioni scientifiche e letterarie, ma anche le pratiche relazionali, gli scambi epistolari, le carriere accademiche e i posizionamenti politici. Gli archivi personali degli studiosi rappresentano fonti privilegiate per questo tipo di indagine, poiché documentano simultaneamente la dimensione intellettuale, professionale e politica delle loro biografie, permettendo di ricostruire i circuiti attraverso cui circolavano idee, saperi e legittimazioni.
Giuseppe Albini (Bologna, 1863-1933) rappresenta una figura paradigmatica per affrontare queste tematiche. Allievo di Carducci, collega di Pascoli, fu docente ordinario di Filologia latina a Bologna (1897-1933), Preside di Facoltà (1912-1917) e Rettore (1927-1930). Traduttore di Virgilio, poeta neolatino pluripremiato, editore critico di Persio e Dante, la sua carriera accademica si intrecciò problematicamente con l'adesione al fascismo: nominato senatore nel 1924, iscritto al Partito Nazionale Fascista dal 1925, mantenne il rettorato in pieno regime, mettendosi anche a disposizione in qualità di rinomato epigrafista per celebrare il potere politico (Funaioli, 1935; Nastasi, 2019). Lo studio dell'archivio di Albini, personalità in grado di rappresentare quella generazione di intellettuali che intrecciò autorità culturale e supporto al regime fascista, permette di indagare criticamente le forme di relazione tra élite culturali italiane e autoritarismo.
Il Fondo Giuseppe Albini,[1] conservato presso il Dipartimento di Filologia Classica e Italianistica dell'Università di Bologna (FICLIT), comprende 42 buste (corrispondenza, appunti, minute filologiche, documenti amministrativi) e più di duecento volumi conservati dalla biblioteca del Dipartimento. Il carteggio, in particolare, documenta la vasta rete intellettuale europea di Albini, permettendo di ricostruire circuiti di scambio culturale, dibattito filologico, dinamiche accademiche e circolazione delle idee. Questa ricostruzione richiede tuttavia strumenti di ricerca adeguati alla complessità e alla ricchezza documentaria del fondo. In quest'ottica, il FICLIT ha avviato un percorso di valorizzazione nell'ambito del Progetto di Eccellenza del Dipartimento 2018-2022,[2] realizzando il recupero dell'elenco di consistenza e la digitalizzazione completa della corrispondenza e degli scritti latini (Fig. 1).



Figura 1. Esempio di lettera consultabile in un repository dipartimentale interno basato su Omeka S, esito delle operazioni digitalizzazione correlate al Progetto di Eccellenza del Dipartimento FICLIT 2018-2022.

Negli ultimi anni diversi progetti hanno sperimentato l'applicazione di tecnologie semantiche agli epistolari storici. Il progetto *Cultures of Knowledge* dell'Università di Oxford ha prodotto *Early Modern Letters Online* (EMLO)[3], un catalogo unificato delle corrispondenze erudite europee dei secoli XVI–XVIII che, attraverso un approccio *linked data*, permette di aggregare e interrogare dati distribuiti in archivi eterogenei, con strumenti di visualizzazione delle reti e analisi prosopografica (Tuominen et al., 2018). Il

[1] Si segnala anche il Carteggio Giuseppe Albini conservato presso l'Archivio Storico dell'Università di Bologna.
[2] Per maggiori informazioni sul riconoscimento MIUR attribuito al FICLIT: https://ficlit.unibo.it/it/dipartimento/premi-e-riconoscimenti/dipartimento-di-eccellenza (cons. 25/01/2026).
[3] https://emlo.bodleian.ox.ac.uk/

progetto finlandese *Constellations of Correspondence* (CoCo)[4] ha costruito un *knowledge graph* di oltre un milione di lettere, combinando raccolte epistolari e dati biografici per ricostruire reti di corrispondenza che vanno oltre il singolo fondo e i suoi corrispondenti diretti (Poikkimäki et al., 2025). Esperienze come l'edizione digitale della corrispondenza della dinastia Bernoulli e di Euler[5] hanno mostrato come la modellazione RDF consenta di facilitare interrogazioni complesse difficilmente realizzabili con strumenti convenzionali (Alassi & Rosenthaler, 2024). Sul versante delle edizioni digitali, il recente progetto *The Letters of Belle da Costa Greene to Bernard Berenson*,[6] sviluppato da I Tatti e dalla Morgan Library, offre trascrizioni annotate e metadatazione strutturata di quasi seicento lettere. In Italia, i progetti dedicati ai carteggi novecenteschi si sono finora mossi prevalentemente sul piano dell'edizione digitale e dell'analisi testuale: ne sono esempi l'Edizione Nazionale dell'epistolario di De Gasperi[7], che ha sviluppato una piattaforma di trascrizione collaborativa con metadati Dublin Core (Tonelli et al., 2019), e la sperimentazione di *topic modeling* e *sentiment analysis* sul carteggio Silone-Mondadori (Baldi, 2021). A differenza di questi progetti, il progetto Albini mira a costruire una struttura semantica che connetta le dimensioni intellettuale, biografica, istituzionale e politica di un singolo fondo, rendendolo interrogabile attraverso approcci storiografici diversi. Partendo dalle digitalizzazioni del carteggio e dall'inventario, il progetto Albini intende dotare il fondo di una struttura semantica che ne faciliti l'interrogazione attraverso diversi approcci, che vanno dalla ricerca di documenti e informazioni specifiche (entità qualificate) all'analisi delle reti di corrispondenti, dalla visualizzazione delle relazioni intellettuali alla mappatura geografica degli scambi. Il progetto intende procedere attraverso fasi successive che integrano modellazione ontologica, annotazione collaborativa delle immagini e sviluppo di strumenti di analisi e visualizzazione. La prima fase, dedicata alla trasformazione dell'inventario in un *knowledge graph* interrogabile e all'implementazione dell'infrastruttura di consultazione di base, è conclusa. Il presente contributo illustra gli esiti di questa fase iniziale e delinea il progetto nel suo complesso.

2. MODELLAZIONE DATI E PIPELINE ETL

La trasformazione dell'inventario archivistico in *knowledge graph* ha richiesto l'adozione di una serie di procedure: la scelta e la selezione delle ontologie, la definizione delle strategie di mappatura e l'implementazione del processo di elaborazione.[8] La selezione delle ontologie è stata effettuata sulla base di tre criteri: (1) riconoscimento come standard di dominio da parte delle comunità scientifiche e istituzionali di riferimento, (2) capacità di modellare la complessità strutturale e relazionale dei dati archivistici, (3) garanzia di interoperabilità con altre risorse del web semantico. Su queste basi sono state selezionate:

- RiC-O (Records in Contexts - Ontology),[9] ontologia di riferimento per il dominio archivistico, per la rappresentazione del materiale documentario e dei relativi contesti.
- CIDOC CRM[10] per la modellazione degli oggetti digitali, ossia le digitalizzazioni dei documenti, in relazione con l'inventario.

A partire dai dati di inventario, conservati in formato tabulare, la mappatura verso RiC-O ha seguito tre principi: preservazione semantica delle informazioni originali; esplicitazione delle relazioni implicite nella struttura dei dati; controllo di autorità, allineamento e arricchimento dei dati dove possibile. L'inventario del Fondo Albini è stato dunque mappato secondo classi e proprietà RiC-O; la Tabella 1 illustra e discute la migrazione dei principali campi descrittivi dal formato tabulare al grafo.[11]

[4] https://coco.rahtiapp.fi/
[5] https://bernoulli-euler.dhlab.unibas.ch/
[6] https://bellegreene.itatti.harvard.edu/resource/Start
[7] https://www.epistolariodegasperi.it/#/
[8] Il repository dei materiali è disponibile al link: https://github.com/dharc-org/digital-albini (cons. 25/01/2026).
[9] https://www.ica.org/standards/RiC/RiC-O_1-1.html (cons. 25/01/2026).
[10] https://cidoc-crm.org/ (cons. 25/01/2026).
[11] La mappatura completa e alcuni esempi di modellazione in UML sono disponibili nel repository del progetto: https://github.com/dharc-org/digital-albini/ (cons. 25/01/2026).

Elemento inventario	Classe RiC-O	Commento	N. entità
Serie/ Sottoserie	rico:RecordSet	Primo e secondo livello di aggregazione (logici); le relazioni gerarchiche tra entità rico:RecordSet sono individuate tramite la proprietà rico:isOrWasIncludedIn e la relativa inversa.	4 (s) 7 (ss)
Busta	rico:RecordSet, rico:Instantiation	Terzo livello aggregazione (fisico); le relazioni gerarchiche tra entità rico:RecordSet sono individuate tramite rico:isOrWasIncludedIn e inversa. A livello di contenitori fisici, le relazioni tra rico:Instantiation che rappresentano buste e fascicoli sono individuate tramite rico:hasOrHadPart e la relativa inversa.	36
Fascicolo	rico:RecordSet, rico:Instantiation	Quarto livello aggregazione (logico e fisico); a livello logico, le relazioni gerarchiche tra entità rico:RecordSet sono individuate tramite rico:isOrWasIncludedIn e inversa. A livello fisico, le relazioni tra rico:Instantiation che rappresentano la gerarchia busta-fascicolo e fascicolo-documenti sono individuate tramite rico:hasOrHadPart e la relativa inversa.	581
Documento	rico:Record, rico:Instantiation	Unità documentaria; relazioni gerarchiche individuate tra rico:RecordSet e rico:Record tramite rico:isOrWasIncludedIn e relativa inversa; relazioni gerarchiche individuate fra rico:Instantiation rappresentanti fascicolo e documenti tramite rico:hasOrHadPart e relativa inversa.	1.852
Identificativo univoco	rico:Identifier	Entità tipizzata come "Internal Identifier" tramite la proprietà rico:hasIdentifierType e collegata alla risorsa tramite la proprietà rico:hasOrHadIdentifier.	1.852
Mittente/ Destinatario	rico:Person rico:CorporateBody	Rappresentato come rico:Person o rico:CorporateBody a seconda che sia una persona fisica o un ente. La relazione con il documento è espressa tramite rico:hasSender o rico:hasAddresse allineati a VIAF laddove possibile.	505 (p) 75 (cb)
Data	rico:Date rico:expressedDate rico:normalized DateValue	Gestione di formati eterogenei attraverso parsing automatico basato su espressioni regolari. Le date sono normalizzate in formato ISO 8601 (xsd:date) tramite la data property rico:normalizedDateValue, mentre le date testuali sono preservate tramite rico:expressedDate; range temporali scomposti in rico:hasBeginningDate e rico:hasEndDate.	1601
Luogo	rico:Place	Collegato al documento tramite rico:isAssociatedWithPlace. Ciascun luogo è georeferenziato mediante linking a URI GeoNames.	225

Tabella 1. Mapping tra elementi dell'inventario archivistico e classi RiC-O

La pipeline *Extract, Transform, Load* (ETL) è stata implementata in Python utilizzando pandas[12] per la manipolazione dei dati e rdflib[13] per la generazione RDF. Il sistema ha permesso di trasformare i dati dell'inventario, organizzati per i livelli Serie, Sottoserie, Busta, Fascicolo e Documento, tramite una serie di esplicite istruzioni di mappatura, che definisce le regole di corrispondenza colonna-predicato per ciascun livello gerarchico dei dati dell'inventario.[14] Il processo di elaborazione (Klímek & Škoda, 2017), in particolare, si è articolato nelle seguenti fasi:

[12] https://pandas.pydata.org/ (cons. 25/01/2026).
[13] https://rdflib.readthedocs.io/en/7.1.1/ (cons. 25/01/2026).
[14] La pipeline e il file di mappatura sono disponibili nel repository del progetto: https://github.com/dharc-org/digital-albini/ (cons. 25/01/2026).

1. Lettura regole di mapping: *parsing* delle corrispondenze definite nel file di mappatura.
2. Iterazione per livelli gerarchici: elaborazione sequenziale dei fogli Serie, Sottoserie, Busta, Fascicolo, Documento.
3. Estrazione di entità e creazione nodi: trasformazione di valori testuali in entità semantiche distinte (URI) secondo logiche specifiche.
4. Allineamento e arricchimento tramite API VIAF[15] e GeoNames[16]: persone ed enti, modellati rispettivamente come rico:Person e rico:CorporateBody, sono stati riconciliati a VIAF tramite collegamenti owl:sameAs laddove possibile. Allo stesso modo, le date topiche sono state trasformate in entità rico:Place e arricchite tramite riconciliazione con GeoNames. Per ciascun luogo identificato sono state aggiunte coordinate geografiche e allineamento owl:sameAs all'URI GeoNames.
5. *Post-processing*: propagazione automatica dei mittenti dai fascicoli ai documenti contenuti, garantendo completezza informativa anche quando i dati non sono esplicitati a livello di singolo documento
6. Serializzazione e validazione: generazione del *knowledge graph* (con serializzazione Turtle), verifica della sintassi RDF e interrogazione via SPARQL al fine garantire la correttezza dei dati prodotti e al contempo validare l'espressività del modello e dei dati.

Il completamento della fase ETL ha restituito un primo *knowledge graph* composto da 112894 triple per la descrizione di 1.852 lettere e dei relativi contesti.[17]

3. DAL KNOWLEDGE GRAPH ALLA PIATTAFORMA OPERATIVA

Nonostante il *knowledge graph*, esito della prima fase del progetto, rappresenti semanticamente l'inventario archivistico, il dataset deve essere trasformato in uno strumento analitico operativo. La novità metodologica di questa sperimentazione risiede nell'integrazione dei materiali archivistici e le informazioni biografiche della rete dei corrispondenti, al fine di tracciare l'evoluzione dei ruoli istituzionali di Albini in relazione alle trasformazioni della rete intellettuale. Il *workflow* generale di progetto si basa su quattro ulteriori fasi, che implementeranno progressivamente: (1) funzionalità di consultazione avanzata dei materiali archivistici, (2) arricchimento dei dati attraverso annotazione collaborativa e informazioni biografiche, (3) analisi quantitativa e (4) visualizzazione dei dati. L'obiettivo finale è di trasformare il grafo in uno strumento di ricerca effettivamente operativo per diverse comunità scientifiche.
La fase attualmente in corso (1) è la pubblicazione dei dati, che si basa sull'ambiente ResearchSpace[18]. Questa applicazione richiederà la configurazione di specifici template di visualizzazione, l'implementazione di query SPARQL per il popolamento automatico delle interfacce, la definizione di permessi utente differenziati e l'integrazione del modulo *Semantic Tree Advanced* per la navigazione gerarchica della struttura archivistica. Parallelamente verrà sviluppato un sistema di ricerca articolato su due livelli: una ricerca semplice attraverso *form* strutturati su campi specifici (mittente, destinatario, data, luogo) e una ricerca avanzata che implementerà un *query builder* visuale, permettendo agli utenti di elaborare interrogazioni complesse senza avere competenze tecniche in SPARQL.
La fase successiva (2), programmata per l'estate 2026, prevede la modellazione formale degli oggetti digitali attraverso CIDOC CRM che implementerà l'accesso interattivo alle immagini digitalizzate attraverso *International Image Interoperability Framework* (IIIF) e lo sviluppo di un sistema di annotazione semantica collaborativa (utilizzando modelli come CRMsci[19] e Web Annotation Model[20]). Il sistema supporterà quattro tipologie di annotazioni: testuali (note filologiche, paleografiche, storiche), di entità (identificazione di entità come persone, luoghi, opere, allineate con i controlli di autorità), di citazioni e riferimenti bibliografici. Una sperimentazione pilota testerà il sistema su 50 lettere

[15] https://developer.api.oclc.org/viaf-api (cons. 25/01/2026).
[16] https://www.geonames.org/export/web-services.html (cons. 25/01/2026).
[17] Il file RDF è disponibile nel repository del progetto: https://github.com/dharc-org/digital-albini/ (cons. 25/01/2026).
[18] https://researchspace.org/ (cons. 25/01/2026). La configurazione dell'ambiente, lo sviluppo dei template di visualizzazione e l'implementazione del sistema di ricerca sono attualmente in corso a cura di Erica Andreose e Nazanin Fakharian (Università di Bologna).
[19] https://cidoc-crm.org/crmsci (cons. 25/01/2026).
[20] https://www.w3.org/TR/annotation-model/ (cons. 25/01/2026).

rappresentative coinvolgendo annotatori che lavoreranno per produrre un *corpus* annotato che servirà sia per validazione tecnica sia come modello metodologico. Il completamento di questo primo ciclo sul carteggio fornirà un modello validato che potrà essere replicato, con gli opportuni adattamenti di template, campi di ricerca e tipologie di annotazione, per altri fondi e serie archivistiche.
Successivamente, il grafo sarà arricchito con informazioni sull'evoluzione temporale dei ruoli rivestiti dai personaggi della rete attraverso il pattern ontologico "Role in Time" (Presutti et al., 2016, p. 119). Quest'ultimo arricchimento, insieme al grafo risultante dalle fasi precedenti che include le annotazioni semantiche sulle immagini, consentirà di dare avvio alla terza fase (3), programmata per inizio 2027, che comprende una serie di analisi quantitative e qualitative della rete intellettuale di Albini. Per portare avanti queste analisi verranno implementati strumenti di visualizzazione che includeranno:

- *network analysis* della rete di corrispondenti attraverso grafi interattivi filtrabili temporalmente, permettendo di osservare le trasformazioni nella composizione della rete a seconda dei vari ruoli rivestiti da Albini durante la sua carriera;
- *timeline* sincronizzate visualizzeranno eventi biografici di Albini in parallelo con eventi storici e intensità della corrispondenza;
- mappature geografiche evidenzieranno centri di maggiore densità relazionale;
- dashboard statistiche e moduli di *Semantic Narrative* completeranno la suite di visualizzazioni, consentendo ai curatori di creare percorsi narrativi guidati attraverso il materiale archivistico (Oldman e Tanase, 2018).

La fase conclusiva (4) si concentra sul consolidamento dell'interfaccia utente sulla base dei riscontri raccolti in ambiente di test, sulla migrazione dall'ambiente di sviluppo a quello di produzione e sulla definizione del piano di sostenibilità.
Complessivamente, il progetto produrrà un dataset RDF interoperabile pubblicato con licenza aperta, una piattaforma web di consultazione con funzionalità IIIF e annotazione collaborativa, una suite di visualizzazioni analitiche interattive. L'impatto atteso si articola su più dimensioni: metodologica, fornendo un modello replicabile per archivi umanistici; storiografica, consentendo approfondimento critico delle relazioni tra filologia classica, istituzioni accademiche e regime fascista; tecnologica, contribuendo al dibattito sull'applicazione di ontologie e *knowledge graph* al patrimonio culturale.

4. CONCLUSIONI

La trasformazione dell'inventario del Fondo Giuseppe Albini in *knowledge graph* semantico rappresenta il primo passo verso uno strumento di ricerca che consentirà di affrontare empiricamente le domande poste in apertura: come si sono intrecciate autorità culturale e potere politico nell'Italia del primo Novecento, e quali trasformazioni hanno subito le reti intellettuali europee in quel contesto. La fase ETL completata ha prodotto un dataset di più di undicimila triple per la rappresentazione di 1.852 documenti, strutturato secondo ontologie standard, garantendo preservazione semantica dell'informazione archivistica e interoperabilità con altre risorse del web. Il mapping sistematico ha permesso di esplicitare relazioni implicite nella struttura tabulare dell'inventario originale e ha arricchito il dataset attraverso riconciliazione con authority file esterni (VIAF per persone, GeoNames per luoghi), trasformando nomi isolati in nodi di una rete informativa interrogabile. Le fasi successive porteranno a una piattaforma operativa di ricerca, annotazione collaborativa e analisi quantitativa che permetterà di studiare empiricamente l'evoluzione della rete di Albini, mappando la geografia degli scambi e le sue trasformazioni durante il regime, analizzando quantitativamente il rapporto tra ruoli istituzionali e pratiche relazionali.

BIBLIOGRAFIA

Alassi, S., & Rosenthaler, L. (2024). Semantic precision: Crafting RDF-based digital editions for unveiling the layers of historical correspondence. Digital Scholarship in the Humanities, 39(3), 813–835. https://doi.org/10.1093/llc/fqae027

Baldi, A. (2021). Text analysis ed editoria digitale: proposte di integrazione con valutazioni critiche. A proposito del carteggio tra Ignazio Silone e la Arnoldo Mondadori Editore. Umanistica Digitale, 10. https://doi.org/10.6092/issn.2532-8816/12762

Funaioli, G. (1935). Giuseppe Albini: Orazione. In Annuario della R. Università di Bologna, 1934-35 (pp. 58–109). R. Università di Bologna.

Klímek, J., & Škoda, P. (2017). LinkedPipes ETL in use: practical publication and consumption of linked data. In Proceedings of the 19th International Conference on Information Integration and Web-Based Applications & Services (pp. 441–445). Association for Computing Machinery. https://doi.org/10.1145/3151759.3151809

Nastasi, A. (2019). Le iscrizioni in latino di Roma Capitale (1870-2018). Quasar.

Oldman, D., & Tanase, D. (2018). Reshaping the knowledge graph by connecting researchers, data, and practices in ResearchSpace. In D. Vrandečić et al. (Eds.), The Semantic Web – ISWC 2018 (pp. 325–340). Springer. https://doi.org/10.1007/978-3-030-00668-6_20

Poikkimäki, H., Leskinen, P., & Hyvönen, E. (2025). Exploring cultural heritage knowledge graphs: Case of correspondence networks in Grand Duchy of Finland 1809–1917. Digital Humanities in the Nordic and Baltic Countries Publications, 7(2). https://doi.org/10.5617/dhnbpub.12289

Presutti, V., Lodi, G., Nuzzolese, A., Gangemi, A., Peroni, S., & Asprino, L. (2016). The role of ontology design patterns in linked data projects. In I. Comyn-Wattiau, K. Tanaka, I.-Y. Song, S. Yamamoto, & M. Saeki (Eds.), Conceptual modeling (pp. 113–121). Springer International Publishing. https://doi.org/10.1007/978-3-319-46397-1_9

Tonelli, S., Sprugnoli, R., Moretti, G., Malfatti, S., & Odorizzi, M. (2019). Epistolario De Gasperi: National edition of De Gasperi's letters in digital format. In Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019) (pp. 253–259).

Traina, A. (1991). Giuseppe Albini Latinista. Con una bibliografia di Giuseppe Albini a cura di P. Ferratini. Eikasmos, 2, 321–344.

Tuominen, J., Mäkelä, E., Hyvönen, E., Bosse, A., Lewis, M., & Hotson, H. (2018). Reassembling the Republic of Letters: A linked data approach. In Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference (DHN 2018) (pp. 76–88). CEUR Workshop Proceedings. http://ceur-ws.org/Vol-2084/paper6.pdf

[8/B] Geografie letterarie e multimediali

‘Spazializzare’ la narrativa ecocritica italiana: la piattaforma dell'Atlante digitale della letteratura ecologica italiana

VALENTINA PASQUAL

DIPARTIMENTO DI FILOLOGIA CLASSICA E ITALIANISTICA, UNIVERSITÀ DI BOLOGNA, ITALIA - VALENTINA.PASQUAL2@UNIBO.IT

ABSTRACT (ITALIANO)

L'atlante digitale del progetto PRIN 2022 PNRR "For an Atlas of Italian Ecological Literature: From the Great Acceleration to the Pandemic (LEDA)" è il primo atlante della letteratura ecocritica italiana, frutto della collaborazione tra le Università IULM, Bologna e Siena.
La piattaforma mappa circa 100 opere narrative pubblicate dagli anni Cinquanta (Grande Accelerazione) agli anni 2020 (era pandemica), identificando 250 spazi narrativi e 200 luoghi. L'approccio metodologico adottato integra l'analisi qualitativa degli studi ecocritici con gli strumenti computazionali. Il modello dati realizzato in seno al progetto connette opere e spazi narrati attraverso annotazioni che includono informazioni descrittive delle opere, temi ecologici e parametri di percezione spaziale (tempo e luogo). La definizione di "spazio" include riferimenti non espliciti interpretati dagli studiosi, spazi fittizi e luoghi non georeferenziabili, formalizzando il quadro interpretativo che delimita gli spazi letterari. L'architettura della piattaforma web comprende tre componenti principali: (1) una mappa interattiva con sistema di ricerca a faccette che interroga opere e spazi geolocalizzati mediante filtri controllati, tassonomie gerarchiche e intervalli temporali, (2) indici strutturati che forniscono accesso alternativo ai dati, (3) percorsi critici redatti dagli studiosi che reintroducono la dimensione qualitativa ed ermeneutica nella piattaforma digitale. L'atlante si configura come ambiente integrato per l'analisi quantitativa e qualitativa della letteratura ecologica italiana, valorizzando il contributo critico attraverso la qualificazione dei dati nei percorsi critici.
Parole chiave: digital atlas; literary studies; spatial mapping; ecocritism; contemporary italian literature

ABSTRACT (ENGLISH)

The digital atlas of the PRIN 2022 PNRR project "For an Atlas of Italian Ecological Literature: From the Great Acceleration to the Pandemic (LEDA)" is the first atlas of Italian ecocritical literature, resulting from collaboration between IULM University, University of Bologna, and University of Siena. The platform maps approximately 100 narrative works published from the 1950s (Great Acceleration) to the 2020s (pandemic era), identifying 250 narrative spaces and 200 locations. The methodological approach integrates qualitative analysis from ecocritical studies with computational tools from Digital Humanities. The project’s data model realized connects works and narrated spaces through annotations that include descriptive information about the works, ecological themes, and spatial perception parameters (time and place). The definition of "space" includes non-explicit references interpreted by experts, fictional spaces, and non-georeferenceable places, formalizing the interpretative framework that delimits literary spaces. The web platform architecture comprises three main components: (1) an interactive map with a faceted search system that queries geolocated works and spaces through controlled filters, hierarchical taxonomies, and temporal ranges, (2) structured indexes providing alternative access to the data, (3) critical pathways authored by scholars that reintroduce the qualitative and hermeneutic dimension into the digital platform. The atlas is configured as an integrated environment for quantitative and qualitative analysis of Italian ecological literature, enhancing critical contribution through the qualification of the data in the critical pathways.
Keywords: atlante digitale; studi letterari; mappatura spaziale; ecocritica; letteratura italiana contemporanea

1. INTRODUZIONE

Il progetto PRIN 2022 PNRR "For an Atlas of Italian Ecological Literature: From the Great Acceleration to the Pandemic (LEDA)" ha sviluppato dal 2022 un'indagine qualitativa sul corpus della letteratura italiana

che tematizza il rapporto uomo-ambiente. Attraverso la collaborazione delle unità di ricerca dell'Università IULM, dell'Università di Bologna e dell'Università di Siena, la ricerca ha analizzato circa 100 testi narrativi dal periodo della Grande Accelerazione (anni Cinquanta) all'era pandemica (anni 2020), approfondendo le modalità di rappresentazione letteraria dello spazio in relazione alla critica ecologica. Il progetto si inserisce nel quadro teorico degli studi ecocritici, approccio critico agli studi letterari che indaga le relazioni tra letteratura, spazio e questioni ecologiche (Scaffai, 2017; Iovino, 2016).
Secondo la teoria letteraria, un mondo narrativo è costituito da tre componenti: personaggi, trama e spazio. Mentre personaggi e trama sono stati oggetto di innumerevoli studi filologici, gli aspetti territoriali e topografici della letteratura sono stati a lungo trascurati. In questo contesto, lo *spatial turn* nelle discipline umanistiche ha rinnovato l'attenzione verso la dimensione spaziale dei testi (Piatti et al., 2009). Dal punto di vista delle possibili interazioni tra la critica letteraria tradizionale, intesa come *close reading*, e gli strumenti computazionali, *il distant reading (Moretti, 2013)* ha promosso l'adozione di pratiche quali la visualizzazione geografica come metodo per rappresentare gli aspetti spaziali dei testi narrativi, rendendo la dimensione spaziale e spazio-temporale una delle caratteristiche centrali nell'analisi dei contenuti letterari (Jänicke et al., 2015). In risposta al dibattito metodologico sul *distant reading,* il concetto di *scalable reading* (Mueller, 2012) supera la dicotomia tra lettura *close* e *distant*, proponendo una sintesi metodologica che non mira a sostituire la lettura tradizionale, ma a creare nuove forme di mediazione tra testo e contesto attraverso strumenti digitali (Krautter, 2024).
In questo quadro metodologico, il progetto ha realizzato il primo atlante digitale della letteratura ecocritica italiana. La piattaforma raccoglie, sistematizza e dissemina gli spazi descritti dai testi selezionati e la loro classificazione in chiave ecocritica. L'atlante è uno strumento interattivo che consente di visualizzare, interagire e ragionare sugli spazi. Il presente contributo presenta l'architettura tecnologica della piattaforma, le funzionalità di visualizzazione e interrogazione dei dati, e il suo posizionamento negli studi letterari computazionali.

2. STATO DELL'ARTE

Nell'ambito della cartografia letteraria digitale, diversi progetti privilegiano un approccio di tipo monografico, finalizzato alla rappresentazione cartografica esaustiva dell'opera di singoli autori. Il progetto *GaddaMap*[1] dell'Università di Bologna mappa le menzioni geografiche di Carlo Emilio Gadda. La piattaforma utilizza OpenStreetMap[2] come base cartografica e integra un *layer* storico della Milano contemporanea all'autore. L'interfaccia implementa filtri per la selezione delle fonti (opere e lettere) e fornisce accesso ai dati attraverso due indici: stradario e luoghi menzionati. Il corpus comprende due opere, trentacinque lettere e più di settecento menzioni geografiche milanesi. Il progetto *Placing Literature*[3] presenta due mappe letterarie della letteratura vittoriana: una dedicata all'opera di Charles Dickens (curata dalla *Dickens Society*) e una focalizzata sul canone di Sherlock Holmes (curata da T. B. Wheeler)[4]. Alternativamente alle soluzioni cartografiche, Atlante Calvino[5] (Ghidinelli, 2019) affronta la narrazione della spazialità nella produzione calviniana attraverso tre percorsi di *data-storytelling* sulla spazialità: i luoghi, il concetto di trasformazione dello spazio e la rappresentazione della realtà. Ciascuna tappa è corredata da un saggi critico che contestualizza ed approfondisce la visualizzazione dei dati[6].
Alcuni atlanti seguono un approccio tematico che raccoglie diversi autori per identificare tematiche ricorrenti. E' il caso del recente progetto *Literary Atlas of Volcanic and Seismic Activities* (LAVA), sviluppato dall'Univeristà di Pisa, che ha l'obiettivo di mappare le narrazioni letterarie di terremoti ed eruzioni vulcaniche attraverso diversi secoli ed autori. Il progetto adotta un approccio interdisciplinare che integra dati geologici con fonti letterarie attraverso percorsi tematici e mappe interattive[7]. Il progetto MeMo (Memoria e Mezzogiorno)[8] (Giurdanella et al., 2025) si configura come una piattaforma di mappatura letteraria digitale finalizzata alla valorizzazione della memoria culturale territoriale. La mappa geolocalizza le modalità attraverso cui le opere letterarie — dal 1945 a oggi — rappresentano tre

[1] https://lab.dharc.unibo.it/gaddamap/
[2] https://www.openstreetmap.org/#map=18/42.454197/13.467291&layers=P
[3] https://placingliterature.wordpress.com/
[4] Il progetto non risulta accessibile al momento: http://www.placingliterature.com/
[5] https://atlantecalvino.unige.ch/navigation?lang=it
[6] Ad esempio l'approfondimento sulla tappa "Realismo": https://atlantecalvino.unige.ch/space/phase3/focus?lang=it
[7] https://www.unipi.it/en/news/introducing-lava-the-first-digital-literary-atlas-mapping-earthquakes-and-eruptions-through-literature/
[8] https://heurist.huma-num.fr/f_aut_memo/web/137/131

macrotemi, guerra, industria e spopolamento, intesi come spazi simbolici di una memoria in continua trasformazione. In questa prospettiva, i testi analizzati sono interpretati alla luce del loro rapporto con il territorio, in quanto custodi di un valore memoriale e dell'interazione tra comunità antropiche e contesto spaziale. *The Emotions of London*[9], sviluppato dal *Literary Lab di Stanford*, sistematizza le emozioni esperite dai personaggi letterari in opere ambientate a Londra nei secoli XVIII e XIX. Il progetto applica tecniche di *sentiment analysis* e *Named Entity Recognition* per l'estrazione e geolocalizzazione automatica di riferimenti emozionali da un corpus generato attraverso *crowdsourcing* (Heuser et al., 2016).
Un ulteriore approccio alla cartografia letteraria è rappresentato da piattaforme che aggregano luoghi narrativi senza vincoli autoriali o tematici predefiniti, con finalità prevalentemente divulgative e orientate all'accessibilità pubblica piuttosto che all'analisi sistematica. *Mappa Letteraria*[10] visualizza più di cinquemila luoghi letterari italiani (città, quartieri, strade, edifici) narrati in romanzi e opere di *fiction*. La piattaforma collega testi, autori e ambientazioni su una mappa interattiva, che è possibile filtrare per luogo, autore, genere, anno, valutazione e percorso di lettura. Il progetto privilegia l'accessibilità a un pubblico non specialistico ed è possibile contribuire all'inserimento delle opere tramite un sistema di log-in.
Nonostante l'efficacia delle soluzioni tecnologiche adottate, la dimensione qualitativa della ricerca risulta poco integrata nelle piattaforme digitali (ad eccezione degli approfondimenti in Atlante Calvino). Inoltre, il quadro interpretativo che sottende la definizione e la delimitazione degli spazi letterari mappati non viene formalizzato né esplicitato, riducendo la componente ermeneutica dell'analisi spaziale a una pratica implicita, non formalizzata, né spesso documentata.

3. APPROCCIO ALLA RICERCA E PRODUZIONE DEL DATASET

La ricerca condotta nel progetto LEDA adotta un approccio metodologico fondato sull'analisi qualitativa e interpretativa dei testi letterari, in cui la dimensione spaziale costituisce un costrutto critico, non un mero elemento descrittivo. Tale studio ha portato le tre unità di ricerca a redigere un indice di circa un centinaio di opere della letteratura italiana pubblicate dagli anni Cinquanta — soglia cronologica motivata dal primo emergere di opere dedicate a questioni ecologiche — a oggi, selezionate per la loro rilevanza rispetto alla spazialità letteraria e ai temi ecocritici.
L'atlante digitale LEDA si è quindi sviluppato a partire da questo indice già criticamente definito. La fase di modellazione dei dati si è sviluppata attraverso un processo collaborativo tra le tre unità, finalizzato a sistematizzare gli studi condotti, ed ad arricchire le informazioni ecocritiche e spaziali rilevanti. Il modello dati dell'atlante si fonda su una struttura relazionale che connette due entità principali—opere e spazi narrati—a cui sono associate le seguenti annotazioni:

- **Opere**: titolo (es. *Le otto montagne*), autore (es. Cognetti, Paolo), anno di prima pubblicazione (es. 2016), editore (es. Einaudi), modo narrativo (es. Realistico), temi ecologici (e.s., Non umano > Piante, Non umano > Animali, Isolamento > Isolamento volontario e Alterazione antropica del territorio > Urbanizzazione, Rovine o resti), breve sinossi dell'opera che ne sottolinea i temi ecocritici.
- **Spazi narrati**: luogo (es. Milano), tipologia nel contesto dell'opera (e.s., città), stato di realtà (es. spazio reale), temporalità (es. presente), breve descrizione testuale dello spazio, ed infine, il tipo di identificazione dello spazio da parte dello studioso (e.g. Menzione esplicita, dove Milano viene esplicitamente citato nell'opera di riferimento).

La definizione del concetto di spazio si è rivelata fondamentale per includere riferimenti non totalmente espliciti interpretati dagli studiosi. Ad esempio, in *La vita in tempo di pace* di Francesco Pecoraro (2013), "la città di Dio" viene ricondotta alla città di Roma grazie a indicatori urbanistico-architettonici e sociali descritti nel testo. Il dataset dell'atlante registra, oltre al lugo menzionato (luogo: Roma) ed alla sua menzione specifica (note), sia lo stato di realtà dello spazio (reale con nome fittizio) sia il tipo di identificazione (Tipo di identificazione: deduzione dalla descrizione). Il modello accoglie anche spazi fittizi non referenziabili a luoghi reali: in *Il pianeta irritabile* di Paolo Volponi (1978) non viene specificato un luogo preciso (annotato come "Luogo non specificato sul pianeta Terra"), ma si deduce dalla descrizione

[9] https://litlab.stanford.edu/projects/emotions-london/
[10] https://www.mappaletteraria.it/

che l'ambientazione è una città (tipologia: città) nel futuro (temporalità: futuro). In termini metodologici, il lavoro di annotazione è stato svolto prevalentemente in forma manuale, a partire da un indice di opere selezionate dagli studiosi ed ampliato in base alle scelte di modellazione sopra descritte.
I luoghi sono stati geolocalizzati semiautomaticamente attraverso il servizio API Nominatim[11], estraendo latitudine, longitudine, OSM_id e poligono; i risultati sono stati verificati e corretti manualmente. I poligoni sono utilizzati per rappresentare aree geografiche estese che non possono essere identificate da un singolo punto (e.g., l'Himalaya ne *Le otto montagne*, dove non viene specificato un luogo preciso). Ai luoghi non georeferenziabili sono state assegnate coordinate simboliche che rimandano a "luogo indefinito"[12].
In seguito, si è proceduto alla popolazione del dataset da parte delle tre unità e, parallelamente, all'implementazione dell'applicazione web *client-side* ed *open-source* dell'atlante[13]. La distribuzione cronologica delle opere nel dataset è asimmetrica e presenta due picchi di concentrazione negli anni Ottanta (15 opere) e nell'ultimo ventennio (56 opere). Coerentemente con la vocazione cartografica di LEDA, tale distribuzione riflette la dinamica effettiva di propagazione del campo, che conosce una fase di consolidamento a partire dagli anni Dieci del Duemila. Questo consolidamento si inscrive in un più ampio incremento di attenzione per le questioni ambientali, di natura non esclusivamente letteraria, e si colloca in parallelo al radicarsi dell'ecocritica come campo di studi in Italia (Iovino, 2016; Scaffai, 2017).

4. L'ATLANTE DIGITALE

Attualmente, l'atlante ospita attualmente circa 250 spazi (menzioni) che descrivono 200 luoghi tratti da 99 fonti (riferimenti letterari). Oltre alle sezioni di homepage e le descrizione del progetto, la piattaforma offre tre principali punti di accesso ai dati: mappa, indici e percorsi critici.
La mappa interattiva (Figura 1), fulcro del progetto, integra opere (fonti), spazi narrativi (menzioni) ed i rispettivi luoghi georeferenziati (luoghi). L'interfaccia si articola in tre componenti principali: il pannello sinistro ospita il sistema di ricerca a faccette, l'area centrale visualizza la mappa con i marcatori geografici, il pannello destro presenta il catalogo delle opere (espandibile per visualizzare le la scheda informativa che include le informazioni associate a ciascuna fonte e ai relativi spazi narrativi). Il sistema di ricerca a faccette comprende una selezione di campi descrittivi definiti in fase di modellazione, relativi sia alle opere che agli spazi rappresentati. L'applicazione dei filtri agisce simultaneamente su entrambe le componenti dell'interfaccia: i marcatori visualizzati sulla mappa e le opere elencate nel catalogo vengono filtrati in modo coordinato secondo i criteri selezionati. Il sistema prevede tre tipologie di filtri: *checkbox* per liste di termini controllati (es. tipologie di spazi come città, paese, fiume); tassonomie gerarchiche (es. categorie ecocritiche come *Non umano > Piante, Non umano > Animali; Isolamento > Isolamento volontario; Alterazione antropica del territorio > Urbanizzazione, Rovine o resti*); e filtri temporali a intervallo (es. date di pubblicazione come 1950-2000).

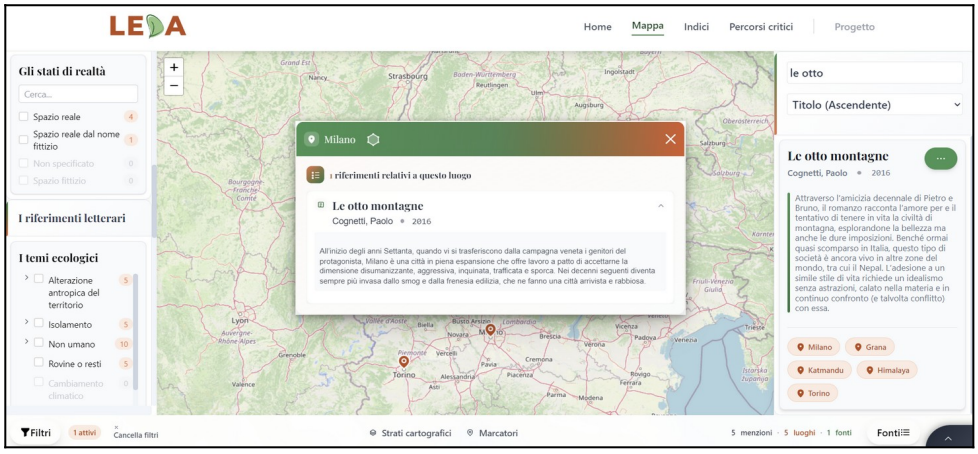


Figura 1. Sezione mappa dell'atlante LEDA, selezione dell'opera "Le otto montagne" di Paolo Cognetti (2016)

[11] https://nominatim.openstreetmap.org/ui/search.html
[12] lat:38.7200, long:-24.2200
[13] https://github.com/progettoleda2022prin/atlante/tree/main/public/data

La sezione Indici fornisce un ulteriore strumento di accesso ai dati. Ciascun indice è strutturato sui medesimi campi del sistema di filtri della mappa interattiva (Figura 1). Analogamente ai filtri, gli indici si articolano in tre tipologie: indice semplice, indice gerarchico e lista degli intervalli temporali, ciascuna con una visualizzazione personalizzata per accomodare queste diverse tipologie di informazione. A titolo esemplificativo, come illustrato in Figura 2, l'indice dei titoli costituisce un indice semplice che presenta: l'elenco delle opere, il numero di spazi descritti per ciascuna opera, il rimando alla relativa scheda descrittiva e un collegamento diretto alla mappa (icona mappa posizionata a sinistra). Una barra laterale consente di interagire con i contenuti attraverso funzionalità di ricerca, filtro e ordinamento.

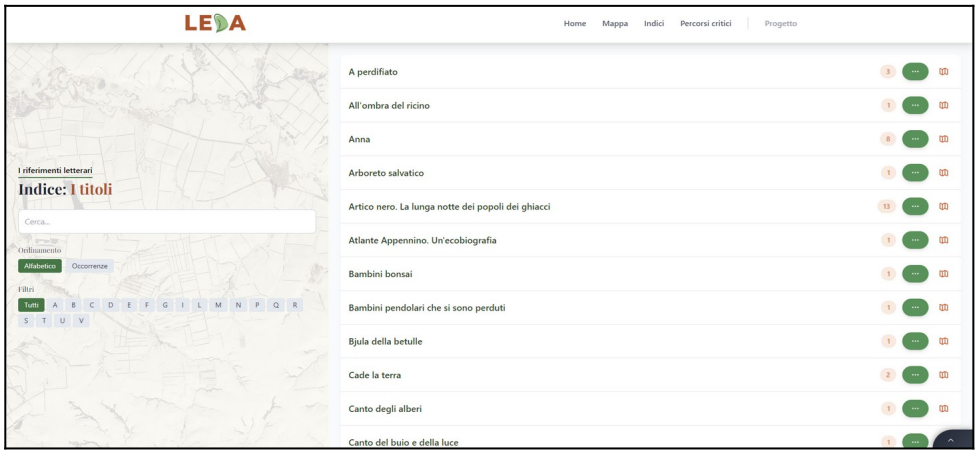


Figura 2. Screenshot dell'indice dei titoli delle opere dell'atlante LEDA

I percorsi critici costituiscono l'ultimo punto di accesso agli spazi e alle opere. Ciascun percorso è redatto da un membro del progetto ed è corredato da titolo, autore, parole chiave e abstract. Le entità menzionate all'interno del percorso sono semanticamente marcate e collegate alle rispettive istanze sulla mappa. Inoltre, per ciascuna opera è possibile visualizzare la corrispondente scheda descrittiva. Figura 3 illustra il percorso critico scritto da Lavinia Torti dal titolo "Spazi reali, spazi fittizi, spazi generici o specifici: una tassonomia", pensato per approfondire e contestualizzare la natura delle scelte di modellazione del modello dati. La componente fondamentale dei percorsi critici risiede nella reintroduzione della narrativa qualitativa elaborata dagli studiosi, che ricontestualizza i dati, esplicita le connessioni latenti, le scelte di modellazione e i criteri di inclusione delle opere, nonché il percorso critico affrontato dai ricercatori nel corso del progetto. In questa prospettiva, i dati dell'atlante forniscono uno strumento di supporto e raccordo per i contributi testuali degli studiosi in un ambiente integrato e interattivo, che consente l'analisi sia qualitativa che quantitativa dei materiali dell'atlante, valorizzando al contempo il contributo fortemente ermeneutico e critico degli studiosi.

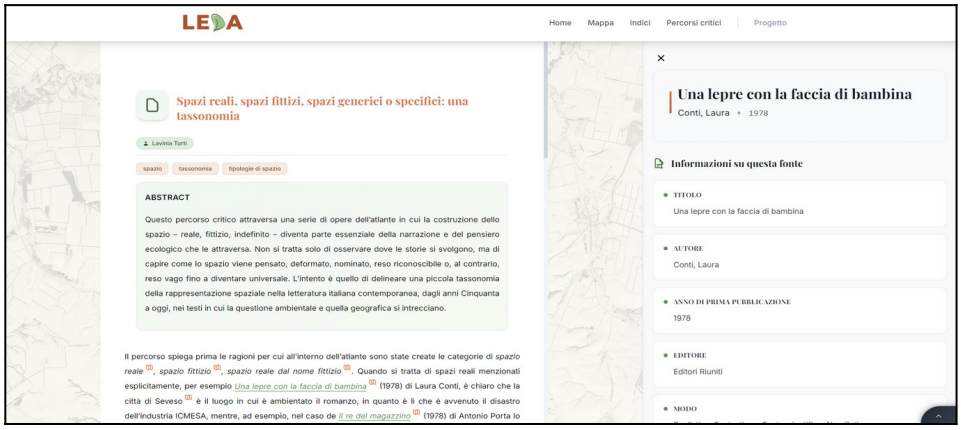


Figura 3. Screenshot del percorso critico "Spazi reali, spazi fittizi, spazi generici o specifici: una tassonomia" ospitato nell'atlante LEDA

5. CONCLUSIONI

Il progetto PRIN 2022 PNRR *For an Atlas of Italian Ecological Literature* ha realizzato il primo atlante digitale della letteratura ecocritica italiana, proponendo un approccio metodologico che integra analisi qualitativa, strumenti computazionali e pratiche interpretative per lo studio della spazialità letteraria in chiave ecologica. L'architettura tecnologica implementa un doppio obiettivo: da un lato, supportare l'analisi quantitativa attraverso l'interrogazione strutturata dei dati spaziali ed ecocritici; dall'altro, valorizzare e disseminare i risultati degli studi qualitativi condotti dalle tre unità di ricerca del progetto. In questo senso, l'atlante fornisce una soluzione al problema dell'accesso alla conoscenza sulle opere in contesti vincolati da restrizioni di copyright. Pur non consentendo la lettura integrale delle opere, la piattaforma documenta e rende interrogabile il punto di vista critico degli studiosi, trasformando l'interpretazione critica in un dato strutturato e visualizzabile. L'accesso visuale a tale conoscenza si configura così come forma di mediazione tra testo e contesto che, nella prospettiva dello *scalable reading*, non sostituisce ma integra la lettura tradizionale, creando nuove possibilità di esplorazione e ragionamento critico sugli spazi letterari dell'ecocritica italiana.
Infine, dal punto di vista della sostenibilità a lungo termine, l'atlante adotta un'architettura *client-side,* ospitata su GitHub Pages, che azzera i costi di manutenzione e *hosting* al fine di garantine l'accessibilità immediata e la persistenza nel tempo dei risultati della ricerca. Il codice sorgente è rilasciato con licenza open source, a garanzia della riproducibilità del progetto.
Possibili sviluppi futuri includono: l'affinamento delle annotazioni spaziali mediante l'adozione di categorie narratologiche derivate dalla teoria di Franco Moretti (Moretti, 1999; Reuschel, 2009), con particolare attenzione al ruolo dello spazio nel dispositivo narrativo; l'implementazione di strumenti di visualizzazione dedicati a sottoinsiemi specifici dei dati, quali gli spazi non geolocalizzabili e le configurazioni narrative, ad esempio, legate ai concetti di movimento e viaggio.

RINGRAZIAMENTI

PRIN 2022 PNRR dal titolo “For an Atlas of Italian Ecological Literature: From the Great Acceleration to the Pandemic (LEDA)” Codice Progetto P2022XX4H5 CUP J53D23016610001”, finanziato nell’ambito del Piano Nazionale di Ripresa e Resilienza PNRR - Missione 4 – Componente 2 – Investimento 1.1 “Fondo per il Programma Nazionale di Ricerca e Progetti di Rilevante Interesse Nazionale (PRIN)” (Bando indetto con D.D. del MUR n. 1409 del 14/09/2022)

BIBLIOGRAFIA

Ghidinelli, S. (2019). Vedere i testi, leggere le immagini: le visualizzazioni per “la Lettura”. In Tuttestorie di donne (pp. 197-201). Fondazione Arnoldo e Alberto Mondadori-Il Saggiatore.

Giurdanella, L., Palazzolo, G., De Luca, B., Autiero, F., Gatto, M., Peluso, S., et al. (2025). MeMo: Una mappa letteraria digitale per la memoria del Mezzogiorno. Proceedings of AIUCD 2025, 627.

Heuser, R., Algee-Hewitt, M., Lockhart, A., Steiner, E., & Tran, V. (2016). Mapping the emotions of London in fiction, 1700–1900: A crowdsourcing experiment. In Literary mapping in the digital age (pp. 25-46). Routledge.

Iovino, S. (2016). Ecocriticism and Italy: Ecology, resistance, and liberation. Bloomsbury Academic, 208.

Jänicke, S., Franzini, G., Cheema, M. F., & Scheuermann, G. (2015). On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges. EuroVis (STARs), 83-103.

Krautter, B. (2024). The Scales of (Computational) Literary Studies: Martin Mueller’s Concept of Scalable Reading in Theory and Practice. Zoomland: Exploring scale in digital history and humanities, 261-286.

Moretti, F. (1999). Atlas of the European novel: 1800-1900. verso.

Moretti, F. (2013). Distant reading. Verso Books.

Mueller, M. (2012, May 29). Scalable Reading. Accessed January 31, 2026. https://sites.northwestern.edu/scalablereading/2020/04/26/scalable-reading

Piatti, B., Bär, H. R., Reuschel, A. K., Hurni, L., & Cartwright, W. (2009). Mapping literature: Towards a geography of fiction. In Cartography and art (pp. 1-16). Berlin, Heidelberg: Springer.

Reuschel, A. K., Piatti, B., & Hurni, L. (2009). Mapping literature. the prototype of “A Literary Atlas of Europe”. In Proceedings of the 24 th International Cartographic Conference.

Scaffai, N. (2017). Letteratura e ecologia: forme e temi di una relazione narrativa (Vol. 245). Carocci.

MeMo: mappe digitali, memoria letteraria e public engagement nel Mezzogiorno

LAURA GIURDANELLA[1]; GIUSEPPE PALAZZOLO[1]
[1]UNIVERSITÀ DI CATANIA

ABSTRACT (ITALIANO)

Il progetto *MeMo – Memory and Mezzogiorno. A literary map from 1945 to the present* si propone di ripensare lo studio e la valorizzazione della memoria letteraria del Sud Italia attraverso le Digital Humanities, con un approccio geocentrico ed ecocritico. Il Mezzogiorno emerge come spazio fondativo e al contempo marginale nella costruzione dell'identità italiana, le cui trasformazioni sociali, ambientali e culturali sono raccontate dalla letteratura postbellica. MeMo integra analisi testuali e dati spaziali, costruendo una mappa letteraria interattiva che connette luoghi reali, simbolici e immaginati con testi in prosa, poesia e scritture non finzionali, in lingua italiana e dialettale. L'infrastruttura digitale sviluppata include strumenti di visualizzazione geospaziale, filtri avanzati, corpus interrogabili e interfacce accessibili, in linea con i principi FAIR di scienza aperta e riusabilità dei dati. La metodologia del progetto integra la geostoria, la letteratura e le Digital Humanities, considerando spazio, luogo, ambiente, territorio e paesaggio come strumenti interpretativi per leggere le interazioni tra eventi storici, traumi, migrazione, industrializzazione e trasformazioni culturali. MeMo promuove la didattica attiva e laboratoriale, mediante unità modulari di apprendimento che connettono presente e passato e favoriscono competenze critiche, digitali e di cittadinanza attiva. La fase avanzata del progetto si concentra sulla riutilizzazione dei dati per finalità ermeneutiche, educative e partecipative, valorizzando voci minoritarie e testi marginali spesso esclusi dai canoni tradizionali. L'infrastruttura digitale diventa così uno spazio di public engagement, dove ricerca, formazione e pubblico dialogano in maniera interattiva, rafforzando la relazione tra territorio, memoria culturale e patrimonio letterario.
Parole chiave: Mezzogiorno; Patrimonio letterario digitale; Geostoria; Didattica digitale; Public Engagement

ABSTRACT (ENGLISH)

MeMo: Digital Maps, Literary Memory, and Public Engagement in Southern Italy
The *MeMo project – Memory and Mezzogiorno. A Literary Map from 1945 to the Present* – aims to rethink the study and valorization of literary memory in Southern Italy through the lens of Digital Humanities, adopting a geocentric and ecocritical approach. The Mezzogiorno emerges as both a foundational and marginal space in the construction of Italian identity, whose social, environmental, and cultural transformations are narrated through postwar literature. MeMo integrates textual analysis with spatial data, creating an interactive literary map that connects real, symbolic, and imagined places with prose, poetry, and non-fiction texts in Italian and regional dialects. The developed digital infrastructure includes geospatial visualization tools, advanced filters, queryable corpora, and accessible interfaces, in line with the FAIR principles of open science and data reusability.
The project's methodology combines geohistory, literature, and Digital Humanities, considering space, place, environment, territory, and landscape as interpretive tools for reading the interactions between historical events, traumas, migration, industrialization, and cultural transformations. MeMo promotes active, hands-on learning through modular teaching units that link past and present while fostering critical, digital, and civic competencies. The advanced phase of the project focuses on the reuse of data for interpretive, educational, and participatory purposes, giving visibility to minority voices and marginal texts often excluded from traditional canons. The digital infrastructure thus becomes a space for public engagement, where research, education, and audiences interact dynamically, strengthening the relationship between territory, cultural memory, and literary heritage.
Keywords: Southern Italy; Digital Literary Heritage; Geohistory; Digital Education; Public Engagement

1. INTRODUZIONE

Il Mezzogiorno d'Italia occupa una posizione profondamente ambivalente all'interno del discorso culturale e letterario nazionale (Abbate, 1968). Da un lato, esso costituisce uno spazio fondativo nella costruzione dell'identità italiana, radicato in immaginari collettivi, narrazioni storiche e geografie simboliche di lunga durata; dall'altro, è stato frequentemente rappresentato come area marginale, periferica o arretrata rispetto ai processi di modernizzazione economica e culturale. Questa tensione attraversa in modo particolarmente evidente la produzione letteraria del secondo dopoguerra, in cui i paesaggi meridionali emergono come luoghi segnati dalla guerra, dallo sfruttamento ambientale, dall'industrializzazione diseguale e dalle molteplici forme di mobilità forzata o volontaria.
La letteratura prodotta nel e sul Mezzogiorno si configura così come un archivio privilegiato delle trasformazioni sociali, ambientali e culturali che hanno contribuito a modellare la modernità italiana in forme profondamente asimmetriche (Moliterni, 2024). Pur in presenza di un'ampia tradizione critica dedicata ad autori, opere e correnti specifiche, risultano ancora limitati gli approcci sistematici capaci di indagare la memoria letteraria meridionale in una prospettiva geocentrica e digitalmente mediata (Assmann, 2004; Marocco, 2014). Negli ultimi anni, gli studi sul neorealismo, sulla scrittura regionale e sulla narrativa postbellica hanno messo in luce il ruolo della letteratura come strumento di testimonianza del trauma, del cambiamento sociale e della frattura storica, ma raramente tali riflessioni sono state integrate in infrastrutture digitali orientate alla rappresentazione spaziale e alla riusabilità dei dati (Alfano, 2010; Pedullà, 2016).

2. DALL'IMPIANTO TEORICO ALL'INFRASTRUTTURA DIGITALE E CARTOGRAFICA

È in questo contesto che si colloca *MeMo – Memory and Mezzogiorno. A literary map from 1945 to the present*, progetto finanziato dal PRIN 2022 PNRR, che propone un ripensamento delle modalità di studio, conservazione e attivazione della memoria letteraria attraverso le Digital Humanities (Giurdanella et al., 2025). Il progetto adotta una prospettiva ecocritica e geo-letteraria, assumendo i luoghi – reali, simbolici e immaginati – come unità primarie di analisi (*in chiave geocentrica*), piuttosto che i singoli autori o le opere canoniche (*studi egoriferiti*) (Westfal, 2007; Scaffai, 2017; Iovino, 2022). Città, borghi, siti industriali, rovine, aree costiere e spazi narrativi diventano così nodi in cui si intrecciano memoria storica, trasformazioni ambientali ed esperienze collettive.
Dal punto di vista operativo (vd. Fig. 1), tale impostazione teorica si traduce nella costruzione di un'infrastruttura digitale basata sulla piattaforma Heurist, uno strumento open source progettato per la gestione e la pubblicazione di basi di dati nelle scienze umane e sociali. La scelta di questa piattaforma è stata determinata dalla sua elevata flessibilità nella modellazione dei dati e dalla possibilità di definire relazioni complesse tra record, adattando la struttura informativa alle specificità del dominio umanistico. Nel caso di MeMo, è stata progettata una struttura modulare che integra informazioni su autori, opere, luoghi e categorie tematiche – guerra, industrializzazione/deindustrializzazione e migrazione – creando relazioni gerarchiche e trasversali tra tali entità e rappresentando la complessità delle interazioni tra letteratura e spazio geografico.
Il database è stato sviluppato secondo uno schema che rispecchia le esigenze analitiche del progetto. I luoghi costituiscono l'unità centrale del sistema e sono registrati mediante coordinate geografiche precise e classificati secondo tipologie specifiche (ad esempio città, siti industriali o luoghi immaginari). Per ciascun luogo sono archiviate descrizioni, riferimenti testuali e materiali multimediali associati. Le opere includono informazioni bibliografiche dettagliate, tra cui autore, titolo, lingua e genere, nonché collegamenti ai luoghi e alle categorie tematiche; è stata inoltre introdotta una doppia contestualizzazione temporale che distingue tra tempo reale (anno di pubblicazione) e tempo finzionale (periodo narrato nel testo). Le schede degli autori integrano dati biografici e rimandi a authority file nazionali e internazionali (quali ISNI e VIAF), mentre le categorie tematiche consentono di connettere fenomeni storici e rappresentazioni letterarie in modo sistematico. L'adozione della piattaforma Heurist garantisce inoltre la piena aderenza ai principi *FAIR (Findable, Accessible, Interoperable, Reusable)* (Wilkinson et al., 2016) e consente l'esportazione dei dati in formati interoperabili, tra cui XML e SQL, assicurando la sostenibilità del progetto nel lungo periodo.
Parallelamente alla costruzione del database, è stato sviluppato un portale digitale interattivo che consente la visualizzazione e l'esplorazione dei dati attraverso una mappa letteraria basata sul framework open source Leaflet, integrato con OpenStreetMap. Questo strumento consente la visualizzazione

dinamica dei dati geospaziali e offre un'interfaccia intuitiva che permette agli utenti di navigare tra luoghi, opere e autori, accedendo a schede descrittive corredate da estratti testuali e materiali iconografici. La mappa include inoltre layer tematici dedicati ai principali ambiti di analisi – guerra, industrializzazione e migrazione – e un sistema di filtri avanzati che permette di selezionare i dati in base a parametri specifici, quali periodo storico, genere letterario o fenomeno socio-economico.
Questo sistema di interrogazione rappresenta uno degli elementi più innovativi del progetto, poiché consente di individuare pattern ricorrenti e di analizzare correlazioni tra luoghi, testi e fenomeni storici, favorendo l'emersione di nuove prospettive interpretative. La mappa non si configura quindi soltanto come uno strumento di consultazione, ma come un ambiente dinamico di ricerca e scoperta, capace di supportare analisi critiche e di rendere accessibile a un pubblico più ampio il patrimonio letterario del Mezzogiorno.
Le unità di ricerca coinvolte, in linea con gli obiettivi progettuali, hanno inoltre realizzato un convegno conclusivo concepito come spazio di restituzione e confronto metodologico, nel quale sono state approfondite le modalità attraverso cui luoghi e comunità del Mezzogiorno sono stati storicamente trasformati e simbolicamente rielaborati in ambito letterario in relazione ai tre fenomeni chiave individuati: guerra, deindustrializzazione e migrazione.
In questa cornice teorica e metodologica, MeMo si inserisce nel più ampio ripensamento delle modalità di produzione e trasmissione del sapere storico, orientate a pratiche di apertura, condivisione e coinvolgimento del pubblico. Fondato sui principi della public digital history (Dell'Oglio & Triolo, 2018), il progetto concepisce il digitale non solo come strumento di disseminazione, ma come ambiente di lavoro e di interazione, capace di rafforzare il legame tra ricerca accademica, territorio e memoria culturale del Mezzogiorno.



Figura 1. Ecosistema digitale MeMo (Immagine creata con NotebookLM)

3. OLTRE IL DATABASE: UNA METODOLOGIA INTEGRATA PER LA VALORIZZAZIONE DIGITALE DELLA MEMORIA LETTERARIA

Il progetto MeMo è entrato in una fase avanzata, focalizzata sulla riutilizzazione e valorizzazione dei materiali digitali secondo prospettive critico-ermeneutiche, didattiche e partecipative, in linea con le *Public Humanities* e la *Open Science*. Partendo da un corpus strutturato di letteratura del dopoguerra del Sud Italia, codificato in XML-TEI e arricchito da metadati multidimensionali (autori, generi, temi, riferimenti spaziali), il progetto sperimenta pratiche di engagement che combinano lessicografia, ermeneutica e didattica della letteratura. Corpus e mappa digitale diventano ambienti dinamici di apprendimento e interpretazione, favorendo la co-costruzione della conoscenza tra ricercatori, studenti e pubblico non specialistico.

Gli strumenti digitali consentono di esplorare testi e luoghi tramite visualizzazioni tematiche e spaziali, filtri avanzati e layer interattivi, rendendo visibili pattern di memoria e narrazione altrimenti difficili da cogliere. L'infrastruttura digitale supporta la valorizzazione di voci minoritarie e marginali, mettendo in luce le interconnessioni tra letteratura, trasformazioni ambientali, migrazione e industrializzazione, con particolare attenzione alla Sicilia contemporanea.
L'approccio metodologico privilegia sostenibilità, accessibilità e riusabilità dei dati secondo i principi FAIR, combinando conservazione archivistica con interazione educativa e partecipativa. In questo modo, MeMo non solo documenta il patrimonio letterario e culturale del Mezzogiorno, ma sperimenta modelli di generazione della memoria digitale, rafforzando il dialogo tra ricerca scientifica, formazione e pubblico e promuovendo una memoria letteraria viva, aperta e condivisa.

4. INTERDISCIPLINARIETA' TRA GEOSTORIA, LETTERATURA E DH

Nel progetto MeMo, l'analisi della letteratura post-bellica del Sud Italia segue una prospettiva geostorica e interdisciplinare, in cui luoghi, città, villaggi e paesaggi – reali o simbolici – connettono eventi storici, trasformazioni sociali e ambientali con le narrazioni letterarie (De Lucia, Gallo, Marino, 2014; Peterle, 2016). I concetti fondamentali della geografia diventano strumenti interpretativi capaci di generare percorsi di indagine interdisciplinari secondo la metafora delle "cinque porte" (Westfal, 2007; Rocca, 2012):

a) *Spazio*: dato naturale e fisico, antecedente all'intervento umano, fornisce il contesto e gli strumenti di orientamento per comprendere la configurazione territoriale.
b) *Luogo*: spazio vissuto soggettivamente, dove emozioni, memoria, identità e percezioni sensoriali conferiscono significato personale e collettivo alle ambientazioni narrative.
c) *Ambiente*: descrive le relazioni tra sistemi naturali e attività umane, includendo ecosistemi, risorse, cambiamento climatico e sostenibilità, permettendo di leggere la letteratura in chiave uomo-natura.
d) *Territorio*: spazio antropizzato, prodotto socio-culturale di una comunità (economia, religione, potere, miti), collegato alla geostoria e all'azione storica dell'uomo sullo spazio.
e) *Paesaggio*: sintesi del territorio percepito dalle popolazioni, integra dimensioni materiali e immateriali, consentendo una lettura diacronica di passato, presente e futuro.

Attraverso questi concetti, la letteratura è concepita come archivio di esperienze, traumi e trasformazioni, utile a comprendere dinamiche storiche, sociali e culturali, con particolare attenzione alla geografia dei luoghi narrativi e al rapporto tra dimensione locale, nazionale e mediterranea, integrando la geostoria digitale e la didattica attiva del progetto (Mattozzi, Rabuiti, Di Tonto, 2012).

5. DIDATTICA, DH E PUBLIC ENGAGEMENT

L'integrazione con le Digital Humanities trasforma i testi censiti da MeMo in strumenti didattici e di partecipazione pubblica: attraverso codifica XML-TEI, mappe interattive, visualizzazioni geospaziali e unità di apprendimento modulari, studenti, insegnanti e cittadini possono esplorare relazioni tra autori, luoghi e tematiche, costruendo percorsi interpretativi personalizzati e favorendo un apprendimento attivo, laboratoriale e inclusivo. Seguendo il modello delle UDA di geostoria sviluppate da (Rocca, 2012), ciascun percorso si articola nelle tre fasi Presente – Passato – Ritorno al presente, in cui la geografia fornisce strumenti analitici per leggere il presente, la storia indaga i processi e le trasformazioni nel tempo, attraverso il ricorso alle esperienze letterarie, e la riflessione finale ricollega conoscenze e competenze a problemi contemporanei, promuovendo consapevolezza critica e capacità di progettare interventi di cittadinanza attiva.
La struttura modulare delle UDA, la loro autosufficienza e trasferibilità, così come l'integrazione di risorse digitali (manuali, archivi online, carte tematiche, dati statistici, immagini e mappe interattive) consentono la riutilizzabilità dei materiali per la didattica universitaria e secondaria, la sperimentazione laboratoriale e la costruzione di percorsi personalizzati. Le attività didattiche privilegiano la costruzione di competenze trasversali, includendo la capacità di tematizzare e localizzare fenomeni, produrre e rielaborare dati, comunicare informazioni geostoriche in forma multimediale, analizzare testi e fonti diversificate e utilizzare consapevolmente strumenti digitali per ricerca, mappatura e rappresentazione delle conoscenze (*Data Visualization*). Il modello prevede anche unità di geostoria locale, in cui l'osservazione diretta dei

paesaggi urbani e territoriali, la visita a musei e la schedatura delle informazioni diventano esperienze formative concrete, capaci di sviluppare senso di responsabilità, collaborazione, curiosità e cittadinanza attiva (vd. Fig.2).



Figura 2. Workflow metodologia didattica MeMo (Immagine creata con NotebookLM)

Nel progetto MeMo, la dimensione didattica e di *Public Engagement* si colloca oltre i confini dell'educazione formale, intercettando pratiche di apprendimento non formale e informale rese oggi centrali dall'ecosistema digitale. La disponibilità online dei materiali letterari, cartografici e metadatati amplia in modo significativo le occasioni di accesso alla conoscenza, favorendo processi di apprendimento diffuso e continuo che coinvolgono studenti, docenti, lettori non specialisti e cittadini. In questo quadro, le *Digital Humanities* non operano come meri strumenti di semplificazione o divulgazione, ma come ambienti di mediazione critica, capaci di riorientare il rapporto tra sapere esperto e pubblico, in senso più dialogico e democratico (Panciroli, 2021). La fruizione digitale dei testi e delle mappe di MeMo consente infatti di esplorare in autonomia continuità e discontinuità storiche, relazioni spaziali e trasformazioni ambientali, promuovendo una partecipazione attiva alla costruzione del significato e superando modelli comunicativi verticali fondati su una trasmissione unidirezionale del sapere.
In questa prospettiva, MeMo si colloca all'interno del paradigma dell'*engagement*, inteso non come semplice diffusione di contenuti, ma come pratica di co-costruzione della conoscenza e di responsabilizzazione culturale. L'infrastruttura digitale del progetto crea uno spazio condiviso in cui rigore scientifico e accessibilità dialogano, permettendo di problematizzare i testi letterari e i contesti storici anziché ridurli a narrazioni semplificate o esemplari. La valorizzazione di voci minoritarie, marginali o poco note al grande pubblico diventa così un'azione culturale e civica, che contribuisce a riequilibrare il rapporto tra centro e periferia, tra canone e corpus, e a rafforzare una consapevolezza critica del patrimonio letterario e territoriale.
Attraverso strumenti digitali aperti, interoperabili e riutilizzabili, coerenti con i principi della scienza aperta e FAIR, MeMo trasforma l'analisi letteraria in un laboratorio di *Public Humanities*, dove didattica attiva, ricerca e partecipazione pubblica si intrecciano. La letteratura diventa così un dispositivo di lettura del presente e del territorio, capace di attivare competenze critiche, storiche e digitali, e di favorire una cultura del patrimonio inteso non come deposito statico, ma come spazio dinamico di confronto, memoria condivisa e cittadinanza culturale.

6. PROPOSTA DI PERCORSO DIDATTICO FONDATO SULLA METODOLOGIA MEMO

Sulla base dei criteri metodologici del progetto MeMo, il percorso didattico "Paesaggi della Memoria: il Mezzogiorno tra letteratura e trasformazione digitale" utilizza il modello delle Unità di Apprendimento (UDA) di geostoria e le metodologie delle Digital Humanities per trasformare il testo letterario in un

archivio di esperienze sulle trasformazioni storiche, sociali e ambientali del Sud Italia dal secondo dopoguerra a oggi. L'attività si articola in tre fasi integrate:

I. *Fase 1* – Analisi del presente: osservazione diretta del territorio (aree industriali dismesse, borghi rurali, contesti urbani) tramite carte tematiche, dati statistici e immagini satellitari, per identificare segni visibili di industrializzazione e migrazioni.
II. *Fase 2* – Indagine del passato: utilizzo della letteratura come strumento di accesso alla memoria storica e ai traumi collettivi. Gli studenti esplorano il portale MeMo e la mappa letteraria interattiva, analizzano testi legati ai luoghi osservati o opere non incluse nel progetto, applicano la metodologia delle cinque "porte" geostoriche (spazio, luogo, ambiente, territorio, paesaggio) e sperimentano strumenti DH come la codifica XML-TEI per l'analisi di riferimenti spaziali, temi e toponimi.
III. *Fase 3* – Ritorno al presente: confronto tra memoria letteraria e percezione attuale del territorio, riflessione critica sulle voci marginali emerse dal corpus, produzione di visualizzazioni di dati o narrazioni geospaziali e restituzione pubblica dei risultati alla comunità locale, promuovendo cittadinanza attiva e valorizzazione del patrimonio letterario come risorsa viva (Granata, 2013; Berger, 2022).

Il percorso (vd. Fig.3) favorisce lo sviluppo di competenze trasversali, tra cui alfabetizzazione digitale tramite strumenti di mappatura geospaziale e codifica dei testi, pensiero critico nell'analisi storica e letteraria, e cittadinanza culturale attiva nella tutela e valorizzazione del patrimonio territoriale e letterario.



Figura 3. Struttura UDA geostoria, secondo la metodologia MeMo (Immagine creata con ChatGPT)

Sulla base di questa metodologia, si potrebbe elaborare un'UDA su Giuseppe Bonaviri e l'entroterra siciliano, grazie alla quale gli studenti potrebbero leggere il paesaggio come archivio vivente di memoria storica, sociale ed ecologica, in coerenza con MeMo (Zangrilli, 1998; Musarra, 1999; Carta, 2006). I testi selezionati diventerebbero così strumenti per intrecciare letteratura, geografia e Digital Humanities, integrando dati territoriali, codifica dei testi e riflessione ecocritica.

BIBLIOGRAFIA

Abbate, M. (1968). L'alternativa meridionale. Roma-Matera: Basilicata Editrice.

Alfano, G. (2010). Paesaggi, mappe, tracciati. Cinque studi su letteratura e geografia. Napoli: Liguori.

Assmann, A. (2004). Ricordare. Forme e mutamenti della memoria culturale. Bologna: Il Mulino.

Berger, J. (2022). Ways of seeing. London: The British Broadcasting Corporation-Penguin Books.

Carta, A. (2006). Giuseppe Bonaviri. Le forme del racconto tra memoria e utopia fantastica. Palermo: Gruppo Editoriale Kalos.

De Lucia, S., Gallo, C., & Marino, D. (Eds). (2014). Landscapes and mindscapes: metodologie di ricerca, percorsi geocentrati e poetiche dello spazio in una prospettiva comparata. Napoli: Marchese.

Dell'Oglio, P., & Triolo, E. (2018). L'Archivio degli Iblei. Umanistica Digitale, 3, 85-104.

Giurdanella, L., Palazzolo, G., De Luca, B., Autiero, F., Gatto, M., Peluso, S., Pagliuca, C. M., & Schembari, A. (2025). MeMo. Una mappa letteraria digitale per la memoria del Mezzogiorno. Proceedings of XIV Convegno Annuale della AIUCD 2025, Verona 11-13 giugno 2025. Rebora, S., Rospocher, M., Bazzaco, S. (Ed.). 627-637. DOI 10.6092/unibo/amsacta/8380.

Granata, E. (2013). Lo sguardo e quello che guarda. Leggere, esplorare e riconoscere i segni del mondo. Le città invisibili. Per Giovanna Franci. Ragonese, R. (ed.). Ocula, 14, 3.

Iovino, S. (2006). Ecologia letteraria. Una strategia di sopravvivenza. Milano: Ambiente.

Marocco, M. (2014). Un luogo esiste solo se raccontato. La dimensione letteraria come strumento di indagine dello spazio periurbano. Territorio, 68, 1, 117-124.

Mattozzi, I., Rabuiti, S., & Di Tonto, G. (Eds). (2012). Il tema: Geostoria e Didattica. Il Bollettino di Clio, XIII, 0, maggio.

Moliterni, F. (2024). Finzioni meridionali. Il Sud e la letteratura italiana contemporanea. Roma: Carocci.

Musarra, F. (1999). Scrittura della memoria. Memoria della scrittura. L'opera narrativa di Giuseppe Bonaviri. Leuven-Firenze: Leuven University Press-Franco Cesati Editore.

Panciroli, C. (2021). Elementi di didattica post-digitale. Bologna: Bonomia University Press.

Pedullà, W. (2016). Il mondo visto da sotto. Narratori meridionali del '900. Soveria Mannelli: Rubbettino.

Peterle, G. (2016). Leggere, scrivere, abitare: proposte di ricerca e didattica tra geografia e letteratura. AST, LXI, 4, ottobre-dicembre, 26-30.

Rocca, L. (2012). Uno sguardo storico-geografico alle cinque porte della geografia. Il Bollettino di Clio, maggio, 9-14.

Scaffai, N. (2017). Letteratura e ecologia. Forme e temi di una relazione narrativa. Roma: Carocci.

Stuto, S. (2024). La sicilianitudine di Giuseppe Bonaviri, il cantore del mondo contadino amato da Vittorini, La Repubblica, 11 luglio 2024, https://www.repubblica.it/cultura/2024/07/11/news/la_sicilianitudine_di_giuseppe_bonaviri_il_cantore_del_mondo_contadino_amato_da_vittorini-423386691/.

Westfal, B. (2007). La Géocritique. Réel, fiction, espace. Paris: Éditions de Minuit.

Wilkinson, M., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N. et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3, 160018, https://doi.org/10.1038/sdata.2016.18.

Zangrilli, F. (1998). Sicilia isola-cosmo. Conversazione con G. Bonaviri. Ravenna: Longo Editore.

Una proposta di approccio computazionale alle geografie del cinema italiano

ALBERTO SAVI[1], LUCA GIOVANNINI[2]
[1] UNIVERSITÀ DEGLI STUDI DI MILANO / UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA – ALBERTO.SAVI@UNIMI.IT
[2] UNIVERSITÀ DI POTSDAM, GERMANIA - GIOVANNINI@UNI-POTSDAM.DE

ABSTRACT (ITALIANO)

Questo contributo modella la banca dati *Filmitalia* per estrarre e processare quasi 4000 sinossi di film italiani usciti negli ultimi 25 anni. I dati raccolti vengono utilizzati per affrontare questioni di ricerca relative alla rappresentazione della geografia nel cinema italiano, esaminando ad esempio menzioni e distribuzione dei toponimi, uso di lessico geografico e tasso di 'mobilità narrativa'.
Parole chiave: cinema italiano; film studies; modellazione; sinossi; Wikipedia

ABSTRACT (ENGLISH)

A computational approach to Italian film geographies.
This paper leverages the *Filmitalia* dataset to retrieve and perform text mining on about 4000 synopses from Italian films released in the last 25 years. The data collected is used to investigate the representation of geography within Italian cinema, examining e.g. toponym mentions and distribution, usage of geographic lexicon, and narrative mobility.
Keywords: Italian cinema; film studies; modelling; synopses; Wikipedia

1. INTRODUZIONE

Fin dalle origini, il cinema italiano ha svolto una funzione fondativa nell'immaginario nazionale (Brunetta, 2020; Patriarca, 2010), misurandosi con una geografia frammentata tra centralità e perifericità produttive (Morreale, 2015). Nel continuo sforzo di rappresentare questa diversità territoriale, il cinema italiano ha spesso esposto «conflitti irrisolti» (Zucconi, 2014: 499), restituendo un'immagine territoriale composita e un'identità collettiva in continua ricontrattazione.
Questa tensione tra unità e differenza, che esprime una ambivalenza strutturale tra un'idea nazionale omogenea e una più regionalizzata ed eterogenea, attraversa in profondità la storia culturale del Paese, ed è ben riassunta da Pietro Clemente quando osserva come l'immaginario italiano si sia costruito attorno al «nesso tra piccola e grande patria» (Clemente, 1997: 6). Tale nodo appare oggi particolarmente rilevante almeno su due piani. Da un lato, nel quadro di una globalizzazione intesa non come omogeneizzazione, ma come disgiunzione di flussi e immaginari (Appadurai, 1996), risulta evidente l'indebolimento dello Stato-nazione come cornice unificante (Giaccardi, 2000). Dall'altro lato, sul piano più strettamente disciplinare, questa rinnovata attenzione alla scala locale e subnazionale è presente anche all'interno del dibattito nei film studies, dove si discute la necessità di un *regional turn* capace di affiancare alle prospettive macro-comparative un i micro-mapping subnazionale (Marlow-Mann, 2017). Questa rinnovata prospettiva epistemologica, applicata al cinema italiano, ha informato diverse linee di ricerca che hanno esaminato ad esempio la regionalizzazione dei consumi audiovisivi (Avezzù, 2022) e dei sostegni alla produzione attraverso la diffusione delle *film commission* (Cucco e Richeri, 2013) e o la persistenza, sul piano testuale, di immaginari locali profondi (Zucconi, 2014: 495), paesani (Brunetta 2015: 296) e subnazionali.
Il presente contributo affronta il tema della regionalizzazione del cinema italiano con metodi computazionali, esaminando un componente del sistema-film finora largamente ignorato, ossia la sinossi 'ufficiale' rilasciata di volta in volta dalle case di produzione. Si tratta di un oggetto di studio affascinante, un 'epitesto editoriale' (Genette, 1989) che veicola contenuti autorizzati e istituzionalmente legittimati, capaci di condensare coordinate tematiche e culturali – incluse quelle geografiche – attraverso cui il film viene presentato, reso riconoscibile e posizionabile.
L'analisi qui presentata si basa dunque su un corpus di sinossi relative a tutti i film italiani distribuiti dal 2000 al 2025. Per evitare di ricorrere a fonti generaliste e anglocentriche (come l'*Internet Movie*

Database, *IMDb*) si è scelto di fare riferimento a una tra le più longeve, complete e sistematiche banche dati nel panorama nazionale, ossia *Filmitalia*.[1] Questa scelta risponde all'esigenza di disporre di una fonte stabile e comparabile su larga scala. Il portale cataloga costantemente tutti i lungometraggi (e cortometraggi) prodotti e distribuiti in Italia (e riconosciuti da criteri ministeriali). A ciascuna opera è dedicata una scheda che fornisce varie informazioni tra cui titolo (in italiano e nella versione internazionale), anno, regia, cast principale, durata, genere, crediti di produzione e distribuzione, nonché una sinossi. Le modalità di raccolta e validazione dei dati (incluse le sinossi) integrano fonti istituzionali, comunicazioni dirette con le società di produzione, comunicati stampa e materiali provenienti dal circuito festivaliero.[2]
Questo tipo di sinossi rappresentano una fonte preziosa in quanto rappresentative di cosa un film (o chi per lui) ritiene importante dire di sé, rispetto al suo posizionamento pubblico e quindi anche culturale. A queste informazioni verrà affiancate un altro tipo di paratesto, ossia i riassunti delle trame dei film presenti nelle corrispondenti pagine della Wikipedia italiana. In questa maniera intendiamo ricostruire, seppur con approssimazione, due tipi di identità geografica del film: una 'volontaria', espressa nelle sinossi fornite dalla case di produzione stesse, e una 'involontaria', che emerge da crowdsourcing collaborativo.

2. LINEE DI RICERCA E CORPUS

Le domande che guidano questo contributo sono due:

1. *Dove abita il cinema italiano tra 2000 e 2025?* Quali luoghi vengono nominati nelle sinossi, attraverso quali lessici (toponimi e lessici locativi), con quale granularità (locale/regionale/nazionale) e secondo quali gerarchie di visibilità?
2. *Come abita il cinema italiano?* Se la regionalizzazione può essere intesa come una frattura della nazione in unità subnazionali plurali, diventa rilevante osservare come le sinossi configurino geografie più stanziali o più mobili.

Per rispondere a questi interrogativi si è reso innanzitutto necessario convertire il portale *Filmitalia* in un database interrogabile con strumenti computazionali. A tal fine, un primo passaggio ha riguardato la raccolta (tramite *scraping* iterativo) di tutti i metadati presenti nelle schede dei 3762 film presenti su *Filmitalia*. I dati grezzi sono stati arricchiti attraverso il tool *Openrefine*, che ha permesso la riconciliazione dei titoli ai corrispettivi elementi Wikidata, e successivamente, attraverso un nuovo scraping delle relative voci Wikipedia, il recupero del riassunto della trama, quando disponibile.[3] Il dataset risultante[4] comprende 3.762 film, ciascuno con due campi testuali potenzialmente contenenti riferimenti geografici (sinossi e riassunti). Essi non si riferiscono alle location di ripresa, né mirano a censire in modo esaustivo tutte le ambientazioni (più direttamente recuperabili da sceneggiature, dialoghi o dall'analisi del testo filmico); sono piuttosto indicatori di dove un film ritenga importante (e, con buona probabilità, utile) dichiarare di "abitare" nel proprio posizionamento paratestuale. Si ricostruisce così non una geografia filmica "in senso forte", quanto una *geografia dichiarata*, nel convincimento che la menzione dei luoghi e le modalità di relazione con essi costituiscano elementi centrali di un posizionamento identitario, che contribuisce a definire il pubblico implicito dell'opera e alcune coordinate della geografia culturale del cinema italiano. Operativamente, questi indicatori testuali vengono trattati come dispositivi di localizzazione e immaginazione geografica misurabili attraverso toponimi (in senso ampio) e liste di parole legate a luoghi antropici e naturali.
Per quanto riguarda l'estrazione dei toponimi, si è utilizzato uno dei modelli *spaCy* per l'italiano,[5] catalogando tutte le entità classificate come 'locations' (LOC). Un'ispezione manuale dei risultati ha rivelato come la performance del modello non fosse adeguata, sia a livello di tokenizzazione, sia a livello

[1] https://filmitalia.org/. La banca dati e il relativo sito sono gestiti da Cinecittà S.p.A, società pubblica controllata dai ministeri della Cultura e dell'Economia e Finanze.
[2] Gli autori ringraziano Giovanni Marco Piemontese, redattore e curatore del database, per la disponibilità e per le preziose informazioni fornite.
[3] Circa il 60% dei film presenti su *Filmitalia* possiedono anche una pagina Wikipedia, ma solo un quarto del totale (889) prevede al suo interno una sezione "Trama", oggetto di *scraping*.
[4] Disponibile in formato CSV all'indirizzo: https://anonymous.4open.science/r/modelling-italian-films-2337/filmitalia_geocoded.csv.
[5] https://github.com/explosion/spacy-models/releases/tag/it_core_news_sm-3.8.0.

di effettivo riconoscimento di termini geografici. Si è dunque provveduto a una ulteriore pulizia semi-automatica, rimuovendo una serie di *stopwords* e scremando tutti i falsi positivi con il supporto di un modello linguistico di grandi dimensioni (Claude Opus 4.5) e successiva revisione manuale. A ognuno dei 1427 toponimi univoci e validati sono state infine abbinate coordinate geografiche mediante il *geocoder Nominatim*. Ne è risultato un database in formato CSV contenente, per ciascuno dei film, i metadati originali arricchiti con le liste di toponimi estratti e le rispettive coordinate geografiche.

3. ANALISI E PROSPETTIVE

Uno sguardo preliminare alla distribuzione dei toponimi estratti nelle due fonti evidenzia come le sinossi presentino un repertorio di località complessivamente più ampio rispetto alle trame da Wikipedia.[6] Questo è coerente con l'ipotesi introduttiva secondo cui il paratesto editoriale – inteso come posizionamento identitario del film stesso – faccia del situarsi geograficamente un fatto centrale, e venga dunque considerato rilevante nel framing pubblico dell'opera da parte dei suoi produttori o distributori.
La prima domanda di ricerca, ovvero *dove* abiti il cinema italiano, può essere affrontata sia in maniera più sfumata, attraverso *heatmaps* che mostrano l'intensità delle menzioni di un dato toponimo, sia attraverso i dati statistici grezzi (Figura 1). Il primo approccio offre una prospettiva cartografica sul territorio di competenza del cinema italiano contemporaneo, con un prevedibile rilievo dell'area nazionale e europea, oltre a presenze rilevanti anche nel continente africano e americano. La lista dei toponimi più frequenti conferma invece la forte centralità di Roma, coerente con la sua preminenza nell'immaginario e la sua storica centralità a livello di produzioni (Morreale, 2015) Allo stesso tempo, il profilo complessivo delle cinque principali città italiane presenti (Roma, Napoli, Milano, Torino, Palermo), simile tra sinossi e Wikipedia, è compatibile con l'attuale distribuzione demografica della popolazione italiana, mostrando come il cinema italiano, in termini puramente distributivi, abiti l'Italia nelle stesse coordinate dei suoi cittadini.

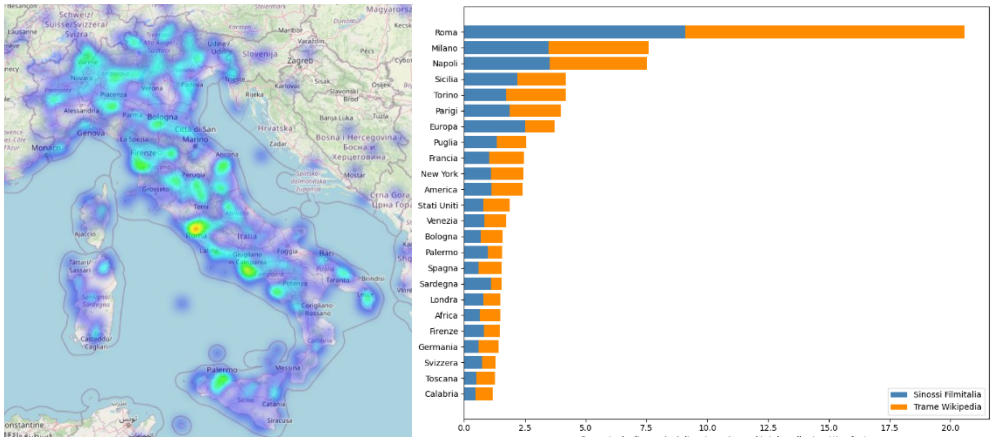


Figura 1. Statistiche di base. A sinistra: mappa di calore con la distribuzione dei toponimi sul territorio italiano. A destra: istogramma dei venti toponimi più frequentemente menzionati.

Accanto ai toponimi urbani emergono ulteriori riferimenti a scale diverse – sovranazionale (Europa, Africa, America), nazionale (Francia, Stati Uniti, Svizzera, Germania e Spagna) e subnazionale (Sicilia, Puglia, Sardegna, Toscana e Calabria) – che evidenziano un immaginario territorialmente stratificato e transcalare. I toponimi sovranazionali e nazionali più citati delineano l'orizzonte culturale del cinema italiano, che se guarda fuori dai suoi confini lo fa verso sud (Africa) e ovest (America, Stati Uniti): due realtà geopolitiche da cui spesso i personaggi principali arrivano e verso dove partono. Ne sono esempio, per l'Africa, le ricostruzioni drammatiche della rotta dei migranti (*Io capitano*, Garrone 2023; *Terraferma*, Crialese 2011), ma anche di contro-migrazioni declinate nella satira (*Tolo Tolo*, Zalone 2020) o nella ricostruzione storica (*Le rose del deserto*, Monicelli 2006). In direzione speculare, l'America si configura ancora come terra di opportunità verso cui si orientano parte degli immaginari migratori nazionali, sia in contesti storici (*Nuovomondo*, Crialese 2006; *Napoli–New York*, Salvatores 2024), sia in cornici

[6] Toponimi presenti solo nelle sinossi *Filmitalia*: 56,9%; solo nelle trame *Wikipedia*: 18,6%; in entrambe le fonti: 24,5%.

contemporanee (*My name is Tanino*, Virzì 2002) o in chiave turistica (*Divorzio a Las Vegas*, Carteni 2020), fino al movimento opposto del turista americano in Italia (*La mia vita a stelle e strisce*, Ceccherini 2003; *To Rome with Love*, Allen 2012).
Contemporaneamente emerge un estero continentale europeo con una maggiore varietà di toponimi nazionali (oltre a Europa e città come Parigi e Londra, nazioni come Francia, Stati Uniti, Svizzera, Germania e Spagna). Ciò riflette una filmografia in cui l'Europa è un repertorio di nazioni che, di film in film, possono diventare luoghi di lavoro (*Una vita tranquilla*, Cupellini 2010), di fuga (*Dopo la guerra*, Zambrano 2017), di "esilio" (*Cetto c'è, senzadubbiamente*, Manfredonia 2019), di turismo (*Natale a Londra. Dio salvi la Regina*, De Biasi 2016) o ambientazioni storiche di confine (*Lubo*, Diritti 2023).
In questo quadro, la relativa bassa frequenza di "Italia" può essere letta come un effetto di focalizzazione: la scala nazionale tende a rimanere implicita in quanto l'ancoraggio si effettua a livello urbano o regionale. Tra i toponimi subnazionali (Sicilia, Puglia, Sardegna, Toscana e Calabria) ricorrono invece regioni i cui capoluoghi non compaiono tra le città più citate. Questo può indicare che, in alcuni casi, l'identificazione geografica opera per "area" (la regione come unità simbolica o come termine ombrello per contenere più unità), mentre in altri casi per specifiche centralità urbane. Per esempio, se in un film come *Anna* (Amenta 2023) la localizzazione è sarda, in *La grande bellezza* (Sorrentino 2013) l'ancoraggio passa invece attraverso la metropoli romana, più che attraverso una cornice regionale; analogamente, se *Una femmina* (Costabile 2022) si colloca in "un paesino della Calabria", *Napoli velata* (Özpetek 2017) impiega la sola città come riferimento identitario.
In una fase successiva, si è provveduto a categorizzare ciascuno dei toponimi estratti secondo una semplice tassonomia in dodici categorie[7] al fine di verificarne la granularità. Nel grafico aggregato delle sinossi e trame (Figura 2) prevalgono nettamente gli *insediamenti* (come ad esempio nomi di città o paesi), indicando che il posizionamento geografico del cinema italiano, a livello paratestuale, risulta soprattutto ancorato a coordinate precise, localizzate e riconoscibili. Seguono due scale più ampie come *region*i, intese soprattutto come categorie subnazionali interne (es. Sicilia, Toscana) e *nazioni*, che funzionano prevalentemente come riferimenti esterni (es. Francia, Svizzera, Germania).
Nel complesso, la tassonomia conferma come l'ancoraggio dominante sia quello micro-topografico degli insediamenti, mentre regioni e nazioni intervengono come livelli di generalizzazione e apertura dell'orizzonte geografico.[8]

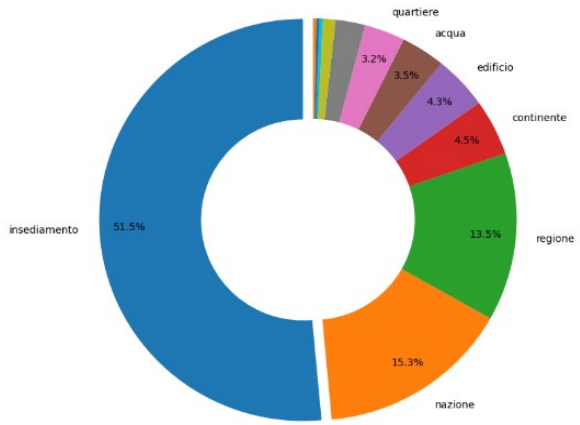


Figura 2. Distribuzione delle dodici tipologie di località.

Al di là dei toponimi propriamente detti, le sinossi e le trame contengono anche un più ampio lessico di tipo geografico, costituito da termini legati all'aspetto antropico ('città', 'periferia', 'piazza'…) e a quello

[7] Categorie impiegate: 'settlement', 'nation', 'region', 'continent', 'building', 'water', 'neighborhood', 'elevation', 'street', 'valley' 'generic_area', 'natural_area'. La classificazione è stata condotta in maniera semi-automatica (assegnazione tramite LLM e revisione manuale).
[8] La salienza dei toponimi regionali vista in precedenza dipende anche da questo effetto compositivo: la "regione" funziona come etichetta-ombrello che aggrega molte micro-menzioni di insediamenti, ciascuna poco ricorrente se presa singolarmente.

naturale (‘mare’, ‘montagna’, ‘isola’...). Al fine di meglio delineare il ‘paesaggio’ dei cinema italiano si è dunque provveduto a estrarre le occorrenze di una serie di liste di parole semanticamente interessanti. I risultati (Figura 3) mostrano come i termini del lessico ‘di insediamento’ occorrano con frequenza maggiore rispetto a quelli del lessico ‘ambientale’. Nel complesso, queste evidenze sono coerenti con quanto emerge dall’analisi toponimica e suggeriscono un cinema che si identifica in prevalenza con le sue geografie umane (città/paesi/quartieri) rispetto a quelle ambientali (mare, campagna, ecc.).

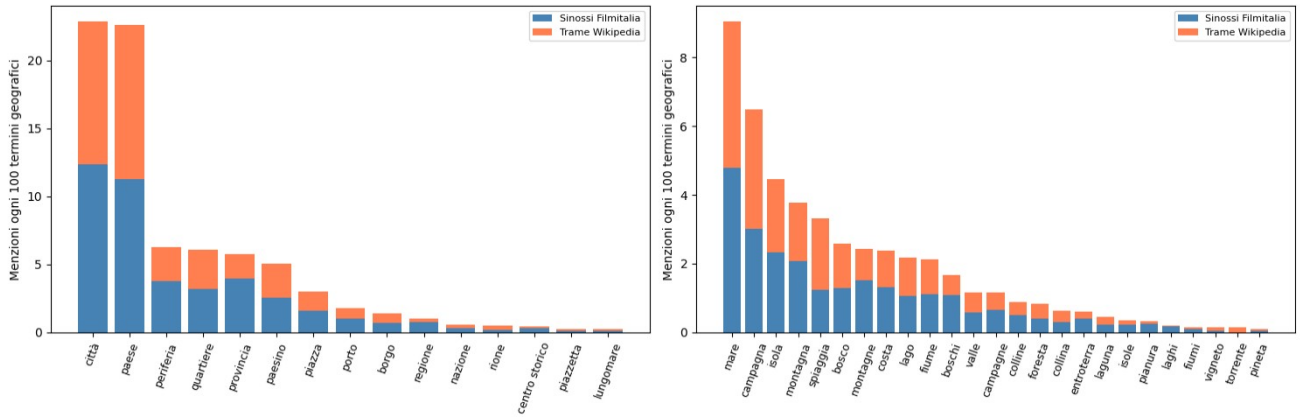


Figura 3. Frequenze relative di termini del lessico di insediamento (a sinistra) e ambientale (a destra).

Per quanto riguarda invece la seconda domanda di ricerca – *come* abita il cinema italiano? – si è ritenuto opportuno investigare se le sinossi configurino geografie narrative prevalentemente stanziali oppure orientate alla mobilità geografico-culturale.
Per operazionalizzare questa dimensione abbiamo adottato ancora una volta un approccio *dictionary-based*, distinguendo due campi semantici: da un lato uno legato alla stanzialità (intesa come radicamento, ambientazione e permanenza, e espressa da verbi quali ‘vivere’, ‘abitare’, ‘risiedere’, ‘stabilirsi’, ‘svolgersi’), dall’altro uno legato alla nozione di movimento (intesa come transizione e spostamento, e espressa da verbi quali ‘partire’, ‘arrivare’, ‘tornare’, ‘trasferirsi’, ‘fuggire’, ‘viaggiare’).
Per aumentare la precisione dell’estrazione, si è imposto un vincolo di ‘ancoraggio geografico’, includendo solo quei verbi che compaiono in prossimità di un toponimo riconosciuto (entro una finestra di 3 tokens), così da ridurre usi generici o metaforici del lessico (ad es. “fuggire dal lavoro”) e privilegiare occorrenze effettivamente locative (“fuggire da Roma”).
Per ogni sinossi si è poi calcolato un indice di movimento, basato sulla frequenza dei termini stessi e normalizzato da −1 (solo espressioni di stanzialità) a +1 (solo espressioni di movimento). La distribuzione dell’indice (Figura 4) si dimostra nettamente polarizzata: molti casi si collocano agli estremi (vicino a -1 e a +1), mentre risultano relativamente pochi i film in prossimità dello zero. In altri termini, quando nelle sinossi compare un lessico geograficamente ancorato, questo tende a essere dominato da una sola modalità: o prevale un senso di stanzialità, oppure uno di movimento, mentre le configurazioni ibride (che combinano in modo equilibrato stanzialità e mobilità) appaiono minoritarie. Questa polarizzazione è coerente con la natura selettiva dei paratesti brevi – spesso un singolo marcatore, collocato in apertura (“si svolge a…”, “parte per…”), è sufficiente a orientare il posizionamento spaziale del racconto – ma riflette anche l’urgenza, da parte dell’estensore, di trasmettere una certa peculiarità del film stesso.

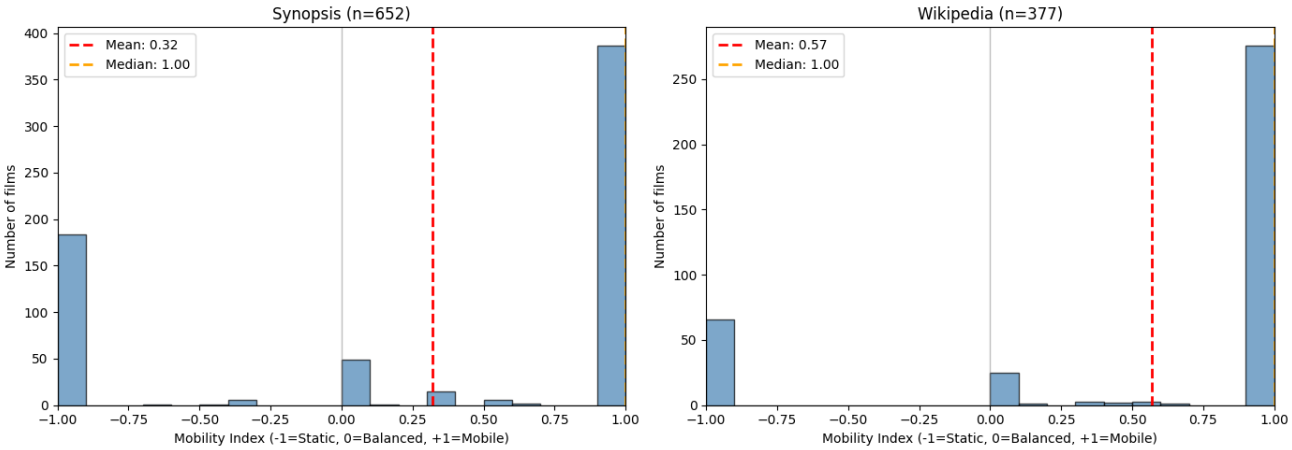


Figura 4. Distribuzione dell'indice di mobilità nelle sinossi (a sinistra) e trame da Wikipedia (a destra).

Come evidenziato dalla linea di media, i termini di movimento sono decisamente più frequenti rispetto a quelli stanziali, sia nelle sinossi sia nelle trame Wikipedia, suggerendo che la mobilità costituisca un dispositivo descrittivo ricorrente per rendere leggibile la geografia narrativa dei film. In questa prospettiva, il cinema italiano – almeno nel suo posizionamento paratestuale – si presenta come meno "stanziale" e chiuso in «microsocietà il più possibile omogenee» (Eugeni, 2016: 400) di quanto si possa presupporre. La dimensione dell'attraversamento e della transizione emerge dunque come risorsa frequente di posizionamento, accanto agli ancoraggi più propriamente insediativi. Tuttavia, questi risultati non escludono che, proprio in quanto paratesti editoriali, le sinossi tendano a privilegiare una strategia discorsiva più orientata all'idea di percorso e di 'viaggio dell'eroe' come schema narrativo sintetico e riconoscibile (Vogler, 1999).
In conclusione, se la geografia narrativa del cinema italiano si organizza spesso in termini di mobilità, un passo successivo consisterà nel ricostruire le traiettorie implicate da questo movimento e individuare le direzioni prevalenti degli spostamenti e verificare se producano effettivamente connessioni ricorrenti tra luoghi. In questa prospettiva, tali connessioni possono essere lette come dispositivi capaci di collegare porzioni del territorio e di generare, per aggregazione, cluster territoriali interpretabili come macroregioni identitarie o immaginate. Parallelamente, poiché l'analisi dei toponimi evidenzia un posizionamento spaziale consistente, potrebbe divenire utile misurarne l'intensità in modo comparabile, ad esempio attraverso lo sviluppo di un indicatore sintetico - un "coefficiente di località" - che riunisca densità, granularità e struttura delle menzioni geografiche, trasformando il posizionamento nello spazio in una variabile misurabile e confrontabile.

BIBLIOGRAFIA

Appadurai, A. (1996). *Modernità in polvere*. Raffaello Cortina Editore.

Avezzù, G. (2022). *L'Italia che guarda: Geografie del consumo audiovisivo*. Carocci.

Brunetta, G. P. (2015). *Il cinema neorealista italiano: Storia economica, politica e culturale*. Laterza.

Brunetta, G. P. (2020). *L'Italia sullo schermo: Come il cinema ha raccontato l'identità nazionale*. Carocci.

Clemente, P. (1997). Paese/paesi. In M. Isnenghi (Ed.), *I luoghi della memoria: Struttura ed eventi dell'Italia unita* (pp. 5–39). Laterza.

Cucco, M., & Richeri, G. (2013). *Il mercato delle location cinematografiche*. Marsilio.

Eugeni, R. (2016). Vacanza. In R. De Gaetano (Ed.), *Lessico del cinema italiano: Forme di rappresentazione e forme di vita* (Vol. 3, pp. 359–403). Mimesis.

Genette, G. (1989). *Soglie: I dintorni del testo*. Einaudi.

Giaccardi, C. (2000). L'identità nazionale nell'era della globalizzazione. *Studi di Sociologia*, *38*(2), 113–129.

Marlow-Mann, A. (2017). Regional cinema: Micro-mapping and glocalisation. In R. Stone, P. Cooke, S. Dennison, & A. Marlow-Mann (Ed.), *The Routledge companion to world cinema* (pp. 323–336). Routledge.

Morreale, E. (2015). Le Italie del cinema: Centralità di Roma, immaginari locali e decentramenti produttivi. In *L'Italia e le sue regioni: Vol. 3. Società* (pp. 505–521). Treccani.

Patriarca, S. (2010). *Italianità: La costruzione del carattere nazionale*. Laterza.

Vogler, C. (1999). *Il viaggio dell'eroe: La struttura del mito ad uso di scrittori di narrativa e di cinema*. Dino Audino.

Zucconi, F. (2014). Geografia. In R. De Gaetano (Ed.), *Lessico del cinema italiano: Forme di rappresentazione e forme di vita* (Vol. 1, pp. 433–503). Mimesis.

Il campo semantico del fuoco nella Commedia: un caso di studio sulla pertinenza lessicale e sulla visualizzazione

RUOCI SONG
CAMBRIDGE UNIVERSITY

ABSTRACT (ITALIANO)

Questo contributo esamina il motivo del fuoco nella *Commedia* dantesca come caso filologicamente controllato di visualizzazione dell'intensità tematica. Il lavoro nasce nel contesto delle ricerche sull'"antica fiamma" e sui rimandi fra Virgilio e Dante, e concentra l'attenzione sulla distribuzione interna del fuoco nel poema. Il lessico del fuoco prende forma da una ricognizione di lessemi effettivamente presenti nella *Commedia*, discussi alla luce della tradizione filologica, in particolare Pertile (1991), e organizzati in tre livelli di pertinenza per fini operativi e interpretativi. Su questa base, il modello assegna a ogni verso un punteggio di intensità del fuoco combinando lessicografia gerarchica, sentence embeddings e una finestra contestuale di tre versi. La sequenza dei valori viene trasformata in una heatmap che rende visibili concentrazioni, gradienti e anomalie locali. La visualizzazione conferma la trasformazione del fuoco dal registro punitivo dell'Inferno alla luce e all'ardore del Paradiso e mette in evidenza micro-pattern da restituire alla lettura ravvicinata e alla discussione esegetica. L'integrazione del modulo nell'ambiente già esistente di commento digitale ddpcommentary.com rende inoltre ispezionabili e documentabili le scelte lessicali e i parametri locali.
Parole chiave: visualizzazione testuale; Dante Alighieri; motivi poetici; rappresentazioni vettoriali; lettura ravvicinata

ABSTRACT (ENGLISH)

The Semantic Field of Fire in the *Commedia*: A Case Study in Lexical Pertinence and Visualisation
This paper examines the fire motif in Dante's *Commedia* as a philologically controlled case of thematic-intensity visualisation. It grows out of earlier work on the "antica fiamma" and on Virgil-Dante echoes, and focuses on the internal distribution of fire across the poem. The fire lexicon is constructed from a survey of lexical items actually attested in the *Commedia*, discussed in light of philological scholarship, especially Pertile (1991), and organised into three levels of pertinence for operational and interpretive purposes. On this basis, the model assigns a fire-intensity score to each line by combining a hierarchical lexicon, sentence embeddings, and a three-line context window. The resulting sequence is transformed into a heatmap that makes concentrations, gradients, and local anomalies visible. The visualisation confirms the transformation of fire from the punitive register of the Inferno to the light and ardour of the Paradiso and highlights micro-patterns that can return to close reading and exegetical discussion. The integration of the fire module into the already existing digital commentary environment ddpcommentary.com further makes lexical choices and local parameters inspectable and documentable.
Keywords: textual visualisation; Dante Alighieri; poetic motifs; sentence embeddings; close reading

1. INTRODUZIONE

Il motivo del fuoco attraversa l'intera *Commedia*. Nell'Inferno brucia i corpi dei dannati, nel Purgatorio purifica sulla soglia del Paradiso terrestre, nel Paradiso si trasforma in luce e ardore spirituale (Barolini, 1992; Singleton, 1958). Questa tripartizione è ben nota agli studi danteschi; resta invece più difficile descrivere con precisione la distribuzione interna del motivo, cioè i suoi addensamenti, le sue interruzioni e le sue riemersioni a livello di canti e di versi.
Nel quadro dell'impiego di metodi quantitativi in letteratura, diverse ricerche hanno affiancato alla close reading un'osservazione di lungo raggio dei testi e dei corpora (Jockers, 2013; Moretti, 2013; Underwood, 2019). In ambito dantesco, il Dartmouth Dante Project (Hollander, 1988-) mette a disposizione testi digitali strutturati e un ricco apparato di commenti che rendono particolarmente praticabili esperimenti computazionali sulla *Commedia*.
Song (2025) ha proposto una mappa dell'intertestualità fra Virgilio e Dante, seguendo la linea dell'"antica fiamma" nella *Commedia* attraverso strumenti di allineamento testuale e misure vettoriali. Il presente contributo si colloca in continuità con quel lavoro e concentra l'analisi sul solo poema, assumendo il fuoco

come caso di studio per osservare la distribuzione di un motivo lungo l'intera *Commedia*. Il saggio mira a rendere visibili gradienti, concentrazioni e anomalie del fuoco e a riportarli poi alla lettura ravvicinata del testo. Il caso del fuoco offre un banco di prova particolarmente solido: il gradiente generale è in larga parte atteso, mentre le variazioni locali permettono di valutare la capacità del modello di segnalare punti che meritano un controllo esegetico.
Un ancoraggio filologico essenziale è offerto da Pertile (1991), che ricostruisce il campo semantico del fuoco e del calore nella *Commedia*, elenca diversi lessemi con la loro distribuzione nelle tre cantiche e descrive la trasformazione semantica e teologica che conduce dal fuoco infernale al fuoco d'amore e di conoscenza del Paradiso. Questa linea di lettura offre un riferimento essenziale per comprendere perché il fuoco sia un caso di studio privilegiato e per giustificare l'organizzazione operativa del lessico adottata qui. Il modello presentato nelle pagine che seguono emerge dunque dal testo e vi ritorna: parte da una ricognizione lessicale interna alla *Commedia*, la riorganizza in forma esplicita e la rimette alla prova attraverso la visualizzazione.
Le visualizzazioni che ne derivano vanno lette come costruzioni interpretative, in dialogo con le riflessioni di Drucker sulla natura congetturale dei grafici umanistici (Drucker, 2011, 2014), con il dibattito sulla modellizzazione dei testi (Flanders & Jannidis, 2019; Rockwell & Sinclair, 2016) e con gli interrogativi epistemologici posti dagli strumenti digitali in area italiana (Ciotti, 2023; Fiormonte, Numerico, & Tomasi, 2010). L'ambiente già esistente di ddpcommentary.com offre inoltre un contesto concreto in cui il modulo sul fuoco può essere reso ispezionabile, con scelte lessicali, pesi e finestre contestuali documentati come parte del lavoro interpretativo.[1]

2. DAL MOTIVO AL PUNTEGGIO DI INTENSITÀ

2.1 Corpus e lessicografia gerarchica

Il corpus di lavoro è costituito dall'edizione Petrocchi della *Commedia* (Alighieri, 1966-1967), segmentata in versi e accompagnata da metadati minimi: cantica, numero di canto e posizione del verso. Il lessico del fuoco è organizzato in tre livelli che riflettono una diversa prossimità al fenomeno della combustione.
La costruzione del lessico rappresenta il passaggio metodologico decisivo dell'intero lavoro. Il repertorio prende forma attraverso una ricognizione dei lessemi effettivamente attestati nella *Commedia* e pertinenti al campo del fuoco. La procedura non ricorre a una generazione automatica a partire da reti semantiche generali di tipo WordNet e lascia ugualmente da parte la definizione di un elenco astratto separato dal corpus. Il punto di partenza coincide quindi con il testo dantesco, sottoposto a vaglio interpretativo e confrontato con la discussione filologica sul fuoco e sul calore nella *Commedia*. In questo quadro, Pertile (1991) offre un ancoraggio particolarmente utile: ricostruisce il campo semantico del fuoco-calore, registra una serie concreta di lessemi, ne osserva la distribuzione nelle tre cantiche e segue la trasformazione che conduce dal fuoco punitivo dell'Inferno al fuoco metaforico, affettivo e conoscitivo del Paradiso. Il lessico adottato in questo studio nasce dunque dal testo della *Commedia* e trova in Pertile un riferimento di controllo e orientamento filologico.
Su questa base, i termini sono stati riorganizzati in tre livelli di pertinenza per fini insieme operativi e interpretativi. Il campo semantico del fuoco assume così la forma di una costruzione esplicita, documentata, vincolata dal testo e interpretativamente negoziata.
Il primo livello comprende termini che indicano direttamente il fuoco fisico: fuoco/foco, fiamma, incendio, brace, insieme ai relativi verbi. Il secondo livello raccoglie lemmi che esprimono calore e ardore, spesso in senso più astratto, come ardore, ardente e scintilla. Il terzo livello contiene forme graficamente affini ma referenzialmente autonome, come etnonimi o nomi propri (Fiamminghi, Focaccia).
Nel punteggio di base, i lemmi del primo livello contribuiscono in modo pieno, quelli del secondo con peso ridotto, mentre il terzo livello viene impiegato in fase esplorativa e poi escluso dal calcolo finale. La gerarchia consente di distinguere il fuoco descritto da semplici somiglianze grafiche e di riconoscere i casi in cui il motivo è evocato in forma più mediata o astratta. La scansione in livelli rende visibili i margini del campo semantico e trasforma una scelta interpretativa in un passaggio metodologico controllabile.
All'interno di ddpcommentary.com, questa stessa articolazione può diventare oggetto di documentazione e confronto, rendendo leggibili eventuali varianti di configurazione del lessico.

[1] La pagina di ricerca dedicata al motivo del fuoco è consultabile all'indirizzo: https://ddpcommentary.com/research/fiamma.html.

2.2 Sentence embeddings e finestra contestuale

La sola presenza di parole legate al fuoco non basta a definire l'intensità del motivo. Il modello integra quindi una componente distribuzionale basata su sentence embeddings. Ogni verso viene codificato mediante SentenceTransformer per l'italiano, una variante di Sentence-BERT (Reimers & Gurevych, 2019), e rappresentato come vettore in uno spazio normalizzato.
L'uso di un modello addestrato su italiano contemporaneo introduce un effetto di embedding drift rispetto al toscano trecentesco (Hamilton et al., 2016). Il framework assume questo scarto come compromesso metodologico locale e lo rende esplicito. Il lavoro cerca una geometria semantica interna sufficientemente coerente per confrontare versi della *Commedia*, e accompagna questa scelta con controlli qualitativi sui vicini semantici dei termini chiave, così da limitare collocazioni estranee al contesto dantesco.
Il vettore prototipico del fuoco viene costruito a partire da un insieme ristretto di versi scelti per la presenza esplicita del lessico del fuoco, per la loro coerenza contestuale locale e per la loro riconoscibilità nella tradizione esegetica, come il muro ardente di Purgatorio XXVII o le fiamme che avvolgono i consiglieri fraudolenti in Inferno XXVI. La similarità coseno tra questo vettore medio e quello di ciascun verso fornisce una componente distribuzionale del punteggio. Versi privi di lessico esplicito di fuoco possono così ricevere un valore positivo quando si collocano in contesti semanticamente affini, per esempio in descrizioni di luce intensa o di calore che corrode.
Come controllo qualitativo, l'elenco dei versi con il punteggio di fuoco più alto mostra un pattern coerente con questa scomposizione. Le prime posizioni sono occupate da passi del Paradiso e del Purgatorio nei quali il lessico del fuoco è particolarmente denso e il contesto mantiene un'atmosfera ardente lungo più linee, come "ogne livore, ardendo in sé, sfavilla" (*Par.* VII, 65), "Al mio ardor fuor seme le faville" (*Purg.* XXI, 94), "L'incendio suo seguiva ogne scintilla" (*Par.* XXVIII, 91). I versi infernali compaiono più in basso nella lista, con esempi di fuoco punitivo concentrato come "fiamme cadere infino a terra salde" (*Inf.* XIV, 33). La scomposizione del punteggio in componente lessicale, distribuzionale e contestuale mostra che, in tutti questi casi, il modello tende a privilegiare scene in cui il fuoco è nominato esplicitamente, semanticamente allineato al prototipo e sostenuto da un contesto immediato che ne prolunga l'effetto.
Il punteggio finale di un verso somma la componente lessicale e quella distribuzionale entro una finestra di tre versi: precedente, corrente e successivo. Questa scelta corrisponde a un compromesso contestuale locale. Molti motivi si dispiegano infatti su unità leggermente più ampie del singolo verso, pur restando ancorati a un passo riconoscibile e filologicamente controllabile. La sequenza di valori che ne deriva attraversa l'intera *Commedia* e costituisce la base delle visualizzazioni.

3. UNA HEATMAP ESEMPLIFICATIVA DEL FUOCO NELLA *COMMEDIA*

Fra le diverse forme possibili di visualizzazione, la heatmap offre la sintesi più compatta della distribuzione del motivo fra cantiche e canti. La Figura 1 mostra una heatmap esemplificativa costruita sui punteggi di intensità del fuoco.
Per renderla adatta a un contributo breve, la risoluzione è aggregata. Ogni cantica è suddivisa in dieci blocchi di canti contigui; per ciascun blocco si calcola la media dei punteggi dei versi che lo compongono. La matrice finale presenta così tre righe, corrispondenti alle cantiche, e trenta colonne, corrispondenti ai blocchi, con valori normalizzati su scala relativa.
Nell'Inferno compaiono blocchi ad alta intensità associati a episodi di fuoco punitivo concentrato, come le tombe infuocate degli eretici o le lingue di fuoco che avvolgono i consiglieri fraudolenti. Nel Purgatorio si osserva un netto innalzamento in corrispondenza del blocco che comprende il XXVII canto, dove il fuoco segna il passaggio verso il Paradiso terrestre. Nel Paradiso, invece, la heatmap mostra celle a intensità medio-bassa, spesso contigue, che suggeriscono una presenza più diffusa e luminosa del motivo, legata alla metaforica della luce e dell'ardore caritativo.
Questo gradiente generale è in linea con la progressiva trasformazione del fuoco descritta dalla tradizione critica; al tempo stesso, la matrice completa, non riprodotta qui, mette in evidenza micro-pattern meno intuitivi. Un primo caso riguarda il tratto di Inferno XXIV-XXV, nel quale la violenza metamorfica e narrativa non coincide con un addensamento altrettanto compatto del campo del fuoco: occorrenze come "com' el s'accese e arse, e cener tutto" (*Inf.* XXIV, 101) o "come procede innanzi da l'ardore" (*Inf.* XXV, 64) restano relativamente isolate, e il modello distingue così un immaginario acceso da una vera concentrazione lessicale e contestuale, che emerge con maggiore nettezza nei canti XXVI-XXVII. Un secondo esempio compare in Paradiso XXXI, canto meno canonico per il motivo del fuoco e tuttavia segnato da un rialzo locale attraverso formule come "Le facce tutte avean di fiamma viva" (*Par.* XXXI,

13) e "porgevan de la pace e de l'ardore" (*Par.* XXXI, 17), nelle quali il lessico del fuoco si riattiva entro la visione finale in modo meno prevedibile rispetto ai luoghi più attesi. Un breve rialzo si osserva anche in Purgatorio XXI: "Al mio ardor fuor seme le faville" (*Purg.* XXI, 94) e "che mi scaldar, de la divina fiamma" (*Purg.* XXI, 95) delineano un lessico del fuoco già interiorizzato, precedente alla soglia del XXVII. In questi casi la visualizzazione funziona come segnale di ritorno al testo: distingue addensamenti forti da presenze sporadiche, separa la violenza figurativa dalla coesione del campo del fuoco e richiama l'attenzione su punti che invitano a una verifica ravvicinata.

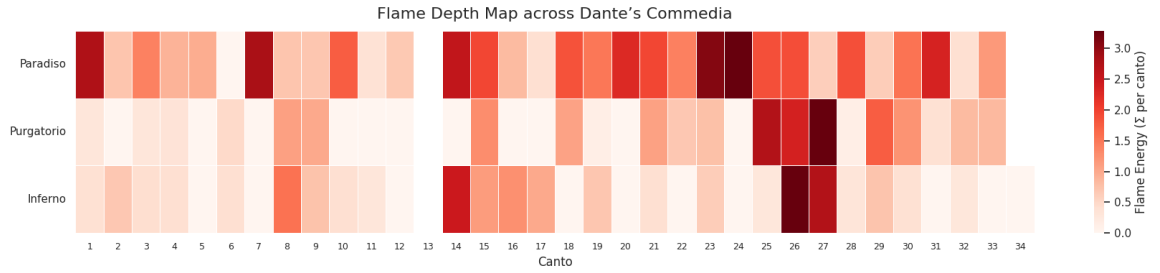


Figura 1. Heatmap esemplificativa dell'intensità del motivo del fuoco nella *Commedia*.
L'intensità cromatica indica il punteggio medio normalizzato per blocchi di canti in ciascuna cantica. Si osservano concentrazioni nell'Inferno in corrispondenza dei blocchi che comprendono le tombe infuocate degli eretici e i consiglieri fraudolenti, un picco presso il muro ardente di Purgatorio XXVII e una distribuzione più uniforme nel Paradiso. La figura suggerisce anche una distribuzione interna irregolare, utile per individuare rialzi locali e zone relativamente più fredde.

4. VISUALIZZAZIONI COMPLEMENTARI E AMBIENTE DOCUMENTARIO

Dalla stessa sequenza di punteggi si possono ricavare altre visualizzazioni. Una curva di intensità sui cento canti, in cui ogni punto rappresenta il punteggio medio di un canto, rende visibili gradini, plateau e picchi secondari. Nel caso del fuoco, la curva mostra il contrasto fra cantiche e l'emergere di zone di transizione interne, nelle quali l'immaginario del fuoco si raffredda o si riaccende in modo inatteso.
Le mappe semantiche, ottenute proiettando in due dimensioni il sottospazio lessicale connesso al fuoco, consentono di osservare la disposizione dei lemmi legati al fuoco infernale e di quelli più vicini all'ardore mistico. Una proiezione UMAP dello stesso sottospazio rende più leggibili alcuni addensamenti locali e distingue con maggiore evidenza i nuclei lessicali della combustione da quelli più prossimi all'ardore e alla luminosità spirituale. Letta alla luce di Pertile (1991), questa configurazione offre un supporto ulteriore nei punti in cui il passaggio da combustione, ardore e luce richiede una descrizione più fine.
In questa sede, la dimensione di ambiente digitale resta secondaria e assume una forma concreta: il modulo sul fuoco può essere integrato in ddpcommentary.com, che costituisce già un ambiente di commento e documentazione. In quel contesto, lessici, livelli di pertinenza, pesi e ampiezza della finestra contestuale possono essere resi visibili, confrontabili e registrabili come parte del lavoro interpretativo. Lo stesso ambiente consente di collegare le diverse configurazioni ai passi e ai punteggi che le sostengono. L'interfaccia interessa qui soprattutto come estensione della commentariazione e come spazio di ispezione delle scelte metodologiche.
Anche sul piano didattico, il valore di queste visualizzazioni dipende dalla loro capacità di accompagnare la lettura: aiutano a formulare ipotesi sulla distribuzione del motivo e a verificarle direttamente sul testo.

5. CONCLUSIONI

Il caso del fuoco mostra l'utilità di una visualizzazione dell'intensità tematica fondata su scelte lessicali esplicite, su livelli di pertinenza dichiarati e su un costante ritorno al testo. Il modello ricostruisce il gradiente che va dal fuoco punitivo alla luce paradisiaca e porta in superficie micro-pattern e anomalie locali capaci di aprire nuove domande esegetiche. Distingue così grandi concentrazioni, riattivazioni locali e presenze più isolate del lessico del fuoco.
La visualizzazione sostiene l'interpretazione, la orienta e talvolta la complica. Il vero snodo metodologico del lavoro risiede nella costruzione controllata del lessico e nella distinzione fra diversi livelli di pertinenza del campo semantico del fuoco, cioè nel punto in cui giudizio filologico e modellizzazione computazionale devono trovare un equilibrio esplicito.

Nel loro insieme, questi passaggi riportano la modellizzazione al suo punto di origine, il testo della *Commedia*. L'ambiente già esistente di ddpcommentary.com offre un luogo concreto in cui rendere ispezionabili e comparabili tali scelte come parte di una pratica di commento documentata. Il contributo principale dello studio consiste nel rendere osservabile il campo semantico del fuoco attraverso una costruzione lessicale esplicita e nel riportarne gli andamenti alla lettura ravvicinata.

BIBLIOGRAFIA

Alighieri, D. (1966–1967). La Commedia secondo l'antica vulgata (G. Petrocchi, Ed.). Mondadori.

Barolini, T. (1992). The undivine comedy: Detheologizing Dante. Princeton University Press.

Ciotti, F. (Ed.). (2023). Digital Humanities: Metodi, strumenti, saperi. Carocci. ISBN 9788829018437.

Drucker, J. (2011). Humanities approaches to graphical display. Digital Humanities Quarterly, 5(1). https://www.digitalhumanities.org/dhqdev/vol/5/1/000091/000091.html

Drucker, J. (2014). Graphesis: Visual forms of knowledge production. Harvard University Press.

Fiormonte, D., Numerico, T., & Tomasi, F. (2010). L'umanista digitale. Il Mulino. ISBN 9788815134257.

Flanders, J., & Jannidis, F. (Eds.). (2019). The shape of data in digital humanities: Modeling texts and text-based resources. Routledge. ISBN 9780367584030.

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (pp. 1489–1501). https://doi.org/10.18653/v1/P16-1141

Hollander, R. (1988–). Dartmouth Dante Project. https://dante.dartmouth.edu/

Jockers, M. L. (2013). Macroanalysis: Digital methods and literary history. University of Illinois Press. https://doi.org/10.5406/illinois/9780252037528.001.0001

Moretti, F. (2013). Distant reading. Verso.

Pertile, L. (1991). L'antica fiamma: La metamorfosi del fuoco nella Commedia di Dante. The Italianist, 11(1), 29–60. https://doi.org/10.1179/ita.1991.11.1.29

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 3982–3992). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410

Rockwell, G., & Sinclair, S. (2016). Hermeneutica: Computer-assisted interpretation in the humanities. MIT Press. https://doi.org/10.7551/mitpress/9780262034357.001.0001

Singleton, C. S. (1958). Dante studies. Harvard University Press.

Song, R. (2025). "Segni de l'antica fiamma": Mapping intertextuality between Virgil and Dante in the Divine Comedy. Annali d'italianistica, 43, 161–190. https://doi.org/10.65266/irst1090

Underwood, T. (2019). Distant horizons: Digital evidence and literary change. University of Chicago Press. https://doi.org/10.7208/chicago/9780226612973.001.0001

[8/C] Annotazioni e markup

Ritagli di teatro. Annotazione semantica della rassegna stampa del Teatro Sant'Erasmo

PAOLO BONORA
UNIVERSITÀ DI BOLOGNA, ITALIA - PAOLO.BONORA@UNIBO.IT

ABSTRACT (ITALIANO)

Nell'ottobre 1953, a Milano, fu inaugurato il Teatro Sant'Erasmo, un piccolo teatro a pianta centrale. Progettato da Carlo De Carli e gestito dalla compagnia fondata da Carlo Lari e Lida Ferro, tra il 1953 e il 1957 vi furono messe in scena 31 opere, in larga parte prime assolute, per un totale di quasi 1.500 repliche. Nonostante la sua unicità, la vicenda del Teatro Sant'Erasmo è stata quasi completamente dimenticata. Per ricostruirne la storia, si è proceduto allo spoglio della rassegna stampa conservata nell'archivio Lari-Ferro, utilizzando un modello descrittivo basato sulle ontologie formali della famiglia CIDOC CRM e realizzando un'applicazione basata sulla piattaforma ResearchSpace. L'analisi dei risultati ha rivelato la significativa copertura stampa ricevuta negli anni di attività, evidenziando la rilevanza dell'esperienza nel panorama del teatro italiano degli anni Cinquanta. Allo stesso tempo si è potuto ricostruire il repertorio e la cronologia degli spettacoli, un passaggio essenziale per procedere allo studio della genesi della singolarità teatrale rappresentata dal Sant'Erasmo.
Parole chiave: annotazione semantica; Teatro Sant'Erasmo; CIDOC CRMSci; ResearchSpace.

ABSTRACT (ENGLISH)

In October 1953, the Teatro Sant'Erasmo, a small arena theatre, was inaugurated in Milan. Designed by Carlo De Carli and managed by the company founded by Carlo Lari and Lida Ferro, starting from 1953 to 1957, 31 works, most of which were world premieres, for almost 1,500 performances were staged. Despite its uniqueness in Italy, the story of the Teatro Sant'Erasmo has been almost forgotten. In order to reconstruct its history, a review of the press coverage preserved in the Lari-Ferro archive was conducted using a descriptive model based on the formal ontologies of the CIDOC CRM family and creating an application based on the ResearchSpace platform. The analysis revealed the significant press coverage obtained during its initial years of activity, highlighting the cultural importance of the experience in the Italian theatre scene of the 1950s. At the same time, establishing the repertoire and performance chronology was essential for analysing the emergence of the Sant'Erasmo Theatre as a singular theatrical phenomenon.
Paper Title for AIUCD2026. Theatre clippings. Semantic annotation of the press review of the Sant'Erasmo Theatre
Keywords: semantic annotation; San'Erasmo Theatre; CIDOC CRMSci; ResearchSpace.

1. INTRODUZIONE

Il Teatro Sant'Erasmo è stato un teatro milanese a pianta centrale costruito nel 1953 in un seminterrato di via dei Giardini 7. Fu il frutto della singolare convergenza di tre figure professionali intorno ad un'idea di teatro: Carlo Lari, il regista; Lida Ferro, la prima attrice; Carlo De Carli e Antonio Carminati, gli architetti . La scelta programmatica della pianta centrale, l'essere gestito da una compagnia stabile (la compagnia Lari-Ferro tra il 1953 e il 1957), la quasi esclusiva programmazione di opere contemporanee (in buona parte prime assolute), l'essere aperto tutto l'anno (nelle quattro stagioni di attività della compagnia verranno date oltre 1.500 rappresentazioni per oltre 180.000 presenze)[1] sono alcuni dei tratti che ne fanno un'assoluta singolarità nel panorama teatrale italiano.
La compagnia fu estromessa nel giugno del 1957 e, dopo alterne vicende, il teatro fu demolito nel 1969[2] per destinare il seminterrato alla più prosaica funzione di parcheggio .[3] Eppure, i suoi creatori sembrano

[1] Così rileva orgogliosamente Lida Ferro nelle sue note.
[2] Tra il 1957 e il 1968 si avvicendarono due gestioni: la Compagnia di Gianfranco De Bosio (1957-1959) e l'impresario Maner Lualdi (1959-1968).
[3] Manzella e Pozzi, in chiusura delle pagine dedicate al Sant'Erasmo (pp. 218-222), scrivono: «Di teatri a Milano ne sono stati chiusi tanti. La chiusura del Sant'Erasmo riteniamo sia stata una delle più notate, poiché privava la città di

aver voluto lasciare un'ampia documentazione di quell'avventura, evidentemente convinti della sua rilevanza. A Bologna, presso la Fondazione Lyda Borelli, è conservato l'archivio Lari-Ferro[4] che raccoglie i lasciti documentari dei due artisti con un'ampia sezione dedicata a Sant'Erasmo; presso il Politecnico di Milano è invece conservato l'archivio De Carli.[5] L'archivio Lari-Ferro raccoglie documentazione relativa all'attività di Lari come regista, critico e traduttore. Consta di ritagli di stampa, relazioni, appunti di regia e, in particolare, di copioni e delle relative traduzioni. La peculiarità dello spazio teatrale a pianta centrale rendeva infatti necessario rivolgersi a testi originali o adattati. Ampia documentazione della ricezione dell'attività della compagnia è presente nei materiali raccolti dalla Ferro: rassegne stampa, foto, contratti e documenti amministrativi, copioni con appunti e annotazioni. Stranamente nella storiografia del teatro italiano del secondo Novecento se ne è quasi del tutto persa la memoria.[6] Per valutare l'effettiva rilevanza culturale di questa singolare iniziativa teatrale, ci si è rivolti alla rassegna stampa raccolta e conservata da Lida Ferro durante gli anni di attività della compagnia. L'obiettivo è studiare l'eco delle attività del teatro sulla stampa coeva, identificando, ove possibile, i nessi con il repertorio, gli spettacoli e gli interpreti al fine di ricavare un'indicazione, per quanto parziale, della risonanza del Sant'Erasmo sulla scena teatrale italiana degli anni Cinquanta. Operazione preliminare all'approfondimento dello studio multidisciplinare necessario a recuperare la memoria storica e ricostruire la genesi di una singolarità teatrale.



Figura 1 - Foto di scena di Nora Seconda (Il Dramma, n.202, 1° aprile 1954, p.7, foto: Giancolombo, Milano)[7]

2. I MATERIALI E GLI STRUMENTI

Nell'archivio Lari-Ferro sono conservati tre faldoni in cui Lida Ferro ha raccolto ritagli di giornali e periodici relativi alle stagioni 1953 e 1954, alla trasferta presso il Teatro Eliseo di Roma (stagione 1955-56) e alle "Serate Musicali" in cui vennero allestite presso il Sant'Erasmo due opere di Giancarlo Menotti.[8] Nel

una sala unica nel suo genere in Italia, e una delle pochissime tutt'ora esistenti in Europa. In compenso, al suo posto, Milano ha in più una rimessa privata per auto.»

[4] https://siusa-archivi.cultura.gov.it/cgi-bin/siusa/pagina.pl?TipoPag=comparc&Chiave=363348

[5] https://siusa-archivi.cultura.gov.it/cgi-bin/siusa/pagina.pl?TipoPag=comparc&Chiave=196485

[6] Fabrizio Cruciani, citando il Sant'Erasmo, rileva la singolarità della scelta programmatica della pianta centrale attribuendola al De Carli, senza però approfondirne gli esiti sul piano drammaturgico né della messa in scena (Cruciani & Ruzza, 2022).

[7] In particolare, dall'annotazione sul retro dell'originale si evince che la foto ritrae la prima rappresentazione del 12/2/1954 (si ringrazia il Centro Studi e l'Archivio del Teatro Stabile di Torino per l'accesso alla fonte).

[8] Presentate dal programma di sala come "due serate eccezionali", il 22 e 23 giugno del 1955, videro l'allestimento sulla "pista" del Sant'Erasmo di *La Medium* e *Il telefono*, ad opera della compagnia del Teatro Musicale di Villa Olmo

complesso si tratta di oltre quattrocento ritagli che l'attrice aveva scelto di conservare annotando data e testata. Pur essendo una fonte sicuramente parziale, frutto di una selezione, restituisce nel complesso una copertura stampa decisamente variegata non solo delle attività della compagnia, ma anche del teatro in quanto spazio e delle figure artistiche protagoniste di quell'avventura. Lo spoglio dei faldoni consente di avere una prima misura del fenomeno Sant'Erasmo e di sondare la ricezione degli allestimenti proposti in cartellone. Per rendere agevoli la consultazione e l'estrapolazione delle informazioni relative a ciascun ritaglio, si è innanzitutto proceduto alla digitalizzazione dei faldoni.
Per la consultazione delle copie digitali e l'annotazione dei singoli ritagli, è stata sviluppata un'applicazione basata sulla piattaforma ResearchSpace . In parallelo si è ricostruita la cronologia degli spettacoli, gli interpreti e il repertorio. Questa attività si è basata in larga parte su una bozza di resoconto delle attività della compagnia che la stessa Lida Ferro aveva redatto, ma non pubblicato, e sullo spoglio degli altri materiali presenti in archivio. L'applicazione per l'annotazione è stata organizzata in quattro sezioni: l'area relativa ai documenti, dove è possibile consultare i faldoni e procedere con l'annotazione; il repertorio, con l'elenco delle opere messe in scena; la cronologia degli spettacoli; la sezione relativa alle annotazioni inserite, dove è possibile visualizzare e ricercare i singoli ritagli e le entità annotate.

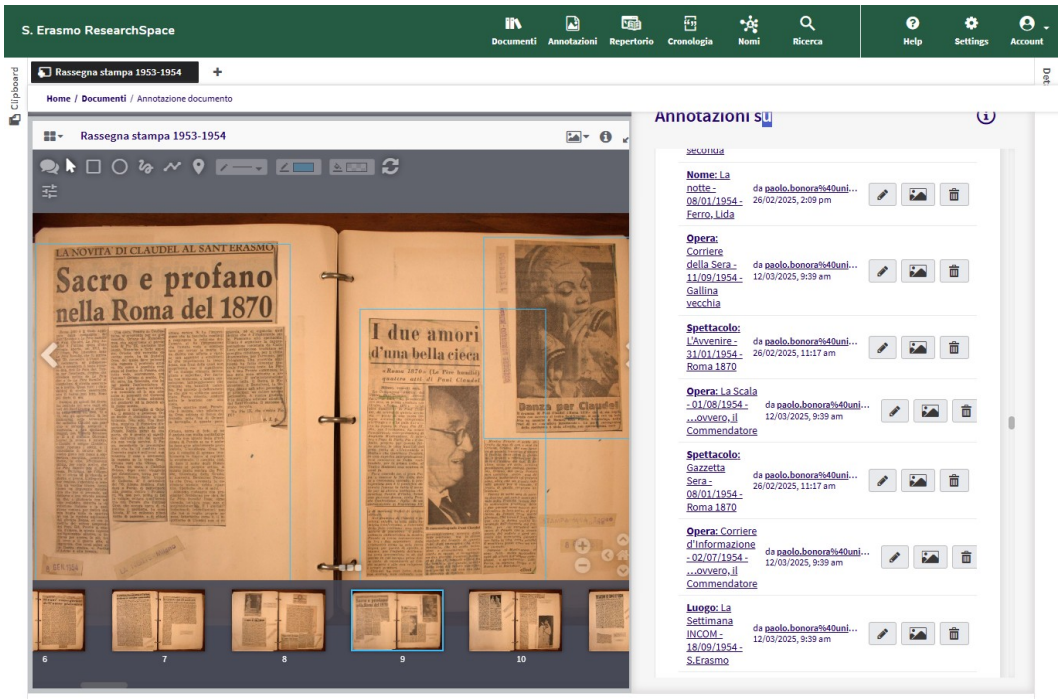


Figura 2 - Interfaccia utente utilizzata per l'annotazione

3. METODOLOGIA DI ANNOTAZIONE

Il primo obiettivo dell'annotazione è stato il censimento delle testate per ottenere una prima indicazione quantitativa del fenomeno. Per ciascun ritaglio, individuato mediante la marcatura della corrispondente area dell'immagine, sono stati annotati nome della testata e data di pubblicazione. In un secondo passaggio si è provveduto ad annotare i riferimenti al repertorio, al relativo allestimento, agli interpreti, alla sede o località in cui si è tenuto lo spettacolo, oltre ad eventuali note in forma libera da sottoporre a successiva analisi.
Le annotazioni sono state raccolte impiegando un modello descrittivo basato sulle ontologie di riferimento che la piattaforma ResearchSpace utilizza nativamente per la rappresentazione delle osservazioni scientifiche basate sulla documentazione caricata nel sistema. In particolare, CRMSci e CRMDig (Doerr et al., 2016), oltre alla ResearchSpace Execution Ontology (EX) dedicata alla gestione delle risorse digitali all'interno della piattaforma .
A queste tre ontologie di base sono state integrate le proprietà delle ontologie CIDOC Conceptual Reference Model e IFLA FRBRoo per rappresentare i rinvii tra le singole annotazioni e le entità associate:

diretta da Giulio Paternieri.

le opere in repertorio (rappresentate come frbroo:F15_ComplexWork), gli spettacoli (frbroo:F31_Performance), i luoghi (frbroo:F9_Place) e le persone (crm:E39_Actor).
La scelta di mantenere il set di ontologie che la piattaforma utilizza nativamente per consentire l'annotazione delle risorse deriva dalla necessità di limitare l'impegno necessario ad una customizzazione degli strumenti già disponibili stante la perfetta aderenza alle esigenze del progetto.
In particolare il modello risultante è articolato in quattro aree: una dedicata alla rappresentazione della risorsa digitale, una al suo valore documentale, una dedicata alla rappresentazione dell'osservazione con la relativa provenienza e, infine, quella relativa all'entità citata.
Quest'ultima è articolata nei grafi corrispondenti alle diverse tipologie di entità. Il diagramma (Figura 3) illustra la struttura del modello in forma sintetica. In una prima fase sono stati popolati i grafi delle entità del dominio oggetto delle annotazioni (le opere, gli spettacoli e i nomi) oltre ai metadati descrittivi dei documenti da annotare.
Ciascun faldone è rappresentato a livello fisico da un'istanza della classe crm:E18_Physical_Thing e, nel suo ruolo di documento, come crm:E31_Document.
Le riproduzioni delle singole pagine sono definite come istanze della classe rs:EX_Digital_Image, a sua volta associata al corrispondente file (rs:EX_File). Per mantenere la corretta sequenza delle pagine anche in fase di visualizzazione, è stata utilizzata la proprietà co:index per specificarne l'ordinale.
In fase di annotazione vengono create sia un'istanza di EX_Digital_Image_Region, per specificare la porzione dell'immagine oggetto dell'annotazione, sia un'istanza di crmsci:S4_Observation che rappresenta il contenuto dell'annotazione.
Quest'ultima è associata all'entità individuata attraverso la proprietà crm:P129_is_about che a sua volta la collega all'istanza crm:E31_Document mediante la proprietà crm:P70i_is_documented_in.
La *provenance* di ogni annotazione è specificata attraverso le proprietà crm:P14_carried_out_by, che traccia l'autore dell'annotazione (sotto forma di identità dell'utente), e crm:E52_Time_Span che specifica la data (in forma di *timestamp*).

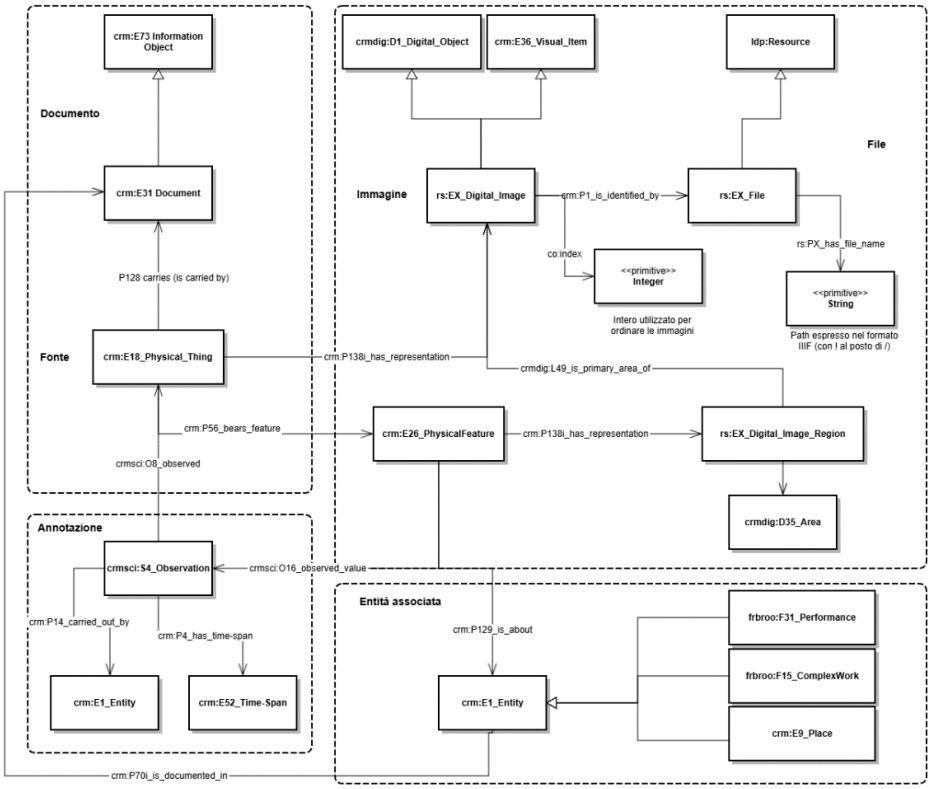


Figura 3: Diagramma del modello di annotazione

La base di conoscenza è suddivisa in tre *named graphs* distinti: il primo è relativo ai documenti, il secondo descrive le entità del dominio (opere, spettacoli, nomi, etc.), il terzo raccoglie le annotazioni. La suddivisione è progettata per mantenere distinte la componente dinamica (le annotazioni) e la componente statica (documenti e entità) durante il processo di annotazione. Una volta consolidata l'informazione acquisita, le nuove triple potranno essere facilmente migrate nel grafo principale. Nel suo complesso, il modello risultante sfrutta l'espressività delle ontologie sottostanti per esprimere l'informazione estratta durante il processo, generando una struttura semplice ma al tempo stesso efficace.[9]

Per lo sviluppo dell'applicazione utente si è cercato di utilizzare i componenti visuali e le funzionalità già previste dalla piattaforma. Questo ha consentito di sveltire notevolmente la fase di sviluppo, di sfruttare l'integrazione nativa con il framework IIIF (International Image Interoperability Framework) e le funzionalità di annotazione del visualizzatore di immagini Mirador. Il risultato è un'interfaccia utente semplice ed efficace che ha consentito di procedere all'annotazione in modo efficiente. L'applicazione è composta dalle maschere dedicate alla consultazione delle diverse entità (documenti, repertorio, spettacoli, nomi e annotazioni), dalla maschera di annotazione vera e propria e dalle funzionalità di ricerca. In questa fase iniziale pensate per accedere rapidamente ai documenti digitalizzati e alle annotazioni già inserite al fine di verificare lo stato di avanzamento del processo di annotazione e la sua correttezza.

Dal punto di vista operativo, l'utente, dopo aver effettuato il login, accede al menu principale che presenta le diverse funzionalità. Aprendo la maschera relativa ai documenti, può selezionare l'album che vuole annotare e individuare la pagina su cui operare. Selezionando una pagina, si apre la maschera di annotazione che consente di disegnare l'area dell'immagine che riproduce l'articolo e di inserire le relative annotazioni. La maschera è suddivisa in due parti. Sulla sinistra, il visualizzatore Mirador consente di navigare tra le pagine, selezionare quella di interesse e individuare le aree dell'immagine che verranno annotate. Sulla destra sono elencate le aree già individuate, ma ancora prive di annotazione, oltre alle

[9] Al termine delle attività il *named graph* delle annotazioni (<http://ficlit.unibo.it/annotations/>) conta 11424 triple, il grafo delle entità (<http://ficlit.unibo.it/serasmo/>) 3852, il grafo dei documenti (<http://ficlit.unibo.it/resource/serasmo>) 32.

annotazioni già inserite per lo stesso documento. Nel caso specifico della rassegna stampa, ciascuna area individua un ritaglio, di norma corrispondente ad un singolo articolo. Per questo, si è scelto di denominare l'area con il nome della testata e la data. Ciò ha consentito di censire le testate e di ottenere la cronologia della copertura stampa. A valle del censimento, si è proceduto annotando le entità citate nell'articolo: lo spettacolo, la pièce e i nomi. Dove rilevante è stato annotato anche il luogo o il teatro, oltre ad eventuali appunti in forma di testo libero. Il risultato è stata la correlazione tra la rassegna stampa e le tre dimensioni fondamentali per lo studio della storia del teatro: la cronologia degli spettacoli, il repertorio e l'indice degli artisti e delle maestranze.
Durante tutta l'attività di annotazione, l'utente può verificare l'avanzamento accedendo alla maschera con l'elenco delle annotazioni (proposte come *card* del corrispondente ritaglio), oppure consultando le maschere relative a repertorio, spettacoli e nomi per ciascuno dei quali vengono riportati gli articoli cui sono associati. La piattaforma ResearchSpace consente di profilare ed autenticare gli utenti. Questo consente di tracciare la provenienza delle annotazioni e il controllo incrociato (a volte critico) degli interventi fatti dai diversi utenti.
Le funzionalità di ricerca implementate consentono di ricercare le annotazioni sia in base a un criterio testuale, sia attraverso interrogazioni ottenute componendo criteri diversi: il periodico, l'opera o lo spettacolo a cui l'articolo si riferisce, oppure i nomi citati o il tipo di annotazione. Infine, sia le funzionalità di consultazione che di ricerca sono state pensate per supportare gli utenti durante le fasi iniziali dell'analisi dei risultati come conteggi e statistiche.

4. RISULTATI

Dallo spoglio dei tre faldoni presenti in archivio, sono stati individuati 444 articoli relativi a 146 testate, di cui 140 italiane e 6 straniere. Sono presenti quotidiani nazionali come Corriere della Sera, La Stampa, L'Unità, Il Messaggero, Il Giorno, Il Tempo; testate locali: Corriere Lombardo, Gazzetta del Mezzogiorno, Il Resto del Carlino, La Nazione, Il Piccolo, L'Adige, La Provincia (diverse edizioni); riviste specializzate o d'attualità: Dramma, Sipario, Epoca, L'Espresso, Bella, Settimana INCOM. Come copertura, tra i quotidiani spiccano il Corriere Lombardo con 22 articoli, seguito dal Corriere della Sera (11) e dal Tempo di Milano (10). Tra le riviste di settore: Il Dramma (8), Teatro Scenario (6) e Sipario (6). La distribuzione temporale è la seguente: 1953, 0,30%; 1954, 59,7%; 1955, 34,6%; 1956, 6,8%.
In totale sono state prodotte 653 annotazioni, di cui 210 relative a persone, 189 a spettacoli, 159 ad opere e 95 a luoghi. Tra i nomi più ricorrenti spiccano Lida Ferro (120 citazioni) e Carlo Lari (94), tra gli attori: Gastone Bartolucci (44), Renata Seripa (43) e Lucio Rama (38). Tra gli autori: Cesare Giulio Viola (28) e Paul Clodel (8). Gli spettacoli più citati sono: *Nora Seconda* con un totale di 57 articoli, le due riprese di *Roma 1870* (33) e l'allestimento di *Gallina Vecchia* (18). Caso a parte sono le due "Serate Musicali" coperte da 40 articoli. Gli articoli di critica dei testi vedono *Nora Seconda* spiccare (63), seguita da *…ovvero il Commendatore* (25) e dal repertorio dialettale. Gli aspetti legati allo spazio teatrale e alla messa in scena sono oggetto di 60 articoli relativi al Teatro Sant'Erasmo e agli allestimenti della "pista" presso il Ridotto dell'Eliseo di Roma (26) e nel foyer Rossini al Teatro Comunale di Bologna (9).
Al completamento dell'attività di annotazione, l'authority file dei nomi conta 108 entrate; 31 sono le opere censite e 48 gli spettacoli[10].

5. CONCLUSIONI

L'analisi della rassegna stampa ha consentito di ricostruire la produzione della compagnia Lari-Ferro tra il 1953 e il 1957. L'utilizzo di una piattaforma semantica in grado di gestire una base di conoscenza che combina la base informativa iniziale (repertorio, cronologia e nomi) con i dati derivati dalla documentazione ha reso l'attività di annotazione più efficiente e strutturata. Ciò ha permesso di consolidare i dati già acquisiti alla luce delle nuove informazioni raccolte con l'annotazione dei documenti. Il modello semantico, basato sulla famiglia di ontologie CIDOC CRM, ha consentito di rappresentare efficacemente sia le entità del dominio che le relazioni tra fonti ed entità osservate. La piattaforma ResearchSpace si è rivelata uno strumento efficace per la realizzazione dell'applicazione dedicata al caso

[10]Il riscontro con la rassegna stampa ha permesso di assestare le forme dei nomi e di verificare le date degli spettacoli, correggendole e integrandole dove mancanti. Una sintesi dei dati prodotti è accessibile all'indirizzo: https://rs.ficlit.unibo.it/resource/?uri=http://www.researchspace.org/resource/ErasmoNarrative

di studio, consentendo un'annotazione partecipativa dei documenti, la gestione della *provenance* e una prima fase di analisi delle informazioni ricavate dalla rassegna stampa.
La base di conoscenza ottenuta costituirà il punto di partenza per le successive attività di spoglio dei materiali presenti nel fondo. Si inizierà con lo studio delle annotazioni manoscritte presenti sui copioni conservati in archivio, come attività propedeutica all'indagine sul rapporto tra il repertorio e la messa in scena in pianta centrale. D'altra parte, gli elementi raccolti finora evidenziano chiaramente come la sola "novità" della scena centrale non avrebbe giustificato la visibilità ottenuta sulla stampa coeva in assenza di una precisa idea di teatro in cui lo spazio diventa il fulcro tra repertorio e regia, attori e pubblico.[11] [12]
Resta quindi da capire perché si sia persa traccia di questa singolare esperienza teatrale, una vicenda ancora tutta da indagare.

BIBLIOGRAFIA

Cruciani, F., & Ruzza, L. (2022). *Lo spazio del teatro: Con tracce grafiche di Luca Ruzza* (26 ed). Laterza.

De Carli, C. (1982). *Architettura, spazio primario*. Hoepli editore.

Doerr, M., Bekiari, C., LeBoeuf, P., & nationale de France, B. (2008). FRBRoo, a conceptual model for performing arts. *2008 Annual Conference of CIDOC, Athens*, 15–18.

Doerr, M., Kritsotaki, A., Rousakis, Y., Hiebel, G., & Theodoridou, M. (2014). *CRMsci: The scientific observation model*. FORTH, Tech. Rep., 2014.[Online]. Available: https://projects. ics. forth ….

Manzella, D., & Pozzi, E. (1971). *I teatri di Milano*. U. Mursia & C.

Michon, P., Ore, C.-E., Stead, S., & Velios, A. (2025). *Reference Model*. https://cidoc-crm.org/sites/default/files/cidoc_crm_version_7.3.1_1.pdf

Oldman, D., & Tanase, D. (2018). Reshaping the knowledge graph by connecting researchers, data and practices in ResearchSpace. *International Semantic Web Conference*, 325–340.

Riva, P., Le Boeuf, P., Žumer, M., & others. (2020). *IFLA Library Reference Model: Un modello concettuale per le informazioni bibliografiche*.

Snydman, S., Sanderson, R., & Cramer, T. (2015). The International Image Interoperability Framework (IIIF): A community & technology approach for web-based images. *Archiving conference*, *12*, 16–21.

The ResearchSpace System—Underlying Principles. (2023). Kartography CIC. https://kartography.org/researchspace.html

[11] Che l'iniziativa partisse col favore della stampa professionale è attestato dal numero di novembre 1953 della rivista Il Dramma che riproduce in copertina il logo del Sant'Erasmo, disegnato da Enrico Ciuti, accompagnato, nell'editoriale di apertura del numero, dagli auspici del direttore Lucio Ridenti.
[12] Per inquadrare il tema della scena centrale nel contesto teatrale novecentesco si veda (Cruciani, 1992).

Annotazione semantica delle formule poetiche del Medioevo germanico

ROBERTO ROSSELLI DEL TURCO,[1] LETIZIA VEZZOSI[2]
[1] UNIVERSITÀ DI TORINO, ITALIA – ROBERTO.ROSSELLIDELTURCO@UNITO.IT
[2] UNIVERSITÀ DI FIRENZE, ITALIA – LETIZIA.VEZZOSI@UNIFI.IT

ABSTRACT (ITALIANO)

Gli studi pionieristici di M. Parry e A. Lord sulla composizione e la trasmissione orale di testi poetici hanno introdotto i concetti di "formula" e "tema", divenuti in seguito centrali nello studio delle tradizioni orali. La teoria di composizione orale-formulare è stata successivamente applicata alle letterature germaniche del Medioevo, con particolare attenzione allo studio della letteratura in inglese antico a partire dal noto articolo di F. Magoun (Magoun, 1953), per essere poi estesa ad altre tradizioni germaniche del Medioevo. La vasta produzione critica che ne è scaturita ha messo in luce il ruolo fondamentale delle formule, soprattutto quando combinate con altri stilemi quali *kenningar* e composti immaginifici, all'interno di un processo di composizione orale altamente sofisticato e flessibile. Con l'introduzione di metodi computazionali applicati allo studio dei testi letterari, sono stati sviluppati vari strumenti per l'analisi automatica dei componimenti poetici, in modo da individuare le formule ed elaborarle in termini statistici. Per quanto tali approcci siano utili a fini di ricerca sul piano quantitativo, si ritiene tuttavia necessario affiancare loro una metodologia di annotazione semantica, capace di cogliere non soltanto gli aspetti distributivi delle formule, ma anche la loro struttura interna, le modalità di variazione e adattamento nel tempo, nonché eventuali caratteristiche riconducibili a singoli autori. Questo paper propone un metodo di annotazione basato sugli schemi XML/TEI tale da permettere non solo l'elaborazione del markup relativo alle formule a fini di visualizzazione, ma anche di utilizzarlo come base per interrogazioni complesse all'interno del software EVT, in modo da offrire uno strumento di ricerca integrato ed efficace allo studioso.
Parole chiave: letterature germaniche medievali; composizione orale-formulare; inglese antico; annotazione semantica; markup XML/TEI.

ABSTRACT (ENGLISH)

Semantic Annotation of Poetic Formulas in the Germanic Middle Ages. The pioneering studies of M. Parry and A. Lord on the oral composition and transmission of poetic texts introduced the concepts of *formula* and *theme*, which subsequently became central to the study of oral traditions. The theory of oral-formulaic composition was later applied to the medieval Germanic literatures, with particular attention to Old English literature following the influential article by F. Magoun (Magoun, 1953), and was subsequently extended to other medieval Germanic traditions. The extensive body of scholarship that emerged from this line of research has highlighted the fundamental role of formulas, especially when combined with other stylistic devices such as *kennings* and imaginative compounds, within a highly sophisticated and flexible process of oral composition. With the introduction of computational methods in the study of literary texts, a variety of tools have been developed for the automatic analysis of poetic compositions, aimed at identifying formulas and processing them statistically. While such approaches are undoubtedly valuable for quantitative research purposes, it is nonetheless necessary to complement them with a methodology of semantic annotation capable of capturing not only the distributive aspects of formulas, but also their internal structure, their modes of variation and adaptation over time, and any features attributable to individual authors. This paper proposes an annotation method based on XML/TEI schemas that allows not only for the visualization of formula-related markup, but also for its use as a foundation for complex queries within the EVT software, thereby providing scholars with an integrated and effective research tool.
Keywords: Medieval Germanic literatures; oral-formulaic theory; Old English; semantic annotation; XML/TEI markup.

1. INTRODUZIONE

Il fondamento teorico della formula poetica risale all'opera pionieristica di Milman Parry (Parry, 1932), i cui studi sull'epica greca arcaica, in particolare sull'opera omerica, hanno rivoluzionato la comprensione delle espressioni ripetute nel verso orale. Parry sostiene che espressioni quali ῥοδοδάκτυλος Ἠώς ("l'Aurora dalle rosee dita") in apertura di verso, o πόδας ὠκὺς Ἀχιλλεύς ("Achille piè veloce") in posizione finale, non costituiscono ornamenti stilistici, bensì elementi strutturali del sistema versificatorio. Tali elementi funzionano come moduli linguistici prefabbricati, caratterizzati da coerenza nella condizione metrica e da stabilità semantica.
La celebre definizione di formula poetica data da Parry ("a group of words which is regularly employed under the same metrical conditions to express a given essential idea" Parry [1932] 1971: 272) riconosce le formule come unità compositive sistemiche, attinte da un repertorio trasmesso tradizionalmente. Questo repertorio consente la composizione in tempo reale durante la performance orale e funziona come un insieme di unità lessico-grammaticali prefabbricate, selezionate in base a vincoli metrici (come la quantità sillabica e gli schemi accentuali), all'adeguatezza semantica nel contesto narrativo e alla possibilità di alternanza paradigmatica, ossia la sostituzione di una formula con un'altra di valore metrico e semantico equivalente. In altri termini, le formule sono il prodotto di un sistema compositivo parallelo al linguaggio naturale, caratterizzato da regolarità proprie e sovradeterminato da esigenze prosodiche, cognitive e tradizionali. L'epica orale deve dunque essere concepita come un registro grammaticale specializzato, nel quale le formule agiscono come elementi sintagmatici che interagiscono fra loro per formare strutture superiori e rappresentano le unità costruttive dell'enunciato poetico.
In questa linea si collocano anche le osservazioni di Albert Lord sulla composizione orale dei cantori epici slavi meridionali (Lord, 1960), che mostrano come la performance orale vada ben oltre la semplice memorizzazione. Le formule si inseriscono in modelli narrativi più ampi, detti "temi", che forniscono l'impalcatura per una creazione narrativa spontanea e coerente. Come afferma Lord, le formule non si limitano a riempire spazi metrici, ma assolvono simultaneamente una funzione semantica e narrativa, operando sia per il poeta sia per il pubblico nella costruzione della comunicazione poetica. Secondo Lord, una formula è significativa non solo per ciò che denota, ma anche per il modo e il contesto in cui è stata utilizzata in passato: il pubblico riconosce non solo le parole, ma anche il ruolo narrativo più ampio e la memoria culturale che esse evocano.
In altre parole, le formule non sono semplici sequenze di parole metricamente regolari, ma costruzioni discorsivamente funzionali, dotate di una funzione illocutoria (ossia capaci di svolgere ruoli narrativi e pragmatici specifici) e di una funzione strutturale. Esse si combinano per formare temi e strutture schematiche di ordine superiore che guidano la composizione e l'interpretazione. Una formula, ad esempio, può introdurre il discorso diretto, catturare l'attenzione del pubblico o segnalare una svolta narrativa; a un livello più alto, le formule contribuiscono alla macro-sintassi poetica, combinandosi in unità schematiche ricorrenti, come il vanto dell'eroe, il lamento della vedova o il motivo del viaggio, che fungono da strumenti narrativi e da cornici cognitive e culturali per l'organizzazione del racconto.

2. IL SISTEMA FORMULARE GERMANICO

Fin dai primi tentativi di applicazione della teoria di Parry e Lord alla poesia germanica antica (Magoun, 1953), sono emerse difficoltà legate a differenze metriche fondamentali. Come osserva Niles (1983), il ricorso all'analogia greca si è rivelato al tempo stesso una risorsa e un limite. Se da un lato ha fornito un quadro teorico iniziale, dall'altro il contrasto tra l'esametro quantitativo greco, fisso e rigidamente strutturato, e il metro germanico, allitterativo, accentuale e flessibile, ha reso necessaria una ridefinizione del concetto di formula. Magoun ha dimostrato che il *Beowulf* era costruito attraverso una dizione formulare, ma anche che il verso in inglese antico presentava una maggiore flessibilità, probabilmente dovuta ai diversi vincoli imposti dal metro allitterativo, in quanto lessemi differenti potevano sostituire parole chiave preservando lo stesso significato e la medesima struttura metrica, attraverso un principio di equivalenza semantica.
Da qui è nata la consapevolezza dell'esistenza di un sistema formulare specificamente germanico, caratterizzato da dipendenza dalla tradizione, adattabilità metrica e regolarità compositiva all'interno di convenzioni poetiche culturalmente determinate. Gli studiosi hanno progressivamente spostato l'attenzione dalle stringhe lessicali fisse al concetto di "formula in the story" introdotto da Lord, ossia a modelli sintattici ricorrenti. Le formule sono state così ridefinite come cornici sintattiche o come sistemi

formulari: strutture grammaticali coerenti, capaci di generare variazione. Questa rianalisi strutturale ha ampliato la teoria orale-formulare oltre il solo ambito metrico, includendo grammatica, semantica e funzione culturale, rafforzando l'idea che le formule non siano residui verbali rigidi, ma schemi produttivi operanti all'interno di grammatiche poetiche specifiche:

> with oral poetry we are dealing with a particular and distinctive process in which oral learning, oral composition and oral transmission almost merge; they seem to be different facets of the same process (Lord, 1960: 5)

Riconoscendo la natura peculiare della poesia germanica, studiosi come Donald K. Fry (1967) e John D. Niles (1983) hanno riformulato il concetto di formula come complesso ritmico-sintattico-semantico. Le formule non sono quindi frasi isolate e rigide, ma appartengono a sistemi formulari, ossia insiemi di espressioni strutturalmente affini, caratterizzate da schemi grammaticali coerenti, come il modello [NOME] + [VERBO] + [OGGETTO]. Tali schemi facilitano la memoria e la composizione, consentendo variazione estetica e arricchendo il tessuto poetico, evitando al contempo la ridondanza. La variazione poetica è dunque motivata da un impulso estetico a evitare la ripetizione meccanica grazie a quella che Russom ha definito "artful avoidance of the useful phrase" (Russom, 1978).
Le formule non sono quindi unità statiche, ma modelli produttivi. Fry arrivò persino a suggerire che "Germanic formulas preceded and ultimately produced Germanic meter" (Fry, 1974: 256), implicando che non sia il metro a generare le formule, bensì il contrario. Un'illustrazione particolarmente chiara di questo concetto si trova nell'analisi dell'elemento finale *-dagum* ("giorni") nei semiversi dell'inglese antico (Niles, 1983). Tale elemento funge da ancoraggio per uno schema formulare astratto, che può essere riempito con diversi modificatori allitteranti, mantenendo invariata la struttura metrica e ritmica.

Esempio in inglese antico	**Traduzione**	**Ruolo sintattico/semantico**
in geardagum	"nei giorni antichi"	Tempo (passato)
næfre he on aldordagum	"mai egli nei giorni della vita"	Tempo (arco vitale)
æfter deaðdæge	"dopo il giorno della morte"	Tempo (evento)
on ðyssum windagum	"in questi giorni di fatica"	Tempo (presente/difficoltà)

La formula non coincide con l'intero verso, ma con lo schema X + DAGUM/DÆGE (dove X è un elemento variabile e allitterante) che soddisfa i requisiti dell'accentuazione metrica ed esprime invariabilmente un concetto di tempo, durata della vita o epoca storica.
In questo senso, il materiale germanico ha costituito un terreno empirico privilegiato per sviluppare le intuizioni di Parry: ciò che nell'epica omerica appariva come espressione metricamente fissa emerge qui come modello di produttività a livello di grammatica delle costruzioni. Ne è derivato uno spostamento teorico dalla formula intesa come sequenza ripetuta alla formula come infrastruttura compositiva, considerando le formule non come semplici frasi ereditate, ma come costruzioni convenzionali, immagazzinate, attivate e variate all'interno di repertori culturalmente situati. Questa impostazione è stata ulteriormente formalizzata attraverso i concetti di "tradition dependency" (Foley, 1985: 68) e di "traditional referentiality" (Foley, 1991; Foley, 2002), applicati alle formule poetiche, le quali veicolano una risonanza culturale e richiamano esecuzioni precedenti anche quando risultano lievemente modificate.
Secondo Diamond, infatti, la formula in inglese antico "seems to be the equivalent in Old English poetry of the phoneme in linguistics, the basic unit associated with meaning" (Diamond, 1975: 304). Tale prospettiva costituisce il fondamento per gli sviluppi metodologici successivi, inclusi gli approcci basati su corpora, l'annotazione formale e la modellizzazione computazionale dei sistemi formulari.

3. APPROCCI COMPUTAZIONALI ALLO STUDIO DELLE FORMULE POETICHE

L'applicazione di metodi computazionali allo studio delle formule poetiche ha aperto nuove prospettive nella ricerca filologica e linguistica (tra i precursori si veda Stevick, 1962). Se in passato le espressioni formulari venivano analizzate prevalentemente attraverso il *close reading* e il confronto manuale dei testi, oggi è possibile studiarle in modo sistematico su corpora di grandi dimensioni, con un ampliamento significativo della scala di osservazione e un aumento della precisione analitica. Gli strumenti computazionali consentono di individuare ricorrenze lessicali, sintattiche e metriche attraverso più testi, spesso a un livello di dettaglio difficilmente raggiungibile con l'analisi tradizionale. Tali metodi permettono di tracciare la frequenza e la distribuzione delle formule, di misurarne la variazione interna e di valutare

la loro produttività e stabilità all'interno di una determinata tradizione poetica. Mappando le formule su strutture sintattiche astratte (ad esempio [NOME] + [VERBO] + [OGGETTO]), diventa possibile testare empiricamente ipotesi relative alla loro regolarità grammaticale e alla loro flessibilità compositiva. Numerosi studi (Frog, 2014; Pagán Cánovas e Antović, 2016) hanno dimostrato l'efficacia di questi approcci. Più recentemente R. Ginevra e altri ricercatori (Ginevra *et al*., 2024) hanno applicato i principi della grammatica delle costruzioni alle formule poetiche dell'antico inglese, utilizzando lo *York–Helsinki Parsed Corpus of Old English Poetry* per analizzare il modello di introduzione del discorso basato su *maþelode*. I risultati mostrano una significativa variazione interna al sistema formulare, confermando l'interpretazione delle formule come costruzioni schematiche e produttive, piuttosto che come espressioni rigidamente fissate.
Nel loro insieme, gli studi computazionali rafforzano una concezione linguistica delle formule poetiche come costruzioni flessibili e schematiche, anziché come sequenze verbali immutabili. Tali metodologie consentono una validazione quantitativa di affermazioni stilistiche e linguistiche, favoriscono confronti intertestuali e diacronici nell'uso delle formule e permettono un'integrazione sempre più stretta tra linguistica cognitiva, grammatica delle costruzioni e teoria della tradizione orale.

4. OBIETTIVI DELLA RICERCA

Al di là di alcuni limiti intrinseci dell'approccio puramente NLP, come la difficoltà di catturare la variabilità semantica delle formule e le funzioni discorsive non direttamente osservabili tramite pattern automatici, questo tipo di metodologia presuppone uno schema di lavoro e di analisi che non necessariamente risponde alle esigenze di tutti gli studiosi, né risulta adeguato a ogni possibile obiettivo di ricerca. Ciò appare particolarmente evidente negli studi dedicati alla poesia in inglese antico e, più in generale, alle altre lingue germaniche medievali. Le ricerche sul cosiddetto *"maþelode system"* (Creed, 1957: 525), ad esempio, basate sull'individuazione di schemi sintattici finalizzati a riconoscere le introduzioni formulari al discorso diretto, hanno mostrato come tale "sistema" non possa essere inteso come una formula da applicare rigidamente, ma debba piuttosto essere concepito come un continuum di costruzioni formulari. Se alcune espressioni risultano infatti fortemente cristallizzate (come nel tipo *Beowulf maþelode / bearn Ecgþeowes*), altre presentano un grado maggiore di flessibilità e ricorrono a sinonimi o a stilemi alternativi per evitare un'eccessiva ripetizione o schematicità; è il caso, ad esempio, di *Wulfgar maþelode / þæt wæs Wendla leod* (*Beowulf*, v. 345), dove il secondo semiverso rappresenta una significativa variazione della versione standard della formula con *maþelode*.
Un ulteriore limite dell'approccio esclusivamente computazionale risiede nel fatto che esso richiede competenze tecniche e risorse non sempre facilmente accessibili. Ne deriva che tale metodologia tende a rimanere circoscritta all'ambito della linguistica dei corpora, risultando meno fruibile per quei ricercatori che adottano approcci diversi. In una prospettiva più ampia, tra i soggetti indirettamente esclusi si possono annoverare anche gli utenti delle edizioni scientifiche digitali. Le *Digital Scholarly Editions* (DSE) più recenti, infatti, si orientano sempre più verso un modello di *knowledge base*, nel quale i dati testuali non sono concepiti soltanto per la visualizzazione, ma anche per forme avanzate di elaborazione che consentono interrogazioni complesse e la produzione di nuova conoscenza. In questo contesto, il nostro obiettivo è definire un modello di annotazione semantica per la marcatura di formule e temi nei testi poetici in inglese antico. La codifica in formato XML/TEI consente innanzitutto una maggiore aderenza alle tipologie individuate rispetto a ricerche generiche su corpora; inoltre, essa si integra pienamente nel processo di allestimento delle DSE, dal momento che gli schemi TEI rappresentano lo standard di riferimento per questo tipo di edizioni. Tale annotazione costituisce infine il presupposto necessario per successive elaborazioni dei dati. L'impiego di uno strumento ormai maturo come EVT (Rosselli Del Turco, 2019; Rosselli Del Turco, 2025), che prevede l'integrazione di funzionalità di *elaborazione di dati semi-strutturati* (EDSS) tra quelle offerte agli utenti (Rosselli Del Turco, 2021), rende possibile arricchire le capacità di ricerca di una DSE, consentendo interrogazioni mirate volte a individuare strutture sintattiche ricorrenti e variabili, schemi d'uso delle formule e, più in generale, le modalità di innovazione poetica.
Il punto di vista qui adottato non è dunque quello dello studioso che utilizza strumenti di tipo NLP per condurre analisi automatiche sui testi, bensì quello del lettore o dell'utente di un'edizione digitale che intende svolgere ricerche articolate direttamente all'interno dell'edizione stessa. Ciò non implica una svalutazione degli strumenti NLP, che possono anzi rivelarsi utili in fase di supporto alla codifica, ad esempio per l'individuazione preliminare di formule e temi. Tuttavia, è solo attraverso un approccio di *close reading* che diventa possibile analizzare in modo approfondito il passaggio da modelli formulari

relativamente stabili a schemi strutturati in cui la variazione assume un ruolo centrale: dalle ripetizioni meccaniche, spesso determinate dai vincoli metrici, a costruzioni semanticamente ricche e dinamiche, che si trasformano e si affinano nell'uso.

5. LA DEFINIZIONE DI UN MODELLO DI CODIFICA BASATO SUGLI SCHEMI TEI

Il sistema di marcatura proposto si fonda su una definizione di formula volutamente ampliata rispetto alla versione tradizionale (Parry, 1932) al fine di includere non solo la dimensione metrico-lessicale, ma anche le dimensioni costruttiviste, semantiche e paradigmatiche emerse negli studi successivi sulla poesia germanica medievale. L'obiettivo è identificare la funzione della formula in relazione al tema narrativo, alla sua conformità ai vincoli del verso lungo allitterante germanico, alla struttura sintattica e alle possibilità di variazione all'interno di uno stesso paradigma orale-formulare.

Il modello di codifica si articola in tre livelli distinti, ma interconnessi:

1. **livello testuale** (occorrenze), in cui le formule sono annotate direttamente nel testo senza imporre una classificazione semantica o lessicale rigida usando l'elemento <seg type="formula">;
2. **livello lessicale** (repertorio delle formule), in cui le formule vengono rappresentate come oggetti astratti mediante l'elemento <entry>, che permette di descriverne la forma canonica, i componenti e le eventuali varianti;
3. **livello concettuale** (pattern formulari), in cui le formule sono modellate come schemi astratti tramite <item> e <seg type="formulaPattern">, integrati da strutture di *feature* (<fs>) che ne descrivono le proprietà semantiche e funzionali.

Il primo livello prevede la marcatura delle occorrenze formulari nel testo, come nel seguente esempio:

```
<l n="32">
 <seg ref="#witig_dryhten" type="formula" xml:id="DEOR_witig_dryhten">
  <w pos="ADJ">witig</w>
  <w pos="NOUN">dryhten</w>
 </seg> wendeþ geneahhe,</l>
```

L'uso di <seg> consente di mantenere un approccio descrittivo e non prescrittivo, evitando di interpretare a priori la formula come unità terminologica o concettuale. L'attributo @ref collega l'occorrenza testuale a una voce di glossario, rappresentata mediante <entry>:

```
<entry corresp="#adj_noun" xml:id="witig_dryhten">
 <form type="lemma">
  <orth>witig dryhten</orth>
 </form>
 <form type="components">
  <form>
   <orth><w lemmaRef="#https://bosworthtoller.com/36291" xml:id="witig">witig</w></orth>
   <gramGrp><pos>ADJ</pos></gramGrp>
  </form>
  <form>
   <orth><w>dryhten</w></orth>
   <gramGrp><pos>NOUN</pos></gramGrp>
  </form>
 </form>
```

Ogni <entry>, infine, è collegato a un più ampio schema concettuale che permette di gestire le formule sul piano semantico, in quanto costruzioni versatili collegate a un tema, e non solo metrico-sintattico:

```
<item xml:id="eval_phrase">
 <label>EVALUATION PHRASE</label>
```

```
<seg type="formulaPattern" subtype="phrase" function="evaluation" met="halfline">
  <term type="pos" ref="https://universaldependencies.org/u/pos/PRON.html">PRON</term>
  <term type="pos" ref="https://universaldependencies.org/u/pos/VERB.html">VERB</term>
  <term type="pos" ref="https://universaldependencies.org/u/pos/ADJ.html">ADJ</term>
  <term type="pos" ref="https://universaldependencies.org/u/pos/NOUN.html">NOUN</term>
</seg>
<fs>
  <f name="semantic_field">
    <vColl>
      <string>appraisal</string>
      <string>eulogy</string>
    </vColl>
  </f>
  <f name="theme"><string>good vs. bad rulers</string></f>
</fs>
</item>
```

Nello stato attuale, il markup proposto presenta un carattere ancora parzialmente sperimentale, in particolare per quanto riguarda la definizione del quadro concettuale di riferimento. Gli schemi TEI, infatti, non sono stati concepiti per la modellazione sistematica di categorie concettuali astratte, pur offrendo alcuni strumenti in tal senso, peraltro piuttosto limitati. Per questo motivo, il livello concettuale viene qui implementato in forma deliberatamente flessibile.
Le proprietà semantiche (ad esempio semantic_field e theme) non sono tuttavia intese come categorie libere, ma come valori appartenenti a un vocabolario controllato definito a livello di progetto. Tale vocabolario può essere implementato mediante strumenti nativi della TEI, come le tassonomie (elemento <taxonomy>), garantendo coerenza annotativa e limitando il rischio di proliferazione arbitraria delle categorie.
L'uso delle *feature structures* consente inoltre di rappresentare la dimensione paradigmatica delle formule, ad esempio specificando quali elementi di uno schema siano soggetti a variazione. In questo modo il modello non si limita a registrare occorrenze, ma permette di descrivere il comportamento sistemico delle formule all'interno della tradizione poetica rendendo esplicite relazioni che rimarrebbero implicite in un'analisi puramente distribuzionale.
In una fase successiva, una maggiore formalizzazione del livello concettuale potrà essere perseguita mediante l'integrazione con ontologie, già supportate, seppur in forma sperimentale, all'interno della TEI (si veda, ad esempio, la *issue* Encoding RDF relationships in TEI), al fine di rappresentare in modo più articolato le relazioni semantiche individuate.

6. CONCLUSIONI

Il modello di annotazione proposto introduce un livello semantico e interpretativo che integra, senza sostituirli, gli approcci automatici basati su tecniche di tipo NLP, permettendo di rappresentare in modo esplicito la struttura interna delle formule, la loro funzione discorsiva e le modalità di variazione. In questo senso, la codifica TEI costituisce un complemento agli strumenti computazionali, rendendo possibile un'interrogazione più articolata dei dati all'interno di ambienti come EVT. L'obiettivo finale, infatti, è di permettere all'utente di una DSE di interrogare i dati testuali non solo in termini di ricorrenze formali, ma anche di strutture sintattiche, pattern semantici e schemi formulari ricorrenti, al fine di individuare regolarità, variazioni e relazioni tra le formule e, più in generale, di supportare la produzione di nuova conoscenza critica a partire dal testo annotato.
Un elemento centrale del modello è la distinzione tra livello testuale, lessicale e concettuale, che consente di descrivere le formule come oggetti complessi e stratificati. Questa architettura rende il sistema intrinsecamente scalabile, permettendo di estendere la stessa logica di annotazione a unità di ordine superiore, quali scene tipo, motifemi ed episodi narrativi. La definizione di categorie semantiche attraverso vocabolari controllati garantisce coerenza annotativa e costituisce un primo passo verso forme più avanzate di formalizzazione e interoperabilità.

Nel suo insieme, il modello proposto non rappresenta un sistema chiuso, ma un'infrastruttura aperta, orientata alla formalizzazione e all'interrogazione delle strutture formulari, e suscettibile di ulteriori sviluppi, in particolare per quanto riguarda la modellazione delle relazioni concettuali.

BIBLIOGRAFIA

Creed, R. P. (1957). The andswarode-System in Old English Poetry. Speculum, 32(3), 523–528. https://doi.org/10.2307/2849894.

Foley, J. M. (1985). Oral-formulaic theory and research: An introduction and annotated bibliography. Garland.

Foley, J. M. (1991). Immanent art: From structure to meaning in traditional oral epic. Indiana Univ. Press.

Foley, J. M. (2002). How to read an oral poem. University of Illinois Press.

Frog, M. (2014). Degrees of Well-Formedness: The Formula Principle in the Analysis of Oral-Poetic Meters. RMN Newsletter 8: 68-70.

Fry, D. K. (1967). Old English Formulas and Systems. English Studies, 48(1–6), 193–204. https://doi.org/10.1080/00138386708597274.

Fry, D. K. (1974). Cædmon as a Formulaic Poet. Forum for Modern Language Studies, X(3), 227–247. https://doi.org/10.1093/fmls/X.3.227.

Lord, A. B. (1960). The singer of tales. Harvard University Press.

Magoun, F. P. (1953). Oral-Formulaic Character of Anglo-Saxon Narrative Poetry. Speculum, 28(3), 446–467. https://doi.org/10.2307/2847021.

Diamond, R. E. (1975). The Diction of the Old English Christ. In L. E. Nicholson & D. W. Frese (a cura di), Anglo-Saxon poetry: Essays in appreciation. For John C. McGalliard (1 online resource (xvi, 387, pp. 301–311). University of Notre Dame Press. http://books.google.com/books?id=GAtaAAAAMAAJ.

Ginevra, R., Biagetti, E., Brigada Villa, L., & Giarda, M. (2024). A Constructionist and Corpus-Based Approach to Formulas in Old English Poetry. Languages, 9(7), Articolo 7. https://doi.org/10.3390/languages9070237.

Niles, J. D. (1983). Formula and Formulaic System. In Beowulf: The Poem and Its Tradition. Harvard University Press.

Pagán Cánovas, C., & Antović, M. (2016). Formulaic creativity: Oral poetics and cognitive grammar. Language & Communication, 47, 66–74. https://doi.org/10.1016/j.langcom.2015.12.001

Parry, M. (1932). Studies in the Epic Technique of Oral Verse-Making: II. The Homeric Language as the Language of an Oral Poetry. Harvard Studies in Classical Philology, 43, 1–50. https://doi.org/10.2307/310666.

Rosselli Del Turco, R. (2019). Designing an advanced software tool for Digital Scholarly Editions: The inception and development of EVT (Edition Visualization Technology). Textual Cultures, 12(2), 91–111. https://doi.org/10.14434/textual.v12i2.27690.

Rosselli Del Turco, R. (2021). Elaborazione di dati semi-strutturati: Ipotesi implementative e casi d'uso tratti da testi in inglese antico. Umanistica Digitale, (10), Articolo 10. https://doi.org/10.6092/issn.2532-8816/12598.

Rosselli Del Turco, R., Cucurnia, D., & Buzzoni, M. (2025). Embracing flexibility: New EVT features for critical editing, accessibility and inclusivity. In S. Rebora, M. Rospocher, & S. Bazzaco (a cura di), Diversità, Equità e Inclusione: Sfide e Opportunità per l'Informatica Umanistica nell'Era dell'Intelligenza Artificiale, Proceedings del XIV Convegno Annuale AIUCD2025. AIUCD. https://doi.org/10.6092/unibo/amsacta/8380.

Russom, G. R. (1978). Artful Avoidance of the Useful Phrase in «Beowulf», «The Battle of Maldon», and «Fates of the Apostles». Studies in Philology, 75(4), 371–390. http://www.jstor.org/stable/4173979.

Stevick, R. D. (1962). The Oral-Formulaic Analyses of Old English Verse. Speculum, 37(3), 382–389. https://doi.org/10.2307/2852359.

Ça y est : Annotating Arabic Texts for Teaching

ELISA GUGLIOTTA[1], MAURA TARQUINI[2]
[1] UGA, LIDILEM; UNIVERSITÉ TOULOUSE JEAN JAURÈS, LNPL – ELISA.GUGLIOTTA@UNIV-TLSE2.FR
[2] UNIVERSITÀ DEGLI STUDI DI SASSARI - MTARQUINI@UNISS.IT

ABSTRACT (ENGLISH)

This paper critically examines the use of digital tools for processing contemporary Arabic texts from a pedagogical perspective. Through a case study, it explores the challenges arising in the digitization, OCR, and linguistic annotation of texts characterized by graphic variation, register alternation, and code-switching. By combining generative OCR, automatic annotation following the Universal Dependencies framework, and manual validation in the INCEpTION environment, the study shows how errors and annotative ambiguities reveal the limits of existing standards and can be reframed as pedagogical resources. The paper thus argues for a critical approach to Arabic digital textuality as a space of mediation between linguistic complexity and computational practices.
Keywords: Arabic digital textuality; OCR; morphosyntactic parsing; Universal Dependencies; Arabic language teaching

ABSTRACT (ITALIANO)

Ça y est : Annotare testi arabi per la didattica. Il contributo analizza criticamente l'applicazione di strumenti digitali al trattamento di testi arabi contemporanei in una prospettiva orientata alla didattica. A partire da un caso di studio, vengono esaminate le principali criticità che emergono nelle fasi di digitalizzazione, OCR e annotazione linguistica di testi caratterizzati da variazione grafica, alternanza di registri e *code-switching*. Attraverso una pipeline che integra OCR generativo, annotazione automatica secondo lo schema *Universal Dependencies* e validazione manuale in ambiente INCEpTION, il lavoro mostra come errori e ambiguità annotative non costituiscano semplici rumori, ma rendano visibili i limiti degli standard esistenti e possano essere valorizzati come risorse didattiche. Il contributo propone quindi una riflessione sulla testualità digitale araba come spazio di mediazione tra complessità linguistica reale e strumenti computazionali.
Parole chiave: testualità digitale araba; OCR; parsing morfosintattico; Universal Dependencies; didattica dell'arabo

1. INTRODUCTION[1]

The teaching of Arabic at the university level is characterized by a marked heterogeneity of approaches, learning objectives, and disciplinary traditions. Arabic language courses are situated within different academic pathways, including linguistic, literary, historical and cultural studies, or area studies, giving rise to pedagogical models that may privilege the acquisition of communicative competence, formal linguistic analysis, the reading and interpretation of literary texts, or a combination of these dimensions. While this plurality represents a scientific and pedagogical asset, it also complicates the design of shared teaching tools and resources capable of addressing diverse educational needs without flattening their specificities. Within this fragmented landscape, the text, understood both as literary and non literary, continues to occupy a central position. Textual materials constitute the primary means of access to Arabic language, culture, and discursive practices, and play a crucial role in the teaching of language as well as literature and culture. However, authentic texts are rarely used in teaching contexts, especially in the early stages of acquisition, as they often display a level of linguistic complexity that reflects structural features of contemporary Arabic, such as orthographic variation, the coexistence of registers and varieties, and the use of code-switching, elements that are difficult to reconcile with fully standardized linguistic models. Moreover, this textual complexity frequently comes into tension with the digital tools currently available for the analysis and automatic processing of Arabic, which tend to presuppose a

[1] Authors responsibility: Elisa Gugliotta is responsible for funding acquisition, methodology, project administration, supervision, visualization and writing - original draft; all authors contributed to conceptualization, formal analysis, investigation, data curation, validation and writing – review & editing.

normalized and relatively homogeneous language. In particular, Optical Character Recognition (OCR) technologies and linguistic annotation systems show significant limitations when applied to texts that diverge from standard varieties or that exhibit linguistic and graphic hybridity. The result is a gap between the linguistic reality of the materials actually used in teaching and the digital resources that can be employed to support their analysis, mediation, and reuse.
It is precisely within this space of friction between the real linguistic complexity of texts and digital tools that both the RĀBIṬA project[2] and the reflection developed in this contribution are situated, with the aim of critically examining the potential and limitations of computational practices applied to Arabic texts from a pedagogically oriented perspective. Through a case study, we present the obstacles encountered and the solutions identified at different stages of text processing, from the transformation of printed material into digital form to the open access provision of teaching materials designed to support courses in Arabic language, culture, history, and literature.

2. THEORETICAL FRAMEWORK

In recent years, the field of Digital Humanities has devoted increasing attention to the computational study of texts, developing methodologies and tools for the analysis, representation, and annotation of literary corpora. However, much of this reflection has focused on European textual traditions or on languages characterized by relative graphic and normative stability. In non European contexts, and particularly for languages with complex writing systems and strong internal variation, the application of models and tools developed elsewhere raises theoretical and methodological questions that have only been partially explored (Fiormonte, Chaudhuri and Ricaurte, 2022). In the case of Arabic, digital textuality is confronted with a series of structural challenges related to the nature of the writing system, orthographic variability, and the coexistence of registers and linguistic varieties (Alayba, 2025). The literature has extensively documented the difficulties encountered by OCR technologies when applied to Arabic texts, especially in the presence of non-standardized printed editions, heterogeneous fonts, or materials that reflect hybrid linguistic practices (Alghyaline, 2023). Errors produced during the recognition phase are not limited to simple graphic inaccuracies, but significantly affect the morphosyntactic structure of the text, compromising the quality of subsequent stages of automatic processing. These issues are compounded by those related to linguistic annotation and the adoption of shared standards such as Universal Dependencies (UD). Although such standards represent a fundamental reference point for computational linguistics, their application to Arabic texts characterized by variation, code-switching, or other contact phenomena raises important questions (Saadiyeh et al., 2025; Gaser et al., 2023; Riabi, Sagot and Seddah, 2021). The tension between the need for normalization and fidelity to the linguistic reality of the text emerges particularly clearly in contexts where variation is not noise to be eliminated, but a constitutive feature of the linguistic data. While these aspects have been widely discussed in computational and linguistic research, less attention has been devoted to their implications in teaching contexts. In particular, the way in which the challenges of Arabic digital textuality, from OCR to annotation, can be integrated into a pedagogical reflection that takes into account the plurality of approaches to teaching Arabic and the central role of the text as an object of mediation remains largely underexplored.

3. FROM PAGE TO DATA: A PIPELINE FOR ARABIC

The transformation of a printed Arabic text into an annotated digital resource involves a series of steps that cannot be considered purely technical, but instead entail significant theoretical and methodological choices. In particular, in the case of texts characterized by linguistic and graphic variation, each stage of the pipeline, from digitization to annotation, contributes to shaping the data and determining its possibilities for analysis and reuse, especially in teaching contexts.
Printed texts and graphic complexity. Printed texts in Arabic present a graphic complexity that reflects both the editorial history of the materials and the linguistic practices underlying their production. Typography, the non uniform use of diacritical marks, the management of graphic variants, and the layout of the text on the page are all elements that significantly affect readability and subsequent digitization, thus constituting a primary source of complexity that impacts the entire pipeline.

[2] Further information on RĀBIṬA (Resources fo Arabic: Bridging Interoperable Texts and Annotation) is available at: https://github.com/eligugliotta/RABITA.

OCR: errors, limits, and choices. The use of OCR technologies represents a crucial step in the digitization process of Arabic texts, but also one of the most critical points in the entire pipeline. In this study, the OCR process is understood not as a simple automatic transcription, but as a generative image to text transduction carried out through a Vision Language Model. Following comparative tests conducted on different systems, QARI was adopted, an Arabic OCR model obtained through fine tuning of the multimodal model Qwen2-VL-2B-Instruct (Wasfy et al., 2025). This choice departs from traditional deterministic OCR approaches, favoring a model capable of integrating visual information and linguistic generation, particularly suited to materials characterized by heterogeneous scan quality and strong graphic variability. The adoption of a generative OCR system, however, introduces new types of errors that do not concern only the recognition of graphemes, but also the handling of word boundaries, the reproduction of repetitive sequences, or potential interference between writing systems (see Examples 1, 2, and 3, Sec. 4). For this reason, the choice of the OCR model and of the generation parameters, as well as pre and post processing strategies, takes on a central methodological value. These decisions directly affect the final representation of the linguistic data and the possibility of making visible, or conversely attenuating, phenomena of variation and language contact that constitute an integral part of the texts under analysis.
Normalization and annotation. The normalization and annotation phase is the stage at which the tensions between standardization and fidelity to the text become most evident. The texts obtained from the OCR phase are first subjected to automatic linguistic annotation carried out through CamelParser (Elshabrawy et al., 2023), a system based on CAMeL Tools (Obeid et al., 2020) that enables morphological analysis and POS tagging of texts in Arabic script, interpreted according to the morphological model of Modern Standard Arabic while also recognizing lexical and morphological forms typical of dialectal varieties, thus generating an output compliant with the UD scheme. The adoption of UD responds to the need to ensure interoperability, comparability, and reuse of the data, situating the produced resources within widely shared annotation ecosystems. The output of the automatic annotation, exported in CoNLL-U format, is subsequently imported into the INCEpTION infrastructure (Klie et al., 2018), where a phase of manual validation and refinement of the annotation takes place, including the addition of further annotation layers oriented toward teaching purposes. However, applying such standards to Arabic texts characterized by graphic and linguistic variation entails mediating between the requirements of the standard and the linguistic reality of the text (see Fig. 1, Sec. 4). From a pedagogically oriented perspective, these choices acquire additional significance, as they determine the type of access to the text offered to students and the way in which linguistic complexity is either made visible or, conversely, attenuated.

4. CASE STUDY: *WƏLD FAḌĪLA*

The case study selected for this analysis is *Wəld Faḍīla*, a narrative text published by Amira Charfeddine in 2019, which addresses issues of identity, social marginalization, and the normative control of bodies and subjectivities, with particular attention to gender dynamics and the stigma associated with sexuality (Achour, 2022). *Wəld Faḍīla* is particularly well suited to highlighting the challenges discussed in the previous sections. The text displays a linguistic and stylistic stratification that reflects the colloquial discursive practices of contemporary Tunis, positioning itself beyond the conventional boundary between norm and usage, and between standard variety and marked linguistic forms, traditionally drawn in the selection of materials for university teaching contexts. From a linguistic perspective, *Wəld Faḍīla* shows a marked presence of orthographic and morphosyntactic variation, as well as alternation between registers and urban varieties. The use of code-switching, together with the emergence of sociolinguistically marked features such as profanity, contributes to the creation of a heterogeneous textual fabric that is not a marginal element of the text, but a structural component of its narrative construction and communicative function. The application of OCR tools to a text of this kind makes particularly evident the limitations of OCR technologies for Arabic. Linguistic sequences that diverge from expected forms tend to generate a high number of errors that cannot be reduced to simple graphic inaccuracies, but instead affect segmentation, readability, and, in some cases, the interpretability of the text. In particular, phenomena of variation and code-switching are often rendered inadequately, with effects that compromise the possibility of subsequent coherent linguistic annotation, as illustrated by the following examples.

(1) I came to warn you about your son Fadel. (p. 18)

Original Text
- شو، راو جيتك بش نفطنك على فاضل ولدك .

OCR output
- شو، راو جيتك بش نفطنك على فاضل ول dek.

Segmented text
شو راو جيت-ك بش ن-فطّن-ك على فاضل‌ولد-ك

Glosses
INTJ PRAG 2SG.F-come.PFV.1SG FUT 2SG.F-warn-1SG PREP Fadel 2SG.F-son

(2) Oh, poor mother, he has killed himself! (p.24)

Original Text
- يا ميمتيييييييي إنتحر !

OCR output
- يا ميمتيييبيرييييي إنتحر !

Segmented Text
يا ميمت-ي إنتحر

Glosses
VOC 1SG.POSS-Mother.DIM suicide.PFV.3SG.M

(3) Oh Lord! He has thrown himself down, it's over! (p.24)

Original Text
- يا رمى روحو لطيف سايي !

OCR output
- يا и رمى روحو لط ф سايي !

Segmented Text
يا لطيف رمى روح-و سايي

Glosses
VOC INTJ.RELIG throw.PFV.3SG.M self-3SG.M PRAG.FR

In Ex. 1, we observe that during the OCR phase part of the text is transduced into Latin characters (*dek*). This error can be traced back to incorrect segmentation and graphic confusion, generating a token with no linguistic status. This type of error compromises the identification of morphemic boundaries and, consequently, syntactic and semantic annotation. Ex. 2 shows a different but equally significant case. The OCR reproduces the lengthening typical of writing that mimics orality. However, additional Arabic graphemes are introduced that do not correspond to the original text. Ex. 3 highlights how the OCR system produces a sequence of characters in the Cyrillic alphabet (*иф* /if/) instead of a correct rendering in Arabic script. This type of error cannot be attributed to graphemic confusion within the writing system adopted for the novel, but rather signals a more radical failure in the script identification process, with the uncontrolled shift to a set of characters belonging to a different graphic domain. By contrast, it is noteworthy that the element سايي a borrowing from French (*ça y est*), is recognized by the OCR as an integral part of the Arabic text. This observation leads to the subsequent stage of text processing, namely linguistic annotation, which raises important methodological questions.

For the project, the INCEpTION infrastructure was adopted, as it allows the integration of different POS formalisms and supports, among others, the UD scheme, which is the same annotation framework used for the automatic annotation performed with CamelParser. This continuity between automatic annotation and the revision environment makes it possible to frame manual annotation as a process of validation and refinement of the data. This operational simplification does not eliminate, but rather makes more visible, the theoretical choices involved in adopting a standardized annotation scheme such as UD. In particular, the need to map the data onto predefined formal categories creates tension between the abstract representation of the text and fidelity to its linguistic and discursive specificities. Colloquial elements and hybrid or mixed language sequences require compromise solutions that make complexity explicit without sacrificing the internal coherence of the annotation scheme. In this sense, the text becomes a testing ground for reflecting on the limits and possibilities of existing standards when applied to non-standard materials.

Fig. 1 illustrates the methodology adopted during the manual validation and annotation phase of the text. First, the pragmatic particle تي initially associated with a fallback value (NOAN),[3] is reinterpreted as an interjection (INTJ), making its discursive and interactional role visible. This type of intervention is particularly significant, as it shows how INCEpTION allows the standard annotation scheme to be complemented by a context sensitive reading, without compromising the overall coherence of the annotation and while retaining the NOAN tag as a marker of non-standard elements. Second, the POS (in orange), the lemma (in light blue), and the conjugated form (in green) of the Tunisian verb تخافش (where the final grapheme constitutes the second element of the circumfix negation *mā*-VERB-*š*) are corrected and complemented by the verbal root (in purple),[4] thus resolving the ambiguity left unresolved during the automatic annotation phase.

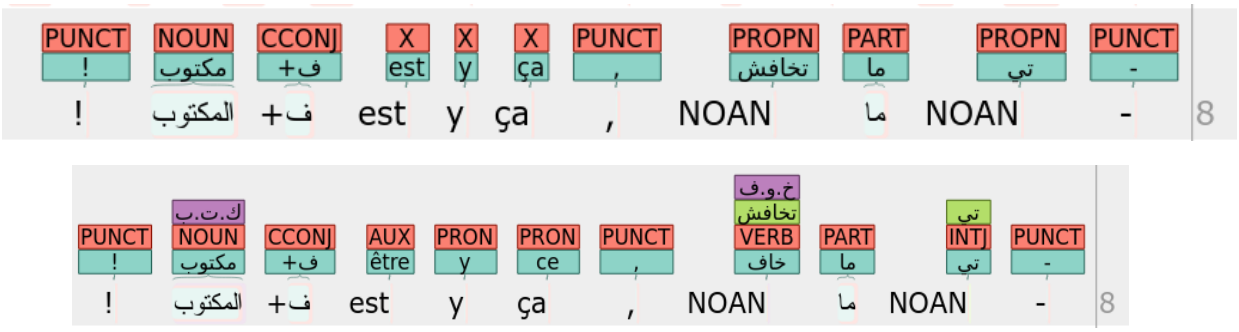


Figure 1: Annotation of an excerpt from *Wəld Faḍīla* in the INCEpTION environment. The sentence means: "Come on, do not be afraid, it is written in destiny!" At the top, the output of the automatic annotation. At the bottom, the same sentence after validation and manual intervention.

Finally, the French sequence *ça y est*, automatically handled as a series of isolated tokens and associated with residual categories (X), is reanalyzed according to its pragmatic function in context. These interventions make explicit an interpretive choice that automatic annotation, by its very nature, is unable to perform when dealing with texts that are so noisy and heterogeneous. In support of this claim, it is sufficient to note that *ça y est* appears elsewhere in the text written in Arabic characters (سايي), revealing the absence of a consistent orthographic treatment of borrowings within the text itself. This graphic oscillation further highlights the difficulty of establishing a clear boundary between what can be considered a borrowing that has become integrated into the language and what continues to be perceived as an exogenous insertion, thereby challenging annotation approaches based on a binary distinction between native and foreign forms. From a pedagogical perspective, such ambiguities do not represent a limitation but rather a resource. The visualization of the annotation allows students to engage with the gradual and negotiated nature of lexical integration processes, as well as with the role played by orthographic and annotation choices in the construction of meaning. In this sense, annotation does not merely structure linguistic data, but becomes a mediating tool that makes the complexity of non canonical texts explicit and encourages critical reflection on the limits and possibilities of existing standards when applied to linguistically and culturally situated materials. *Wəld Faḍīla* thus emerges not only as an object of analysis, but as a teaching resource capable of making the linguistic complexity of Arabic visible and fostering a deeper awareness of the processes involved in constructing textual data in a digital environment.

Indeed, RĀBIṬA resources are designed to function as dynamic pedagogical tools. A pilot phase of classroom implementation is planned for the upcoming academic year. For instance, students may be invited to work directly on OCR outputs, comparing them with the original printed text in order to identify recognition errors and reflect on their linguistic and graphic causes. This activity allows learners to engage critically with the notion of "error", shifting from a prescriptive perspective to an analytical one. A second type of activity focuses on annotation. Using the INCEpTION environment, students may be asked

[3] The NOAN label is a "No Analysis" marker generated during text processing with CamelParser. This tag highlights categorization difficulties and signals lexically problematic segments. Such labels can be useful in the exploratory analysis of the text to identify points of linguistic variation or parsing errors. By contrast, X is an official POS category in the UD scheme, used to indicate elements classified as other or unclassified.

[4] The root was also added for the noun مكتوب "destiny".

to validate or revise automatically generated annotations, particularly in cases involving colloquial forms, code-switching, or pragmatically marked elements. This task is expected to encourage reflection on the limits of standardized annotation schemes such as Universal Dependencies and to highlight the interpretive dimension of linguistic analysis. Finally, selected excerpts from the corpus can be used for task-based activities, such as role-play or guided rewriting, where students must reformulate linguistically complex or hybrid sequences into different registers or varieties. In this way, the annotated text becomes a mediating object that connects computational analysis, linguistic awareness, and communicative practice.

5. CONCLUSIONS AND PERSPECTIVES

This contribution has offered a methodological reflection on digital textual practices applied to complex Arabic texts, highlighting how the transition from the printed page to annotated data cannot be reduced to a sequence of technical operations, but involves interpretive choices with significant implications for teaching. Through the analysis of a pipeline integrating digitization, OCR, and linguistic annotation, it has been shown how the graphic and linguistic complexity of Arabic texts constitutes both a critical challenge and an epistemic resource for the teaching of language, culture, and literature. The case study discussed has concretely illustrated the limitations of currently available computational tools when applied to materials characterized by variation, code-switching, and non-standard features. These limitations have not been interpreted as mere technological shortcomings, but rather as indicators of the tensions between standardization, data representation, and fidelity to the linguistic reality of the text. In this sense, the adoption of shared annotation schemes such as Universal Dependencies appears not as a definitive solution, but as a space of methodological negotiation requiring constant critical reflection.
From a forward-looking perspective, the reflections presented here open the way to further developments both at the methodological and applied levels. The extension of the analysis to additional texts and genres, the experimentation with alternative OCR and annotation strategies, and the construction of interoperable and reusable digital resources are part of an ongoing research trajectory of which this contribution presents an initial methodological application. These developments contribute to fostering a closer dialogue between Digital Humanities, computational linguistics, and the teaching of Arabic, placing digital textuality at the center of educational practices attentive to linguistic complexity and the long-term sustainability of resources.
Beyond its methodological and pedagogical implications, the RĀBIṬA project also contributes to public engagement by promoting access to linguistically complex Arabic texts and making the processes underlying their digital transformation visible to a broader audience. By publishing annotated resources in open-access formats and designing their integration into teaching and training activities, the project fosters a critical awareness of how language technologies operate, their limitations, and their interpretive dimension. In this sense, RĀBIṬA does not only produce data, but also encourages forms of digital literacy that are particularly relevant in the context of under-resourced languages, where technological mediation often remains opaque or inaccessible.

ACKNOWLEDGEMENTS

This research was funded by the Italian National Recovery and Resilience Plan (PNRR), Mission 4 "Education and Research" Investment 1.2 "Funding projects presented by young researchers", supported by the European Union – NextGenerationEU, CUP J83C25000330006. RĀBIṬA is hosted at the Università degli Studi di Sassari and is conducted under the supervision of Elisa Gugliotta.
The project involves a collaboration with the CNR-ILC, in particular with Francesca Frontini and Giulia Pedonese, for monitoring the FAIRification processes of linguistic and educational materials and their dissemination through the CLARIN-IT H2IOSC training platform.

REFERENCES

Achour, M. (2022). La diglossie traversée: La littérature en tunisien et le tunisien dans la littérature. *International Journal of the Sociology of Language*, *2022*(278), 23-50.

Alayba, A. M. (2025). Arabic Natural Language Processing (NLP): A Comprehensive Review of Challenges, Techniques, and Emerging Trends. *Computers, 14*(11), 497.

Alghyaline, S. (2023). Arabic Optical Character Recognition: A Review. *Computer Modeling in Engineering & Sciences* (*CMES*), *135*(3).

Elshabrawy, A., AbuOdeh, M., Inoue, G., & Habash, N. (2023). CamelParser 2.0: A State-of-the-Art Dependency Parser for Arabic. In *Proceedings of the First Arabic Natural Language Processing Conference* (*ArabicNLP 2023*).

Fiormonte, D., Chaudhuri, S., & Ricaurte, P. (Eds.). (2022). *Global debates in the digital humanities*. U of Minnesota Press.

Gaser, M., Mager, M., Hamed, I., Habash, N., Abdennadher, S., & Vu, N. T. (2023, May). Exploring segmentation approaches for neural machine translation of code-switched Egyptian Arabic-English text. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 3523-3538).

Klie, J.-C., Bugert, M., Boullosa, B., Eckart de Castilho, R. and Gurevych, I. (2018): The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In *Proceedings of System Demonstrations of the 27th International Conference on Computational Linguistics* (*COLING 2018*), Santa Fe, New Mexico, USA.

Obeid, O., Zalmout, N., Khalifa, S., Taji, D., Oudah, M., Alhafni, B., Inoue, G., Eryani, F., Erdmann, A. & Habash, N. (2020, May). CAMeL tools: An open source python toolkit for Arabic natural language processing. In *Proceedings of the twelfth language resources and evaluation conference* (pp. 7022-7032).

Riabi, A., Sagot, B., & Seddah, D. (2021, November). Can Character-based Language Models Improve Downstream Task Performances In Low-Resource And Noisy Language Scenarios?. In *Proceedings of the Seventh Workshop on Noisy User-generated Text* (*W-NUT 2021*) (pp. 423-436).

Saadiyeh, O., Ramadan, A., Hajjar, M., & Bernard, G. (2025). A comparative study of Arabic syntactic analyzers*. Frontiers in Artificial Intelligence*, *8*, 1638743.

Wasfy, A., Nacar, O., Elkhateb, A., Reda, M., Elshehy, O., Ammar, A., & Boulila, W. (2025). QARI-OCR: High-Fidelity Arabic Text Recognition through Multimodal Large Language Model Adaptation. *arXiv preprint arXiv:2506.02295*.

Dal markup all'interpretazione: interrogare il linguaggio della critica verista in un corpus XML/TEI

DENISE BRUNO[1], SALVATORE CRISTOFARO[2], ANTONIO DI SILVESTRO[3], LAURA MAZZAGUFO[4], DARIA SPAMPINATO[5], GIUSEPPE ZAPPALÀ[6]

[1]UNIVERSITÀ DI CATANIA – DENISE.BRUNO@PHD.UNICT.IT
[2]CNR - ISTITUTO DI SCIENZE E TECNOLOGIE DELLA COGNIZIONE – SALVATORE.CRISTOFARO@CNR.IT
[3]UNIVERSITÀ DI CATANIA – ANTONIO.DISILVESTRO@UNICT.IT
[4]UNIVERSITÀ DI CATANIA – LAURA.MAZZAGUFO@PHD.UNICT.IT
[5]CNR - ISTITUTO DI SCIENZE E TECNOLOGIE DELLA COGNIZIONE – DARIA.SPAMPINATO@CNR.IT
[6]UNIVERSITÀ DI CATANIA – GIUSEPPE.L.ZAPPALA@GMAIL.COM

ABSTRACT (ITALIANO)

Verbum è un lessico-repertorio tematico del linguaggio della critica verista, sviluppato nel progetto *COVerLeSS* attraverso l'annotazione semantica in XML/TEI di testi critici pubblicati tra il 1872 e il 1890. Il contributo ne descrive l'organizzazione concettuale e la formalizzazione nell'ontologia OntoTerm secondo i paradigmi del Semantic Web. Attraverso analisi quantitative condotte con *NormaTEI*, l'articolo mostra come l'annotazione dei termini consenta di indagare la distribuzione temporale, la polisemia e la densità concettuale del lessico critico verista.
Parole chiave: edizione scientifica digitale; XML/TEI; lessico critico letterario; corpus linguistics; ontologia formale

ABSTRACT (ENGLISH)

From Markup to Interpretation: Querying the Language of Verismo Literary Criticism in an XML/TEI Corpus.
Verbum is a thematic lexicon-repertoire of Verist critical language developed within the *COVerLeSS* project through XML/TEI semantic annotation of reviews, essays, and journal articles published between 1872 and 1890. This paper describes the conceptual and formal structure of *Verbum*, based on the identification of types, subtypes, main words, and semantic categories, and its formalization in the OntoTerm ontology following Semantic Web paradigms. Through quantitative analysis performed with *NormaTEI*, the article demonstrates how term annotation enables the analysis of temporal distribution, polysemy, and conceptual density in Verist critical vocabulary.
Keywords: digital scholarly editing; XML/TEI; literary critical lexicon; corpus linguistics; formal ontology

1. INTRODUZIONE

Il progetto *COVerLeSS* (*Corpus Online del Verismo tra Letteratura, Storia e Società*) si colloca nel quadro delle ricerche di Digital Humanities dedicate allo studio dei periodici e della letteratura secondaria, e propone un ambiente digitale integrato per la conservazione, l'edizione e l'analisi computazionale della ricezione e del dibattito critico sul Verismo italiano. Il progetto nasce con l'obiettivo di restituire in forma strutturata e interrogabile un *corpus* ampio e testualmente eterogeneo di recensioni, saggi e articoli in rivista pubblicati tra il 1872 e il 1890, un arco cronologico cruciale per la formazione dell'identità letteraria e culturale dell'Italia postunitaria.
Il progetto adotta un modello di rappresentazione testuale basato sulla codifica XML/TEI, nella quale integra annotazione strutturale, descrittiva e semantica, al fine di rendere i testi interrogabili a diversi livelli di granularità. In questa prospettiva, COVerLeSS si configura come un'edizione-archivio o *knowledge site* (Tomasi, 2016), in cui la lettura dei testi è affiancata da strumenti di esplorazione sincronica e diacronica dei dati, secondo un approccio che combina *Close* e *Distant Reading* e che mira a valorizzare il potenziale interpretativo dei dati strutturati. A tal fine, un elemento centrale del progetto è la costruzione di *Verbum*, un lessico-repertorio tematico del linguaggio critico verista, basato sull'annotazione di lessemi, strutture sintagmatiche e formule ricorrenti, organizzati in campi semantici e caratterizzati tramite attributi tipologici e concettuali.

Il presente contributo intende mostrare come strumenti di interrogazione di documenti XML/TEI possano supportare una riflessione sul linguaggio della critica letteraria e sulla modellazione formale dei dati in ambito umanistico, incentivando la formulazione di nuove domande di ricerca.

2. DESCRIZIONE E CONTENUTI DI *VERBUM*

Il corpus del progetto, già descritto in Bruno *et al.*, 2025, è stato sottoposto a ulteriore selezione, individuando i testi che, per la specificità del linguaggio in essi adoperato, si prestano a un'analisi granulare qual è quella finalizzata alla costruzione di *Verbum*. In particolare, ad oggi, sono stati selezionati nel complesso 163 documenti, di cui 3 saggi, 97 articoli tratti dalla rivista «Il Momento» e 63 recensioni alle opere di Verga. A livello cronologico, i saggi risalgono al 1879 e al 1883, gli articoli de «Il Momento» al triennio 1883-1885, mentre le 63 recensioni coprono un arco cronologico che va dal 1874 al 1890, pur non in modo omogeneo. Come già dichiarato (Bruno *et al.*, 2025: 9-13),

> La codifica strutturale e quella relativa alle entità nominate costituiscono la base di un'annotazione specifica, comprendente lessemi, combinazioni sintagmatiche, qualche volta strutture fraseologiche, volte a costituire una sorta di dizionario tematico nel quale trovano spazio gli elementi caratterizzanti della dimensione sociale, della poetica, della costruzione narrativa, dello stile della letteratura verista. Non si tratta di approdare [...] ad un vocabolario lemmatizzato, [...] quanto di isolare un campionario rappresentativo della terminologia caratterizzante il discorso critico, la recensione, la polemica letteraria, verificando il tasso di 'tecnicità' con cui quella terminologia ricorre nelle diverse testualità esposte nel corpus. Piuttosto che di un dizionario 'reale', abbiamo preferito caratterizzare *Verbum* come un dizionario 'ideale', nel quale il lemma base, la singola parola, si trovano spesso all'interno di combinazioni sintagmatiche (ad es. sostantivo + aggettivo, verbo + complementi, locuzioni modali) che consentono di circostanziare nelle sue articolazioni il linguaggio specifico della letteratura secondaria.

La creazione di *Verbum* nasceva dalla necessità di individuare degli ambiti concettuali caratterizzanti il campo della creazione letteraria e dell'interpretazione critica nell'ambito del verismo, a cui far corrispondere degli insiemi lessicali abbastanza ampi. Da qui l'individuazione di tre type-guida, ossia "verbum", "forma_letteraria" e "stile", da cui sono stati 'generati' 13 subtype in grado di rappresentare un lessico-repertorio esaustivo atto a consentire una vera e propria 'navigazione' tra autori, testi, poetiche, ricezioni (Fig.1).



Figura 1. Distribuzione dei @type e dei corrispondenti @subtype all'interno del *corpus* analizzato

Dalle prime analisi dei dati ricavati da questa classificazione emerge in modo perspicuo la polisemia di numerosi lessemi e sintagmi, ognuno dei quali si estende ad ambiti semantici diversi. Ad es., il termine "forme" ricopre gli ambiti relativi al campo dell'autorialità-narratività, dello studio-osservazione e del genere letterario. Il termine "sincerità" afferisce sia al subtype "rappresentazione" sia a quello "vero/verità". Ad emergere dalle prime estrazioni e analisi dei dati è l'omogeneità di lessico di un movimento letterario e culturale che, pur estendendosi dalla letteratura alla filosofia alla sociologia, si

caratterizza per una dimensione speculativa fortemente unitaria, e conseguentemente portata a esprimersi mediante un vocabolario altrettanto unitario e compatto. Da qui la necessità di creare delle parole dominanti, ossia dei termini-ombrello in grado di abbracciare elementi afferenti a diversi subtype, che hanno consentito di valorizzare la dimensione polisemica del lessico. Queste parole dominanti costituiscono un repertorio 'ristretto' di poco più di 100 termini, che per la loro ampiezza rappresentano un dizionario tematico di secondo livello. Si tratta però in questo caso di un dizionario 'ideale', strutturato per concetti, con parole a esponente che non corrispondono alle reali occorrenze.
A un livello ulteriore, le parole dominanti sono state ricondotte a 7 categorie, che sono le vere e proprie radici dell'albero lessicale realizzato. All'interno di esse sono stati ricompresi i campi operativi e di pensiero fondanti del movimento verista: "genere_letterario", "creazione_letteraria", "corrente_letteraria", "verismo_sociale", "riproduzione_del_vero", "visualità", "aspetto_linguistico".
L'inserimento di parole dominanti e categorie costituisce pertanto un ampliamento del lessico-repertorio *Verbum*. Occorre tuttavia esplicitare la relazione che sussiste tra type e subtype da un lato, tra parole dominanti e categorie dall'altro. L'individuazione dei subtype, collegati ai tre type "forma_letteraria", "verbum" e "stile", deriva dall'analisi dei contesti d'uso delle occorrenze; questo determina la loro centralità nell'organizzazione del lessico-repertorio, sia verso il 'basso', cioè verso il massimo grado di granularità legato alle occorrenze, sia verso l'"alto', cioè verso il raggruppamento progressivo in entità lessico-semantiche sempre più omnicomprensive. Il rapporto tra categorie e parole dominanti invece mira a costruire una sorta di aggregatore concettuale che consente un approccio olistico al linguaggio del movimento verista, visto nella sua declinazione interdisciplinare. Questa complessa organizzazione di *Verbum* viene rappresentata con un'ontologia formale, illustrata nel paragrafo successivo.
La distribuzione aggregata delle parole dominanti di *Verbum* per categoria semantica consente di osservare il peso relativo dei principali campi concettuali del linguaggio della critica verista (Fig. 2).

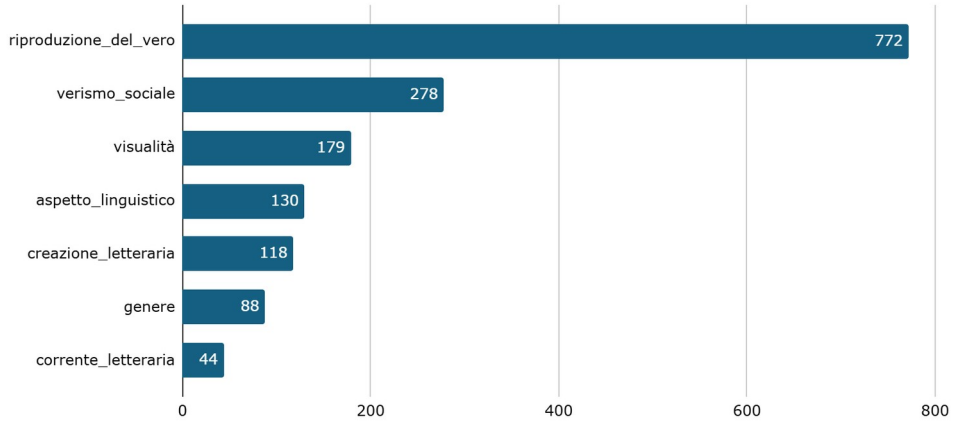


Figura 2. Distribuzione, aggregata per categoria semantica, delle parole dominanti di *Verbum*

L'analisi delle parole dominanti evidenzia la struttura interna delle categorie semantiche, mostrando come alcune aree del lessico siano costituite da numerosi termini di frequenza media, mentre altre presentino un numero più limitato di termini ma con maggiore densità concettuale. La visualizzazione (Fig. 3) mostra come alcune categorie, per esempio "riproduzione_del_vero" (bolle in blu), contengano numerosi termini ad alta frequenza, tra cui *descrizione*, *determinismo* e *studio*, mentre altre, come "visualità" (in violetto), siano caratterizzate da poche parole ma particolarmente centrali, come *pittura* e *colore/colorito*. Allo stesso tempo, la rappresentazione permette di identificare *outlier* concettuali, come *espressione* nell'ambito dell'"aspetto_linguistico" (in verde acqua), che assumono un ruolo predominante nella rete semantica. Categorie di media ampiezza, come "verismo_sociale" (in rosa) e "creazione_letteraria" (in verde), mostrano una distribuzione più equilibrata di termini, con parole rilevanti quali *animazione*, *ambiente*, *classico_romantico* e *intreccio*. In Fig. 3 sono rappresentate anche le categorie "genere" in grigio e "corrente_letteraria" in rosso-bruno.

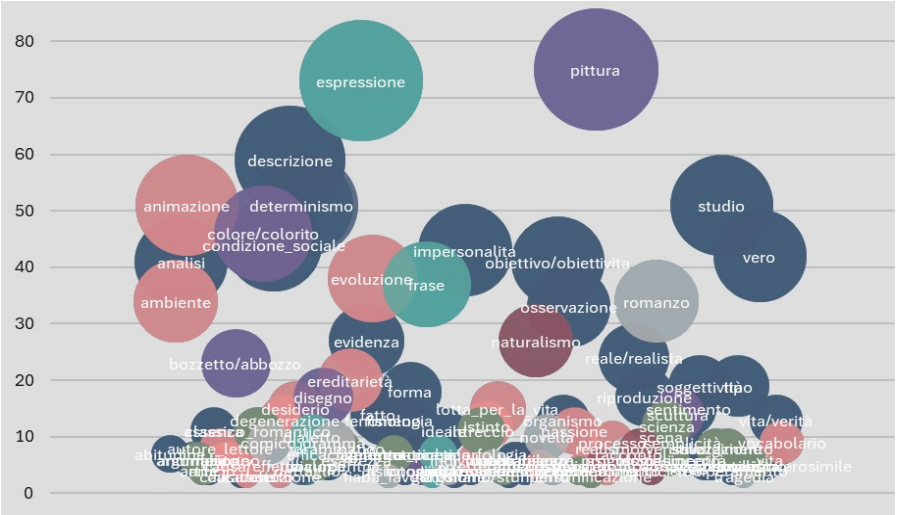


Figura 3. Distribuzione delle parole dominanti di *Verbum*. Ogni bolla rappresenta una parola con dimensione proporzionale al numero di occorrenze nel dataset. L'asse Y indica la frequenza dei termini, mentre i colori distinguono le categorie.

3. ONTOTERM

L'organizzazione concettuale di *Verbum* trova una formalizzazione esplicita nell'ontologia OntoTerm (Fig. 4), concepita secondo il paradigma del Semantic Web applicato all'editoria digitale scientifica (Spadini *et al.*, 2021).

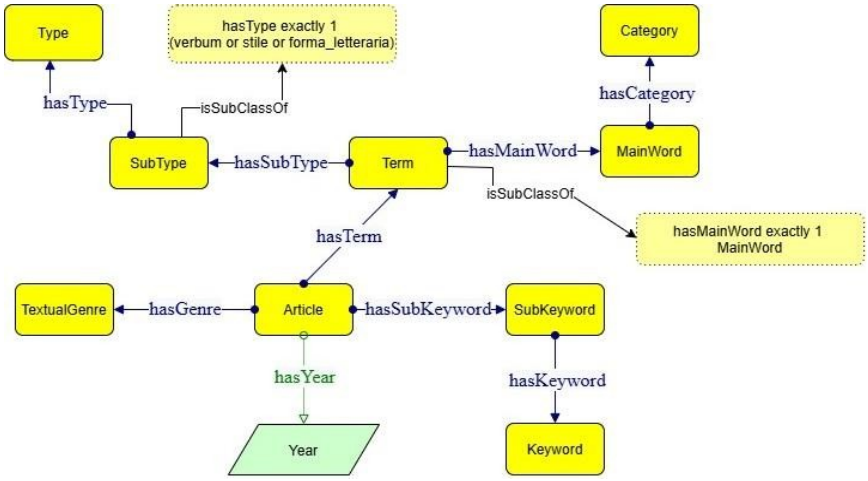


Figura 4. Rappresentazione dell'organizzazione concettuale di *Verbum*

L'ontologia proposta mira a formalizzare e integrare la complessità del sistema concettuale di *Verbum*, articolata attorno alla classe *Term*, che funge da nodo semantico centrale e consente di integrare dimensioni lessicali, tassonomiche e bibliografiche. In questa prospettiva, l'ontologia non si limita a descrivere entità linguistiche isolate, ma costituisce un sistema capace di rappresentare la complessità delle pratiche discorsive e delle classificazioni concettuali.
Accanto ai termini di *Term* si distinguono due diversi livelli di classificazione (vd. par. 2): da un lato, il sistema *Type-SubType*, dall'altro il sistema *MainWord-Category*, che costituiscono due sottografi non direttamente connessi tra loro, ma comunque interferenti attraverso rami di relazione che si estendono dal nodo *Term*. In particolare, il rapporto decisivo è quello che si inferisce tra *SubType* e *MainWord* e – data la relazione 1:1 tra *MainWord* e *Category* – quello fra *SubType* e *Category* (cioè ogni categoria è collegata a diversi type e subtype).
La classe *Term* è definita come un'unità concettuale dotata di una *MainWord*, ovvero una classe dominante che ne costituisce il punto di accesso semantico. Tale relazione, vincolata da una cardinalità 1:1, richiama la distinzione tra lemma e forma canonica presente nei modelli lessicografici

contemporanei, ma ha una valenza maggiormente semantica, configurandosi come ponte fra termine e categoria semantica (*Category*). Accanto alla dimensione formale, il termine è ulteriormente caratterizzato da proprietà che ne specificano la natura tassonomica (*Type* e *SubType*), definita per *Verbum* con specifici riferimenti alle pratiche e alle tematiche del verismo, mentre l'associazione a parole chiave e sotto-parole chiave (*Keyword* e *SubKeyword*) facilita l'annotazione semantica fine e la classificazione dei documenti del corpus all'interno di specifici percorsi didattici (vd. par. 4). Questa articolazione consente di modellare sia la struttura interna del lessico a livello semantico sia le sue relazioni pragmatiche all'interno del *corpus*.
Vi è, inoltre, una componente documentaria, affidata alle classi *Article*, *TextualGenre* e *Year*, che permettono di collocare i termini all'interno di un contesto testuale e cronologico, mettendo in relazione le diverse tipologie testuali presenti nel corpus e la diacronia analizzata.
L'ontologia si presta a un'integrazione naturale con gli standard attualmente più diffusi. In particolare, il modello OntoLex-Lemon, sviluppato dal W3C Ontology-Lexicon Community Group (McCrae *et al.*, 2017), offre un quadro concettuale pienamente compatibile con la distinzione tra forma linguistica, senso e concetto. Le categorie concettuali e le parole chiave possono essere modellate tramite SKOS (Miles & Bechhofer 2009), spesso utilizzato in combinazione con OntoLex per la rappresentazione di tassonomie e vocabolari controllati. Parallelamente, l'ontologia può dialogare con le linee guida TEI Lex-0 (Romary & Tasovac 2018), che mirano a normalizzare la rappresentazione dei dizionari digitali all'interno del framework TEI. L'integrazione con TEI Lex-0 renderebbe l'ontologia compatibile con le pratiche consolidate delle Digital Humanities e con i progetti di linguistica computazionale.

4. CATEGORIZZAZIONE DEL CORPUS DELLE RIVISTE

Il corpus è completato dalle serie integrali di quattro delle riviste dell'epoca impegnate nella diffusione dei principi del naturalismo francese e in una sua 'traduzione' italiana. Il reperimento, l'OCR e la codifica delle riviste contemplate nel progetto hanno permesso di sondare più da vicino la risonanza dei temi che gravitano attorno al Verismo, tanto da proporre un livello di codifica specifico che potesse restituire il ruolo delle riviste stesse nell'architettura tematica e concettuale di *COVerLeSS*. In particolare, data la differente natura della «Rassegna Settimanale» rispetto alle altre tre riviste prese in esame, si è deciso di procedere con una codifica per parole chiave: ambiente urbano, condizione femminile, lavoro, verismo, vita rurale,[1] che restituisse gli interessi comuni fra la linea editoriale della testata e l'interesse dei veristi verso l'oggetto prioritario della rappresentazione artistica, ossia la condizione esistenziale delle masse popolari del meridione. Le parole chiave identificate nella «Rassegna Settimanale» si configurano come naturale prosecuzione della codifica semantica condotta per la costituzione di *Verbum*, fornendo un doppio livello di lettura che potrà essere applicato all'intero corpus e potrà fornire una categorizzazione dello stesso all'interno dei percorsi tematico-didattici progettati. Tale categorizzazione, infatti, pur partendo dalla codifica della «Rassegna Settimanale», verrà applicata all'interno corpus delle riviste assumendo un importante valore per la ricerca all'interno dei testi a disposizione, e risultando utile per sondare le specificità dei testi e delle tipologie testuali all'interno del corpus.

5. ANALISI DI *VERBUM*

In linea con la tradizione della linguistica dei corpora e della statistica testuale (Biber, 1998; Stefanowitsch, 2020), l'annotazione dei <term> con attributi @type e @subtype consente di superare la semplice analisi delle frequenze assolute, e offre la possibilità di studiare distribuzioni condizionate, ovvero la variazione di *Verbum* in relazione a specifiche variabili contestuali o osservata in funzione di una variabile esterna (genere, autore, periodo, tipo testuale, ecc.). A tal fine, le analisi sono state condotte con NormaTEI,[2] uno strumento progettato per la normalizzazione di codifiche in file XML/TEI. NormaTEI permette di eseguire ricerche complesse attraverso query personalizzate sui documenti (Sichera *et al.*, 2025), identificando occorrenze di termini in relazione ai loro attributi tipologici e semantici, e consente di estrarre in maniera sistematica i dati necessari per la costruzione di

[1] Per dar conto della ricchezza e della peculiarità dei temi affrontati, si è proceduto all'individuazione di cinque classi tematiche – *lavoro*, *condizione femminile*, *ambiente urbano*, *vita rurale*, *verismo* –, all'interno delle quali si ritrovano sottocategorie tematiche spesso intrecciate fra loro. Nel rispetto delle linee guida XML/TEI, la rappresentazione degli elementi tematici è stata codificata nella specifica sezione del <teiHeader>, identificando @type e @subtype rispettivamente per classe tematica e sottocategoria tematica.
[2] https://github.com/pierpaolosichera/NormaTEI (cons. 07/02/2026).

visualizzazioni quantitative. Questa impostazione rende possibile visualizzare non solo la frequenza dei termini, ma anche la loro distribuzione nel tempo, evidenziando tendenze e concentrazioni specifiche. Nel corpus, ad esempio il @subtype="studio_osservazione" è stato impiegato per intercettare un lessico che rinvia alla dimensione metodologica del Verismo: il romanzo come pratica di conoscenza del reale e la postura del narratore come osservatore che registra e restituisce, più che commentare. L'annotazione include non solo i termini-base, ma anche concrezioni sintagmatiche fortemente connotate (ad es. *studio sociale*, *studio psicologico*, *studio obiettivo*, *studio coscienzioso*, *studio biologico-sociale*, e formule legate all'*osservazione dal vero*), così da valorizzare non la sola presenza di parole isolate, ma la ricorrenza di un campo semantico coerente.
La Fig. 5 mostra la distribuzione cronologica dei termini[3] appartenenti al subtype="studio_osservazione" (linea violetta) e le occorrenze delle due parole dominanti più rappresentative del subtype, *studio* e *osservazione*: la loro dinamica parallela (incrementi e flessioni che tendono a corrispondere) suggerisce che, nel discorso verista, "osservare" e "studiare" funzionino come una coppia concettuale strettamente integrata, in cui lo sguardo sul reale viene tradotto in una istanza analitica e "metodica". La concentrazione temporale delle occorrenze (più intensa in alcuni snodi rispetto ad altri) indica dunque non soltanto quando questi termini compaiono, ma altresì quando il Verismo insiste maggiormente sulla propria autodefinizione come poetica fondata sull'oggettività e sull'impersonalità. A livello interpretativo, tale subtype intercetta un lessico che, nel dibattito verista, serve a definire la testualità come pratica di conoscenza: non semplice rappresentazione, ma studio e osservazione del sociale e dell'umano. In questo orizzonte si colloca l'idea del romanzo come «studio sociale» (così Francesco Torraca nella recensione ai *Malavoglia*) e la figura del narratore come soggetto che tenta di ridurre al minimo l'intervento valutativo, pur rimanendo vicino ai «deboli che restano per via», ai vinti presenti e futuri.
La rete semantica dell'osservare (e dell'osservatore) si salda così ai principi di oggettività e impersonalità, mentre le occorrenze di studio si cristallizzano spesso in sintagmi che rafforzano l'idea di metodo e rigore. Questo impianto trova un referente teorico esplicito nella prefazione a *I Malavoglia* del 1881, dove l'artista è descritto come colui che sa guardare il movimento della storia senza sovrapporre un giudizio, conquistando una distanza temporanea dalla "lotta" per restituire la scena con nitidezza: «Chi osserva questo spettacolo non ha il diritto di giudicarlo; è già molto se riesce a trarsi un istante fuori del campo della lotta per studiarla senza passione, e rendere la scena nettamente, coi colori adatti, tale da dare la rappresentazione della realtà com'è stata, o come avrebbe dovuto essere» (Verga 2014:129). Letta insieme alla distribuzione cronologica e alla ricorrenza delle parole dominanti, questa cornice conferma che i picchi della Figura 3 corrispondono ai momenti in cui il Verismo elabora e ribadisce con maggiore forza la propria legittimazione "metodologica": osservare il reale e renderlo come fatto, più che interpretarlo dall'esterno.

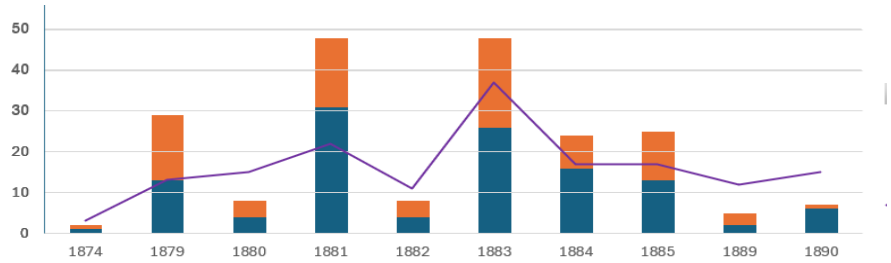


Figura 5. Distribuzione cronologica delle occorrenze di <term> con @subtype "studio_osservazione" (violetto), combinata con quella delle parole dominanti "osservazione" (arancione) e "studio" (blu)

6. CONCLUSIONI

L'analisi di *Verbum* dimostra come l'annotazione semantica e i dati strutturati in XML/TEI possano supportare una lettura integrata del linguaggio della critica verista, combinando frequenza, distribuzione temporale e densità concettuale delle parole dominanti. La metodologia fin qui sperimentata consente di esplorare relazioni tra termini, cluster tematici e pattern sintagmatici, e di confrontare dinamiche lessicali tra autori, generi e periodi diversi.

[3] Si elencano di seguito alcuni dei valori tag appartenenti al suddetto subtype: *reale*, *processo della creazione*, *osservazione del vero*, *finezza d'osservazione*, *fisionomia*, *realtà delle cose*, *studio coscienzioso*, *studio de' costumi*, *osservazione attenta*, *documenti umani*, *analisi lunga e minuziosa*, *esattezza di osservazione*, *realtà naturale*.

L'analisi integrata, oltre a determinare una forte trasversalità dei diversi livelli analizzati (termini, subtype, parole dominanti, categorie), offre gli strumenti utili per la costruzione di un lessico-repertorio esplorabile a differenti livelli di granularità, da quello 'reale' delle occorrenze individuate nei diversi generi testuali, fino a quello 'ideale' disegnato attraverso le categorie concettuali 'apicali'. Si potrebbe dire che, a convergere in questo lessico-repertorio scalabile, siano un approccio top-down (decisione su parole dominanti e categorie) e uno bottom-up (riconduzione delle occorrenze ai rispettivi subtype). In questo senso, fondamentale si rivela il supporto dell'ontologia OntoTerm per la costruzione di queste relazioni multiple. Partendo dai risultati ottenuti categorizzando una porzione del corpus ci si sta muovendo verso l'utilizzo di tecniche di Natural Language Processing (Ježek & Sprugnoli, 2023) che saranno sperimentate sulla restante parte del corpus e nello specifico su articoli tratti da «La Farfalla».
In conclusione, il progetto proposto, facendo interagire modelli di rappresentazione testuale, strumenti computazionali e interpretazione critico-letteraria, dimostra come l'annotazione semantica e l'analisi dei dati possano contribuire non solo alla descrizione di un *corpus*, ma anche alla formulazione di nuove domande di ricerca sulla storia della letteratura, sulla costruzione di un lessico specialistico e sulle dinamiche di circolazione dei concetti estetico-letterari e socio-culturali.

RINGRAZIAMENTI

Progetto PRIN PNRR 2022 COVerLeSS finanziato dall'Unione europea – Next Generation EU, Missione 4 Componente 2 CUP E53D2301888 0001. Si desidera esprimere un sentito ringraziamento a Milena Giuffrida per la consulenza scientifica sui testi di Capuana, a Mariagiusi Polizzi per il lavoro di codifica delle recensioni, e a Giuseppe Arena per il contributo all'ontologia.

BIBLIOGRAFIA

Bruno, D., Canzoneri, G., Cristofaro, S., Di Silvestro, A., Mazzagufo, L., & Spampinato, D. (2025). *Leggere le opere del Verismo attraverso le riviste e i critici del tempo. Edizioni di testi, mappe lessicali e dizionari tematici in un portale* user oriented. Umanistica Digitale, 9(20), 1–30. https://doi.org/10.6092/issn.2532-8816/21302.

Biber, D., Conrad, S. & Reppen, R. (1998). *Corpus Linguistics, Investigating Language Structure and Use*, Cambridge: Cambridge University Press. ISBN 0-521-49957-7.

Ježek, E., & Sprugnoli, R. (2023). *Linguistica computazionale. Introduzione all'analisi automatica dei testi*. Bologna: Il Mulino.

McCrae, J., Bosque-Gil, J., Gracia, J., Buitelaar, P., & Cimiano, P. (2017). *The OntoLex-Lemon Model: Development and Applications*. W3C Community Report.

Miles, A., & Bechhofer, S. (2009). *SKOS Simple Knowledge Organization System Reference*. W3C Recommendation.

Romary, L., & Tasovac, T. (2018). TEI Lex0: *A Target Format for TEI-Encoded Dictionaries*. TEI Consortium.

Sichera, P., Cristofaro, S., D'Agata, C., Del Grosso, A. M., Grasso, M., Mazzagufo, L., & Spampinato, D. (2025). *Streamlining TEI Workflows: Collaborative Editing with NormaTEI*. Atti del convegno "New Territories" - TEI Conference 2025 (TEI2025), Krakow, Poland. Zenodo. https://doi.org/10.5281/zenodo.17171196.

Spadini, E., Tomasi, F.; & Vogeler, G. (2021). *Graph Data-Models and Semantic Web Technologies in Scholarly Digital Editing*. Norderstedt: Herstellung und Verlag.

Stefanowitsch, A. (2020). *Corpus Linguistics: A Guide to the Methodology*. Berlin: Language Science Press. https://doi.org/10.5281/zenodo.3735822.

Tomasi, F. (2016). *Edizioni o archivi digitali? Knowledge sites e apporti disciplinari*. Roma: Sapienza Università Editrice.

Verga, G. (2014). *I Malavoglia*, a cura di Ferruccio Cecco, Novara: Fondazione Verga - Interlinea.

[9] Sessione poster

Questa sezione degli Atti raccoglie i contributi presentati al XV Convegno AIUCD2026 nella modalità poster. Abbiamo scelto di tenere questa sessione in una parte dedicata del volume perché, quest'anno, i poster sono stati 'mappati' al fine organizzare la rete di contenuti rappresentati e poter più agevolmente e con calma visualizzarli e fruirli oltre i tempi delle sessioni a essi dedicati.
Riteniamo che l'esigenza intrinseca delle ricerche della comunità degli umanisti digitali di un continuo scambio e confronto di saperi, di costruzione di reti per la condivisione di contenuti, progetti ed esperienze, trovi nella sintesi grafica dei poster una particolare forma di valorizzazione. La mappa, dunque, non è solo un'interfaccia dinamica di interconnessione, ma apre anche a un terreno interessante e stimolante per sperimentare forme di accesso e di strutturazione di idee e dati complessi, mediati dalla visualizzazione.
Riportiamo quindi il link alla mappa esplorabile dei poster pubblicata nel "Companion di AIUCD 2026" sul sito del Convegno https://www.aiucd2026.unica.it/companion/#poster.
Nelle pagine seguenti si pubblicano i contributi estesi dei contenuti dei poster, indicizzati in ordine alfabetico per autore/autrice.

Archivi letterari nativi digitali: modelli descrittivi ed edizioni sperimentali

Elena Barchielli[1], Simon Willemin[2], Elena Spadini[3].

[1] Universität Bern, Svizzera – elena.barchielli@unibe.ch
[2] Universität Bern, Svizzera - simon.willemin@unibe.ch
[3] Universität Bern, Svizzera - elena.spadini@unibe.ch

ABSTRACT (ITALIANO)

Le opere letterarie contemporanee sono sempre più spesso create direttamente in formato digitale, mentre la filologia, la descrizione archivistica e l'editoria scientifica restano in larga misura ancorate a paradigmi analogici. La crescente presenza di materiali nativamente digitali negli archivi letterari impone quindi una ridefinizione teorica e metodologica delle nozioni di documento, testo, versione e materialità. Il progetto affronta queste sfide considerando i documenti born-digital non solo come oggetti da preservare tecnicamente, ma come fonti filologiche primarie che richiedono nuovi modelli descrittivi, flussi di lavoro editoriali e quadri interpretativi.
Attraverso lo studio di archivi letterari svizzeri italofoni e francofoni, il progetto analizza file conservati su supporti digitali (floppy disk, chiavi USB, hard disk) e documenti acquisiti in formato digitale. Da un lato, mira a sviluppare uno schema di metadati specifico per gli archivi letterari nativamente digitali, basato sugli standard esistenti e su ontologie emergenti, formalizzato in OWL e implementato in RDF, anche mediante l'estrazione automatizzata di dati con strumenti di digital forensics. Dall'altro, propone prototipi di edizioni digitali sperimentali che tengano conto delle specificità delle fonti born-digital, ripensando l'intero workflow filologico: dalla raccolta delle fonti alla collazione, dalla (de)codifica testuale all'apparato critico.
Integrando approcci filologici, archivistici e forensi, il progetto mostra come la descrizione e l'edizione dei materiali digitali siano atti scientifici centrali per la comprensione dei processi di scrittura contemporanei. I risultati attesi sono sia concettuali sia operativi: un quadro teorico per lo studio delle fonti letterarie digitali e strumenti concreti adottabili da archivi, biblioteche e ricercatori. In tal modo, il progetto contribuisce a evitare l'oblio digitale e a garantire il futuro della filologia testuale nell'era del digitale.
Parole chiave: edizione scientifica digitale; archivi nativi digitali; ontologie; RDF; forensica digitale

ABSTRACT (ENGLISH)

Born-digital literary archives: descriptive models and experimental editions.
Contemporary literary works are increasingly created directly in digital format, while philology, archival description, and scientific publishing remain largely anchored to analog paradigms. The growing presence of native digital materials in literary archives therefore requires a theoretical and methodological redefinition of the notions of document, text, version, and materiality. The project addresses these challenges by considering born-digital documents not only as objects to be preserved technically, but as primary philological sources that require new descriptive models, editorial workflows, and interpretative frameworks.
Through the study of Italian- and French-speaking Swiss literary archives, the project analyzes files stored on digital media (floppy disks, USB flash drives, hard disks) and documents acquired in digital format. On the one hand, it aims to develop a specific metadata schema for born-digital literary archives, based on existing standards and emerging ontologies, formalized in OWL and implemented in RDF, including automated data extraction through the use of digital forensics tools. On the other hand, it proposes prototypes of experimental digital editions that take into account the specificities of born-digital sources, rethinking the entire philological workflow: from the collection of sources to collation, from textual decoding/encoding to critical apparatus.
By integrating philological, archival, and forensic approaches, the project shows how the description and editing of digital materials are central scientific acts for understanding contemporary writing processes. The expected results are both conceptual and operational: a theoretical framework for the study of digital literary sources and concrete tools that can be adopted by archives, libraries, and researchers. In this

way, the project contributes to preventing digital oblivion and ensuring the future of textual philology in the digital age.
Keywords: digital scholarly editing; born-digital archives; ontologies; RDF; digital forensics

INTRODUZIONE

Attualmente, gran parte delle opere letterarie viene creata in formato digitale, sebbene i metodi della filologia, della descrizione archivistica e dell'editoria scientifica rimangano prevalentemente radicati in paradigmi analogici. Gli archivi letterari includono sempre più documenti nativi digitali, il che pone nuove sfide per il loro studio (Carbé, 2023). Il progetto Bit Philology posiziona i materiali nati in formato digitale non solo come problemi di conservazione tecnica, ma come oggetti filologici che richiedono nuovi modelli descrittivi, flussi di lavoro editoriali e quadri interpretativi. È opportuno interrogarsi su cosa costituisca un documento in un contesto digitale e quale livello di descrizione sia significativo quando i testi sono instabili, stratificati e dipendenti da software. Il progetto analizza i materiali nativamente digitali di autori italofoni e francofoni conservati negli archivi letterari svizzeri.

1. ÉTAT DE L'ART

Descrizione

La definizione di "documento" nel contesto delle fonti letterarie born-digital pone problemi teorici e pratici rilevanti. Come ha osservato Thorsten Ries, tra i primi studiosi europei a occuparsi in modo sistematico di archivi letterari digitali, nel dominio digitale viene meno l'unità tradizionale tra testo, supporto e tracce del processo di produzione (Ries, 2018). Sebbene la diplomatica digitale richieda, per definizione, una forma fissa e un contenuto stabile per poter parlare di record digitale (Duranti, 2010; Rogers, 2015), nell'uso corrente il termine «documento» è spesso riferito a uno stesso file il cui contenuto, formato e ambiente di conservazione possono mutare nel tempo. Inoltre, la presenza di materiali sommersi o inseriti in fondi tradizionali e collezioni speciali poco documentate rende ormai imprescindibile una riflessione della comunità scientifica volta a individuare buone pratiche (Carbé, 2023).
Esempi di descrizioni solo parzialmente soddisfacenti si riscontrano anche negli Archivi Letterari Svizzeri (SLA): nel fondo di Hans Boesch, ad esempio, è menzionata una scatola che contiene «materiali assortiti, ca. 50 floppy disk», nessuno dei quali è descritto singolarmente; viene menzionato solamente il formato dei file più utilizzato, ovvero WordPerfect (.wps). Nel fondo di Michel Thévoz, invece, sono presenti alcune informazioni aggiuntive sull'hardware e sul software (Microsoft Word), il peso del file, e il numero di pagine, ma esse sono collocate in campi descrittivi generici e non strutturati. Altre informazioni come la data sono estratte automaticamente dai metadati ed è specificato che «non corrispondono necessariamente alla data di redazione del testo».

Edizione

Le ricerche di Ries nell'ambito del progetto «Hard Drive Philology» (2014-2021) e quelle di Crasson et al. nell'ambito del progetto «Derrida Hexadécimal» (2018-2024), nonché i lavori di Mara (2013), Vauthier (2016), Bekius (2021) and Van Hulle (2021) hanno esplorato gli archivi letterari nativamente digitali. Queste ricerche sollevano questioni su cosa debba essere considerato un documento, in un contesto in cui diversi file sono fortemente interdipendenti (Ries, 2018) o su quella che è stata definita «.txtual condition» (Kirschenbaum, 2013). Nell'ambito di queste ricerche, i documenti nativamente digitali sono stati oggetto di un'edizione, in particolare da parte di Ries nell'appendice del suo libro (Ries, 2025) e da parte di Lebrave in un articolo di rivista (Lebrave, 2022). Quest'ultimo esempio include informazioni sulla provenienza e sull'approccio editoriale, una trascrizione annotata, uno screenshot di un'emulazione del documento pubblicato in un contesto simile a quello in cui l'autore lo utilizzava, nonché screenshot di ogni pagina del documento così come appare in un software di elaborazione testi contemporaneo al momento della pubblicazione (Lebrave, 2022). Le edizioni scientifiche che includono fonti born-digital rimangono rare – ad esempio, le ricerche sulla letteratura elettronica (Grigar et al., 2024) – e poche di queste edizioni tengono conto delle specificità dei documenti.
L'edizione *Stoffe* di Friedrich Dürrenmatt (Probst e Weber, 2021) tratta dei testi nativi digitali (*Computertexte*), tuttavia l'utente non ha accesso a tante informazioni sulla fonte rispetto a quante ne ha per i documenti analogici inclusi nella stessa edizione, per i quali vengono fornite informazioni sul supporto, sugli strumenti di scrittura e sull'impaginazione. Per i documenti nativi digitali, non ci sono indicazioni su software o sul formato del documento pubblicato.

2. CORPUS

La selezione degli autori e delle opere da utilizzare come casi di studio nel progetto si basa non solo su criteri estetici di valore letterario, ma anche sulla natura degli archivi di ciascun autore (varietà e quantità di fonti digitali native). Il progetto collabora con archivi a Berna (Archivi letterari svizzeri della Biblioteca nazionale svizzera), a Lugano (Archivio Prezzolini), e a Losanna (Centre des littératures en Suisse romande). Si trovano documenti conservati su floppy disk o dischi rigidi, e documenti acquisiti in formato digitale dai centri archivistici.

3. METODOLOGIA E RISULTATI ATTESI

Il primo passo è stato quello di identificare gli archivi nativamente digitali posseduti da istituzioni svizzere, ottenere i diritti di consultazione e accedervi. Gli scambi con gli archivisti dimostrano come molte questioni rimangano aperte. In primo luogo, le modalità di consultazione degli archivi non sono state sistematicamente oggetto di una regolamentazione specifica e possono variare: alcuni centri di archiviazione propongono di lavorare direttamente sui supporti originali, altri su versioni trasformate (PDF-A) di copie dei file (vedi anche: Pledge e Dickens, 2017). Le prime indagini confermano l'importanza del progetto, poiché molti fondi contengono archivi digitali che non sono ancora stati adeguatamente descritti e sfruttati, e consentono di constatare l'attualità delle questioni sollevate.
È inoltre importante notare come, una volta ottenuta l'autorizzazione dai titolari dei diritti d'autore, l'ottenimento dei testi può risultare in un'operazione tecnicamente complessa a seconda dell'hardware e del software: se il materiale d'archivio include dispositivi fisici e se le questioni etiche e legali lo consentono, è possibile utilizzare un approccio forense digitale per avervi accesso, ma anche per recuperare metadati e oggetti cancellati. La difficoltà di accesso ai documenti è lamentata dagli scrittori stessi riguardo i propri archivi personali. Ne è un esempio l'autore svizzero Vincenzo Todisco (Graubünden Literature Prize 2005; nominato per il Schweizer Buchpreis 2018), il cui archivio va dagli anni '90 ai giorni nostri e include anche supporti obsoleti che lo stesso autore non è più in grado di leggere. Nell'intervista per *I retroscena della scrittura*, ha espresso i suoi timori riguardo all'oblio digitale: "Devo dire però che questa è una cosa che mi ossessiona un po'. Ormai siamo passati al digitale anche nella scrittura. […] I testi che avevo scritto quando c'era ancora il dischetto, oggi rimangono inaccessibili. È una cosa che mi rende triste" (Cordibella, 2022). Per gli oggetti digitali obsoleti, fragili o danneggiati, saranno fondamentali il supporto delle istituzioni archivistiche e la collaborazione con un esperto di informatica forense, il Prof. Dr. Bruce Nikkel dell'Istituto per Cybersecurity e Ingegneria all'Università delle Scienze Applicate di Berna (BFH), disciplina che si occupa del recupero, dell'analisi e dell'interpretazione dei dati contenuti nei dispositivi digitali (Kirschenbaum et al., 2010).
Il poster presenterà i primi risultati dei due sotto progetti relativi alla descrizione e all'edizione dei materiali d'archivio introdotti qui di seguito.

Descrizione
Il primo sotto progetto mira a facilitare la reperibilità e l'esplorazione di archivi nativamente digitali; nella maggior parte dei casi l'accesso e la manipolazione delle fonti è limitato (disponibili solo in loco, soggette a embargo o copyright) e una descrizione dettagliata è l'unico modo per consentirne lo studio preliminare. Le fasi iniziali della ricerca corrispondono all'acquisizione del materiale e all'estrazione dei relativi metadati.
Il corpus a nostra disposizione si caratterizza per un'elevata eterogeneità: comprende computer Apple prodotti nell'arco degli ultimi trent'anni — tra cui un Macintosh del 1994 e un iMac del 2005 — affiancati da modelli più recenti, nonché numerosi floppy disk nei formati da 5,25" e 3,5". Il campione analizzato è costituito da una selezione di estratti relativi ad autori coinvolti nel progetto, i quali hanno operato su differenti tipologie di floppy disk. Le copie forensi sono state ottenute attraverso l'utilizzo combinato di un'unità floppy, del dispositivo Greaseweazle[1] con il relativo software, e di un computer host. Greaseweazle consente di rilevare le transizioni di flusso registrate dalla testina di lettura, salvandole in un file immagine grezzo (.scp). In una fase successiva, tali copie sono state convertite in immagini disco più convenzionali (.img). Il processo ha richiesto l'impiego di diversi strumenti, tra cui HxCFloppyEmulator, software con GUI utilizzabile sia per l'acquisizione che per la conversione, il software

[1] https://github.com/keirf/greaseweazle

Greaseweazle e cpmtools[2], particolarmente efficace nel trattamento dei floppy da 5,25". Una criticità rilevante nella fase di conversione è stata rappresentata dalla corretta identificazione dei supporti, poiché il corpus include una combinazione eterogenea di floppy Single- e Double-Sided, nonché Single-, Double- e High-Density, formattati mediante software differenti.
Al fine di garantire un elevato livello di completezza nell'estrazione dei metadati, sono stati impiegati diversi strumenti, tra cui DROID (Digital Record Objec Identification)[3], sviluppato dai The National Archives del Regno Unito, The Sleuth Kit (TSK)[4], la sua interfaccia grafica Autopsy[5], ExifTool[6] e comandi nativi del sistema operativo. I risultati evidenziano come Autopsy consenta di raccogliere la maggiore quantità di informazioni relative ai supporti, includendo anche file eliminati o temporanei. Le informazioni estratte da DROID sono altrettanto interessanti, il programma riesce infatti a riconoscere molti formati grazie alla sua integrazione con il registro dei formati PRONOM, che ne potenzia le capacità di identificazione. La selezione mirata di specifici campi, variabili a seconda degli strumenti utilizzati, ha permesso di ottenere una visione complessiva e articolata dei metadati associati alle risorse.
Questa fase di esplorazione e selezione dei metadati costituisce il fondamento per quella successiva, in cui ci si interroga sulle modalità di presentazione al pubblico. Una prima esplorazione per la rappresentazione del corpus di materiali born-digital secondo le specifiche IIIF, in particolare usando IIIF-Archive, è in corso.

Edizione
Il secondo sotto progetto mira a sviluppare prototipi di edizioni digitali che tengano conto della natura delle fonti edite, per cui l'intero workflow filologico deve essere ripensato per le fonti born-digital. A tal fine, verrà affrontata sistematicamente la decodifica e la codifica dei caratteri, le funzionalità che influenzano o vincolano la scrittura degli autori, nonché il trattamento di elementi testuali specificamente digitali come i collegamenti ipertestuali integrati.
Diversi floppy disk consultati contengono documenti che si ritiene presentino un interesse particolare per quanto riguarda il lavoro editoriale che può accompagnare la creazione di un testo conservato in un archivio digitale. Da un'analisi preliminare di un floppy disk è emerso che i file sono stati scritti con un editor di testo Brother che ha una stampante 'a margherita' (*Daisy Wheel*) integrata. Questo hardware risale agli anni '80 o '90. Il tentativo di leggere i file con software contemporanei viene interrotto da notifiche che indicano che il file è danneggiato e non può essere riparato (LibreOffice Writer), oppure si apre un documento che presenta problemi di codifica (Blocco note, MS Word), in particolare per i caratteri diacritici. Per avere una visione più completa è utile analizzare i documenti al livello dei byte. Così facendo, è possibile notare che i file non sono codificati secondo gli standard che si sono imposti a partire dagli anni 2000: la codifica dei caratteri è probabilmente quella originale, ma potrebbe anche essere il risultato di un fenomeno di corruzione dovuto a conversioni. Il ricorso alla forma esadecimale può consentire di dedurre il significato di alcune stringhe di byte e di ricostruire il file così come probabilmente appariva all'autore nel contesto in cui era verosimilmente utilizzato. Altri indizi consentono di ricostruire e stabilire il testo, come la documentazione relativa agli strumenti/hardware associati ai file creati dall'autore o le versioni stampate dei file nativamente digitali. La combinazione di queste informazioni consente di effettuare scelte editoriali consapevoli in vista di un'edizione critica. L'edizione critica realizzata nell'ambito del progetto consentirà agli utenti di accedere a differenti versioni e fasi del testo, dal flusso di byte al testo definitivo, raccolte in file XML preparati sfruttando le possibilità offerte dalla TEI. La documentazione che accompagna la preparazione di questi file sarà utile non solo a chi è interessato al testo pubblicato, ma anche a chi dovrà occuparsi dell'edizione di testi originariamente digitali che presentano sfide simili.

[2] https://github.com/lipro-cpm4l/cpmtools
[3] https://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/droid/
[4] https://www.sleuthkit.org/sleuthkit/
[5] https://www.sleuthkit.org/autopsy/
[6] https://exiftool.org

4. CONCLUSIONI

I risultati attesi dal progetto sono sia metodologici che pratici. Da un lato, il progetto contribuisce a creare un quadro concettuale per comprendere le fonti letterarie nate in formato digitale come oggetti filologici, mettendo in discussione le nozioni tradizionali di documenti, versioni e materialità. Dall'altro lato, fornisce strumenti e modelli concreti – schemi di metadati, cataloghi digitali e prototipi di edizioni scientifiche – che possono essere adottati o adattati da archivi, biblioteche e ricercatori che lavorano con archivi letterari contemporanei.
Mettendo in primo piano la descrizione come atto accademico e collegandola direttamente alla pratica editoriale, il progetto vuole contribuire al futuro della filologia testuale, un futuro in cui ci si confronterà sempre più con materiali nati in formato digitale, prima che diventino illeggibili, invisibili o irrimediabilmente persi.

BIBLIOGRAFIA

Bekius, L. (2021). The Reconstruction of the Author's Movement Through the Text, or How to Encode Keystroke Logged Writing Processes in TEI-XML. Variants. The Journal of the European Society for Textual Scholarship, 15–16, 3-43. https://doi.org/10.4000/variants.1245

Carbé, E. (2023). Digitale d'autore. Macchine, archivi, letterature. Firenze University Press.

Cohen, F. B. (2015). Digital diplomatics and forensics: Going forward on a global basis. Records Management Journal, 25(1), 21–44. https://doi.org/10.1108/RMJ-03-2014-0016

Cordibella, G. (2022). I retroscena della scrittura. Come lavorano le scrittrici e gli scrittori in lingua italiana della Svizzera. Armando Dadò Editore.

Crasson, A., Lebrave, J.-L., Pedrazzi, J., & Alonso, L. (2022). Le toucher Touch/to touch him. Genesis. Manuscrits – Recherche – Invention, 55, 121-136. https://doi.org/10.4000/genesis.7747

Duranti, L. (2010). Diplomatics. In Digital Forensics and Born-Digital Content in Cultural Heritage Collections • CLIR (No. 149; CLIR Publication). Kirschenbaum, M. G., Ovenden, R., & Redwine, G. (Eds.). 10–11. Washington, D.C.: Council on Library and Information Resources. https://www.clir.org/pubs/reports/pub149/

Grigar, D., & Pisarski, M. (2024). The Challenges of Born-Digital Fiction: Editions, Translations, and Emulations. In The Cambridge Companion to Twenty-First Century American Fiction. Cambridge University Press. https://doi.org/10.1017/9781009181488

Kirschenbaum, M. G., Ovenden, R., & Redwine, G. (2010). Digital Forensics and Born-Digital Content in Cultural Heritage Collections • CLIR (No. 149; CLIR Publication). Washington, D.C.: Council on Library and Information Resources. https://www.clir.org/pubs/reports/pub149/

Kirschenbaum, M.G. (2013). The .txtual Condition: Digital Humanities, Born-Digital Archives, and the Future Literary. Digit. Humanit. Q., 7. https://doi.org/10.5749/minnesota/9780816680030.003.0003

Lebrave, J.-L. (2011). Computer forensics: La critique génétique et l'écriture numérique. Genesis. Manuscrits – Recherche – Invention, 33, 137-147. https://doi.org/10.4000/genesis.633

Lebrave, J.-L. (2022). « Hymne à Khôra » ou le fichier Diosperi. Genesis. Manuscrits – Recherche – Invention, 55, 147-164. https://doi.org/10.4000/genesis.7589

Mara, M. O. (2013). Nuala O'Faolain: New departures in textual and genetic criticism. Irish Studies Review, 21(3), 342–352. https://doi.org/10.1080/09670882.2013.808873

Pledge, J., & Dickens, E. (2018). Process and progress: Working with born-digital material in the Wendy Cope Archive at the British Library. Archives and Manuscripts, 46(1), 59–69. https://doi.org/10.1080/01576895.2017.1408024

Probst, R. & Weber, U. (Eds.). (2021). Friedrich Dürrenmatt. Das Stoffe-Projekt. https://www.fd-stoffe-online.ch.

Ries, T. (2018). The rationale of the born-digital dossier génétique: Digital forensics and the writing process: With examples from the Thomas Kling Archive. Digital Scholarship in the Humanities, 33(2), 391–424. https://doi.org/10.1093/llc/fqx049

Ries, T. (2025). Hybride Textgenese und digitale Forensik. Universitätsverlag Winter.

Rogers, C. (2015). Diplomatics of born digital documents – considering documentary form in a digital environment. Records Management Journal, 25(1), 6–20. https://doi.org/10.1108/RMJ-03-2014-0021

Van Hulle, D. (2021). Dynamic Facsimiles: Note on the Transcription of Born-Digital Works for Genetic Criticism. Variants. The Journal of the European Society for Textual Scholarship, 15–16, 231–241. https://doi.org/10.4000/variants.1450

Vauthier, B. (2016). Genetic Criticism Put to the Test by Digital Technology: Sounding out the (mainly) Digital Genetic File of El Dorado by Robert Juan-Cantavella. Variants. The Journal of the European Society for Textual Scholarship, 12–13, 163–186. https://doi.org/10.4000/variants.353

Lexicon Philosophicum: una rivista digitale per la condivisione della conoscenza umanistica

Pamela Barletta[1], Maria Cristina Dalfino[2], Pietro Restaneo[3*1]

[1]CNR-Istituto per il Lessico Intellettuale Europeo e Storia delle Idee (CNR -ILIESI), Italia – pamela.barletta@cnr.it,
[2]CNR-Istituto per il Lessico Intellettuale Europeo e Storia delle Idee (CNR -ILIESI), mariacristina.dalfino@cnr.it,
[3]CNR-Istituto per il Lessico Intellettuale Europeo e Storia delle Idee (CNR -ILIESI), pietro.restaneo@cnr.it

ABSTRACT (ITALIANO)

Lo scopo di questo contributo è quello di offrire alcuni spunti di riflessione nell'ambito dell'editoria accademica istituzionale in Diamond Open Access, basati sull'esperienza nella gestione della rivista *Lexicon Philosophicum: International Journal for the History of Texts and Ideas*. In questo contributo si cercherà in particolare di mettere in luce il ruolo del digitale e del public engagement nel caso specifico di *Lexicon Philosophicum*, non solo per la ricerca umanistica in senso stretto, ma in generale per le Digital Humanities. Si illustrerà quanto, nel processo evolutivo della Rivista, si è fatto per tenersi al passo con le pratiche di scienza aperta, quali le criticità affrontate, e ancora da affrontare, e quali gli obiettivi da perseguire nel public engagement.
Parole chiave: rivista scientifica; scienza aperta; condivisione del sapere; strumenti digitali; public engagement

ABSTRACT (ENGLISH)

Lexicon Philosophicum: A Digital Journal for Sharing Humanistic Knowledge.
The purpose of this paper is to offer some suggestions in the field of Institutional Publishing in Diamond Open Access, based on the experience of managing the journal *Lexicon Philosophicum: International Journal for the History of Texts and Ideas*. In this paper, we will seek to highlight the role of digital technologies and public engagement in the specific case of *Lexicon Philosophicum*, not only for humanities research strictly speaking, but also for the digital humanities in general. We will illustrate how we worked, in the evolution of the Journal, to keep up with the practices of open science, which critical issues we encountered and need to be addressed, and which objectives we pursue in the field of public engagement.
Keywords: scientific journal; open science; knowledge sharing; digital tools; public engagement

1. INTRODUZIONE E STORIA DELLA RIVISTA

La rivista *Lexicon Philosophicum*[2] nasce, in occasione della partecipazione dell'ILIESI (Istituto per il Lessico Intellettuale Europeo e Storia delle Idee) al Progetto europeo "Agora: Scholarly Open Access Research in European Philosophy" (2011-2013), dalla trasformazione della omonima collana di volumi a stampa, fondata nel 1985 presso il medesimo Istituto; il primo fascicolo è pubblicato nel 2013.
Lexicon Philosophicum è una rivista digitale, di periodicità annuale, internazionale e a carattere interdisciplinare; la Rivista attua una revisione paritaria, adotta un modello di Diamond Open Access (DOA), ed è pubblicata sulla piattaforma OJS (Open Journal System) dell'Istituto ILIESI del CNR.
Semanticamente collegata agli archivi testuali del portale Daphnet[3] dello stesso ILIESI, la Rivista offre al suo pubblico, fin dal primo fascicolo, strumenti digitali preziosi come l'interconnessione e l'interoperabilità con le banche dati d'Istituto, tramite l'accesso diretto ai testi e alle edizioni di riferimento.
Il suo obiettivo è contribuire con saggi critici, articoli brevi, edizioni di testi e resoconti bibliografici critici nell'ambito del lessico intellettuale; particolare attenzione è dedicata alla terminologia tecnica e

[1]*Gli Autori e le autrici sono riportati in ordine alfabetico.
Tutti gli autori hanno contribuito egualmente e sono da considerarsi primo autore alla pari.
[2] https://lexicon.cnr.it/ojs/index.php/LP (cons. 17/04/2026)
[3] https://scholarlysource.lexicon.cnr.it/ojs/index.php/DDL (cons. 18/01/2026)

specialistica, ai lessici e alle traduzioni. Una sezione specifica è aperta a contributi nel campo delle Digital Humanities, come edizioni accademiche digitali, creazione di corpora, ricerca sul web semantico.
L'Istituto ha continuato a pubblicare la Rivista online, anche dopo la conclusione del progetto "Agora", con l'obiettivo di fornire contenuti accademici di alta qualità e ad accesso aperto. L'ILIESI gestisce infatti, oltre a *Lexicon Philosophicum*, anche la piattaforma OJS della rivista *Aristotelica*, e una serie di collane di volumi in DOA dal titolo "ILIESI digitale". Dalla sua genesi, quindi, *Lexicon Philosophicum* persegue i principi della scienza aperta e ne affronta le sfide.
A partire dal 2021 si è proceduto a un piano di rilancio e rinnovamento della Rivista che ha coinciso anche con la partecipazione dell'Istituto al progetto H2IOSC ("Humanities and Cultural Heritage Italian Open Science Cloud"),[4] occasione favorevole a nuove conoscenze, contatti e opportunità nel settore delle Digital Humanities.
Il piano di rilancio si è posto due obiettivi in particolare: migliorare la visibilità della Rivista e trasformarla in un "laboratorio" di pratiche editoriali utili a coniugare l'elevata qualità scientifica con i principi dell'open access e del public engagement.

2. *LEXICON PHILOSOPHICUM* NELL'ECOSISTEMA DOA

Come noto, i criteri e i requisiti DOA richiesti da varie organizzazioni ed enti, anche a livello europeo, sono stati ampliati e definiti con più precisione (Ancion et al., 2022; Rico-Castro et al., 2024), soprattutto per far fronte alla richiesta di trasparenza, rigore scientifico ed etico. All'interno di questo "ecosistema" di repository, enti, associazioni e certificazioni si inserisce quindi l'intera ristrutturazione della Rivista.
Nel 2021 la natura DOA della Rivista consisteva principalmente in tre fattori: assenza di costi a carico degli autori e dei lettori, licenza open CC (Creative Commons), e gestione da parte dei membri stessi della comunità di riferimento ("community-led"). Questi fattori, che costituiscono certamente l'essenza del DOA, da soli non erano più sufficienti per far fronte alle richieste degli enti interessati all'open access, quantomeno nell'ottica di aderire agli standard più elevati.
In che cosa è consistito, nella pratica, l'ingresso della Rivista nell'ecosistema DOA? Il lavoro si è svolto principalmente su tre livelli strettamente interconnessi: adesione a repository open access, in primo luogo DOAJ (Directory of Open Access Journals); conseguente adeguamento delle politiche editoriali e delle informazioni sulla rivista; formazione del personale.
Esempi di tale adeguamento sono il processo di peer review e le politiche di archiviazione. In previsione della valutazione della Rivista per l'indicizzazione in DOAJ (e altri repository), sono state aggiornate le informazioni presenti sul sito in merito alla peer review, che prima si limitavano a descrivere sommariamente il processo adottato (double blind). Questa revisione ha fatto emergere alcune criticità e ambiguità nella gestione della valutazione. Si è quindi elaborato un nuovo flusso, più dettagliato e rigido ma anche più trasparente. Tale approccio ha ampliato l'orizzonte di valutazione scientifica dei contributi proposti in un'ottica di public engagement, concepito come partecipazione attiva della comunità accademica e non solo.
Un altro parametro di DOAJ nel valutare le riviste consiste nella politica di archiviazione, che ha portato per *Lexicon Philosophicum* alla scelta dell'archiviazione tramite PKP PN. Tutto questo è avvenuto grazie alla partecipazione dell'Istituto al progetto H2IOSC, tramite il nodo italiano dell'infrastruttura OPERAS, che ha offerto inoltre numerose opportunità formative, strumenti e risorse digitali, e un consistente ampliamento della rete di conoscenze nell'ambito del DOA.
Un ulteriore passo in avanti è rappresentato dal completo ripensamento delle politiche della Rivista rispetto alla scelta del tipo di licenza e del copyright, nonché l'elaborazione di un nuovo codice etico.
La Rivista, precedentemente, adottava una licenza CC molto restrittiva, ovvero BY-NC-SA (Attribuzione-NonCommerciale-CondividiAlloStessoModo).[5] Al fine dell'indicizzazione in DOAJ si è analizzata l'interazione fra licenza d'uso e diritto d'autore e sono state ripensate le politiche della rivista sull'accesso aperto. In concomitanza con un più generale ripensamento delle politiche sull'open access per tutte le iniziative editoriali dell'Istituto, la Rivista ha adottato una licenza meno restrittiva, ovvero CC BY-SA (Attribuzione-CondividiAlloStessoModo).[6] Tale licenza regola le modalità del riuso da parte di terzi a beneficio degli stessi autori, favorendo la diffusione dei prodotti della ricerca; basti solo pensare alla

[4] https://www.h2iosc.cnr.it/ (cons. 22/04/2026)
[5] https://creativecommons.org/licenses/by-nc-sa/4.0/ (cons. 30/01/2026)
[6] https://creativecommons.org/licenses/by-sa/4.0/deed.it (cons. 18/04/2026)

possibilità fornita all'autore di condividere la propria opera anche solo in strumenti di analisi e consultazione. Tutto ciò è utile agli autori per condividere i risultati delle ricerche, ma anche un traguardo raggiunto dalla Rivista nell'ottica dei principi FAIR della scienza aperta.
L'adeguamento ai requisiti DOAJ ha comportato la riformulazione del codice etico, che in precedenza si richiamava (senza specificarle) alle linee guida del Committee on Publication Ethics (COPE, 2022). Il codice etico è stato rivisto anche in virtù della nuova strutturazione del COPE in *topics* e declinato in base alle esigenze della Rivista, oltre che integrato ai suoi flussi di lavoro. Sono state infine definite policies specifiche per l'IA.
Sempre in linea con i principi FAIR della scienza aperta si è prestata una particolare attenzione all'arricchimento dei metadati e degli oggetti digitali.

3. LE CRITICITÀ

Nella transizione all'interno di un ecosistema DOA, e quindi nel rivedere le pratiche editoriali precedenti, il team editoriale di *Lexicon Philosophicum* si è trovato ad affrontare una serie di criticità, prima fra tutte la sostenibilità.
L'adozione di pratiche orientate alla trasparenza, alla documentazione dei processi e al public engagement, ha comportato un incremento del lavoro editoriale, un impegno non immediatamente visibile, ma essenziale per garantire la qualità e la coerenza ai principi della scienza aperta. Per un modello editoriale non commerciale basato su finanziamenti pubblici, infatti, la risorsa più importante è il tempo, inteso come ore-persona. Si tratta della forma fondamentale con cui i fondi pubblici si trasformano in pubblicazione. Per questo motivo, la capacità della Rivista di operare – a prescindere dalla componente scientifica – dipende dal tempo di lavoro che il personale dell'Istituto dedica ad essa. La stessa transizione a un sistema DOA richiede al personale di acquisire conoscenze e competenze tramite specifica formazione. Le procedure più rigorose in tutte le fasi del *workflow* editoriale, le operazioni di arricchimento dei metadati, il confronto continuo con l'universo DOA per mantenere elevati gli standard, hanno senza dubbio aumentato il lavoro ordinario per la Rivista, che solo di recente ha potuto aggiungere un nuovo componente al team editoriale.[7]
L'introduzione di nuove pratiche editoriali ha inoltre evidenziato una tensione tra nuovi bisogni organizzativi e abitudini consolidate all'interno della comunità di riferimento. Il modello della rivista commerciale, basato su Call for Papers generali, non è praticabile per la maggior parte dell'editoria istituzionale, spesso rivolta a comunità ristrette e specialistiche. Per questo motivo si è adottata, oltre alla modalità di pubblicazione ordinaria, una modalità per numeri "monografici" attraverso Call for Papers specifiche; in entrambi i casi si assicura un'elevata qualità scientifica, attraverso una peer review estremamente rigorosa e migliorativa, e processi di pubblicazione e di valutazione trasparenti, evitando le cosiddette pratiche "endogamiche" (cf. Nazarovets & Nazarovets, 2026).
Alle criticità sopra esposte si aggiungono naturalmente altre diverse difficoltà cui si è cercato di porre rimedio. Difficoltà ordinarie di gestione di una Rivista, si pensi ad esempio a quelle riscontrate da autori e revisori poco esperti nell'utilizzo delle funzionalità della piattaforma OJS: si sono organizzati seminari d'Istituto dedicati a questo tema specifico. Difficoltà nel rispettare la periodicità e la puntualità di pubblicazione dei fascicoli: si sono tenute periodiche riunioni di redazione e fissate rigide scalette di lavoro. Scarsa attrattività della Rivista ai suoi esordi: si è profuso molto impegno per la diffusione di *Lexicon Philosophicum* nel contesto internazionale, ad esempio tramite piattaforme social e la richiesta di indicizzazione nei database specifici del settore. Traguardi significativi di questo complesso percorso sono stati l'inserimento di *Lexicon Philosophicum* in SCOPUS e la classificazione ANVUR in classe A (11/C5 History of philosophy).

4. GLI OBIETTIVI DA PERSEGUIRE NEL PUBLIC ENGAGEMENT

Alla luce di quanto fino ad ora descritto e delle criticità emerse, obiettivo della Rivista è coniugare la qualità scientifica, l'apertura dei processi e la sostenibilità del lavoro editoriale. In questa prospettiva è

[7] Alcuni studi hanno sottolineato l'inefficienza dell'attuale configurazione dell'editoria istituzionale OA (Frantsvåg, 2010). In particolare, si può menzionare la moltiplicazione di alcune funzioni che potrebbero facilmente essere centralizzate (come ad es. l'hosting OJS). Questi problemi richiederebbero tuttavia soluzioni sovra-istituzionali – nel nostro caso a livello di CNR piuttosto che di singolo istituto – e quindi questa non può essere la sede in cui poterli affrontare.

fondamentale che tutti gli attori, siano essi autori o editori, contribuiscano all'impegno della diffusione pubblica del sapere nelle Digital Humanities.
Nel quadro di un più ampio processo di rinnovamento editoriale, la Rivista intende rafforzare in modo sistematico il valore del public engagement, concepito come partecipazione attiva e dialogica nella produzione e circolazione della conoscenza, attraverso forme di interazione all'interno delle varie comunità di riferimento. *Lexicon Philosophicum* intende dare spazio a contributi che documentino attività partecipative, quali laboratori, organizzazione di eventi aperti — seminari online, forum, workshop — collegati ai numeri tematici, allo scopo di favorire momenti di confronto diretto tra autori e lettori. Questa integrazione tra pubblicazione ed eventi costituisce un tassello fondamentale per configurare la Rivista come infrastruttura di relazione e scambio. *Lexicon Philosophicum* ha infatti promosso alcune iniziative di public engagement attorno ai temi della ricerca umanistica e delle Digital Humanities. Di seguito vogliamo offrire due esempi di tali attività.
Il primo è consistito in un ciclo di incontri dal titolo "Metamorfosi", di cui il primo appuntamento si è svolto il 27 ottobre 2023,[8] e il secondo il 9 maggio 2025.[9] I seminari hanno visto la Redazione della Rivista come promotrice dell'iniziativa, insieme al Gruppo di ricerca Harpocrates (www.harpocrates.eu). I seminari sono stati un'occasione di incontro multidisciplinare fra docenti, ricercatori e ricercatrici che hanno affrontato, da vari punti di vista (letterario, filosofico, storico, digitale, etc.) il tema della "metamorfosi". I proceedings dei seminari del primo incontro sono stati pubblicati nel numero 11 (2023-2024) dalla Rivista in una sezione dedicata.
La seconda iniziativa di engagement prevede la nascita di una nuova sezione della Rivista, che raccolga contributi, sotto forma di note, di giovani ricercatori e ricercatrici, anche attraverso il coinvolgimento di società e istituzioni filosofiche. I prossimi fascicoli dell'annata 2026 di *Lexicon Philosophicum* vedranno la pubblicazione di alcuni contributi selezionati nell'ambito dell'"International Colloquium on Leibniz and Modern Philosophy" (tenutosi a Roma dal 1 al 3 luglio 2025). La selezione è stata compiuta in collaborazione con l'associazione Sodalitas Leibnitiana, coorganizzatrice del Colloquio. Per il 2027 sono in programma brevi note sul pensiero di Pierre Gassendi e, nel fascicolo successivo, nuove ricerche nella storia delle idee linguistiche. Per questi papers si prevede dal 2027 una forma di open peer review che potrebbe scaturire da un seminario/laboratorio dedicato. Si sta infatti prendendo in considerazione la possibilità di adottare forme di open peer review (Ross-Hellauer, 2017), per sostituire o affiancare l'attuale double blind peer review.
Saranno inoltre rafforzate le politiche di equità e inclusione, attraverso l'attenzione al gender balance e alla definizione di linee guida per un linguaggio inclusivo, oltre al coinvolgimento di revisori e autori provenienti da contesti diversi. Queste misure intendono promuovere un ambiente editoriale pluralista, in grado di accogliere e valorizzare prospettive differenziate.
Le azioni delineate contribuiscono a rendere *Lexicon Philosophicum* uno spazio partecipato e aperto, dimostrando come il Diamond Open Access, inteso come ecosistema di pratiche e non solo come modello di accesso, possa diventare un efficace strumento di public engagement, ma anche un'occasione per ripensare il ruolo, la sostenibilità e la responsabilità dell'editoria istituzionale nelle Scienze Umane.

BIBLIOGRAFIA

Ancion, Z., Borrell-Damián, L., Mounier, P., Rooryck, J., & Saenen, B. (2022). Action Plan for Diamond Open Access. Zenodo. https://doi.org/10.5281/zenodo.6282403

COPE DOAJ OASPA WAME (2022). Principles of Transparency and Best Practice in Scholarly Publishing — English. https://doi.org/10.24318/cope.2019.1.12

Frantsvåg, J. E. (2010). The Size Distribution of Open Access Publishers: A Problem for Open Access? First Monday, 15(12). https://doi.org/10.5210/fm.v15i12.3208

Nazarovets, M. & Nazarovets, S. (2026). Prevention of Endogamy in the Editorial Boards of University Journals. J Acad Ethics, 24(1). https://doi.org/10.1007/s10805-025-09687-z

[8] https://www.iliesi.cnr.it/Metamorfosi.pdf (cons. 22/04/2026)
[9] https://www.iliesi.cnr.it/Metamorfosi2.pdf (cons. 22/04/2026)

Rico-Castro, P., Rooryck, J., Melinščak Zlodi, I., Stojanovski, J., Ševkušić, M., & Armengou, C. (2024). D3.2 Extensible Quality Standard in Institutional Publishing (EQSIP) V2.0 for Diamond Open Access (1.0). Zenodo. https://doi.org/10.5281/zenodo.10726732

Ross-Hellauer T. (2017). What is Open Peer Review? A Systematic Review. F1000Res, 6(588). https://f1000research.com/articles/6-588

Intelligenza artificiale, reference digitale e accessibilità: l'esperienza di Alphabetica

Flavia Bruni[1] – Elisabetta Castro[2]
[1]ICCU, Italia – flavia.bruni@cultura.gov.it
[2]ICCU, Italia – elisabetta.castro@cultura.gov.it

ABSTRACT (ITALIANO)

I primi tentativi di applicazione dell'intelligenza artificiale ai cataloghi online si sono concentrati sulla catalogazione piuttosto che sul reference e sulla fruizione. Alphabetica, il portale lanciato dall'ICCU nel 2021 per offrire un'esperienza di ricerca innovativa, propone invece un 'reference occultato', ovvero suggerimenti e percorsi di navigazione basati su criteri biblioteconomici e non sulla profilazione degli utenti tipica dei motori di ricerca commerciali. Questo contributo illustra sinteticamente le scelte adottate in fase di progettazione del portale e gli sviluppi tuttora in corso, in particolare per quanto riguarda la reperibilità delle risorse accessibili (Braille, audiolibri, CAA, alta leggibilità, LIA) anche in un nuovo catalogo afferente in corso di sviluppo grazie a un recente finanziamento di cui l'ICCU è risultato vincitore.
Parole chiave: cataloghi online; reference digitale; intelligenza artificiale; accessibilità; chatbot

ABSTRACT (ENGLISH)

Artificial Intelligence, Digital Reference and Accessibility: the case of Alphabetica
The first attempts to apply artificial intelligence to online catalogues have focused on cataloguing rather than on reference and user access. Alphabetica, the portal launched by ICCU in 2021 to offer an innovative search experience, proposes instead a 'hidden reference', namely suggestions and navigation paths based on library science criteria and not on user profiling, which is typical of commercial search engines. This contribution briefly explains the choices made during the design phase of the portal and the developments still underway, especially those on the findability of accessible resources (Braille, audiobooks, AAC, high readability, LIA) also in a new linked OPAC currently being developed thanks to a grant recently obtained by the ICCU.
Keywords: online catalogues; digital reference; artificial intelligence; accessibility; chatbots

1. INTRODUZIONE

Questo contributo illustra l'approccio adottato da Alphabetica, il portale lanciato dall'Istituto Centrale per il Catalogo Unico delle Biblioteche Italiane (ICCU) nel 2021 e attualmente in corso di ulteriori sviluppi. Alphabetica è il frutto del tentativo di creare uno strumento nuovo, alternativo sia rispetto agli OPAC più tradizionali sia ai più recenti discovery (Bardelli & Verga, 2019; Fusetti & Tagliani, 2019), pensato per utenti inesperti. I risultati e i percorsi di navigazione proposti sono inconsueti ma affidabili in quanto strutturati in base alle competenze degli specialisti dell'informazione piuttosto che sugli algoritmi dei motori di ricerca generalisti di natura commerciale. Grazie a un recente finanziamento, l'ICCU è inoltre attualmente coinvolto nella realizzazione di un nuovo catalogo per facilitare la reperibilità delle risorse accessibili.

2. L'USO DELL'IA NEI CATALOGHI TRA CATALOGAZIONE E FRUIZIONE

Nonostante un ruolo di primo piano nell'adozione di tecnologie innovative negli scorsi decenni, le biblioteche sembrano essere ancora in cerca di una propria via per sfruttare al meglio le potenzialità dell'intelligenza artificiale (IA). Piuttosto che nell'assistenza agli utenti nell'utilizzo dei cataloghi online (OPAC) o nella prospettazione dei risultati, l'IA sembra aver destato maggiori aspettative nell'ambito della catalogazione (Roncaglia, 2025; Roncaglia, 2026). Nell'ambito di un progetto finanziato dal Digital Europe Programme (DIGITAL) dell'UE, tra il 2023 e il 2025 è stata realizzata la piattaforma AI4Culture, che mette a disposizione una serie di strumenti basati sull'IA per il patrimonio culturale, come il

riconoscimento di testi multilingue all'interno di documenti digitalizzati, la creazione e validazione di sottotitoli multilingue, l'arricchimento e la traduzione automatica dei metadati.[1] Particolarmente promettente appare infatti la possibilità di estrarre e arricchire automaticamente i metadati semantici, in particolare delle risorse digitali, con un'incidenza di errore limitata e passibile di raffinamento progressivo: algoritmi di natural language processing (NLP) possono essere impiegati per analizzare il contenuto delle risorse e ricavarne la classificazione, richiedendo solo un intervento di controllo e validazione finale da parte del bibliotecario. Diverse biblioteche nazionali europee stanno da tempo mettendo a punto nuovi strumenti per l'indicizzazione semantica automatica o semiautomatica delle risorse (Manzoni, 2023); la Biblioteca nazionale centrale di Firenze ha intrapreso nel 2022 l'addestramento del programma Annif della Biblioteca nazionale di Finlandia in italiano, mentre all'anno seguente si data l'avvio del progetto V.I.CO. – Vocabolario delle Identità Controllate, per la creazione di un vocabolario italiano controllato e partecipato dei termini relativi a entità (concetti, persone ed enti, opere, luoghi e altro) per l'indicizzazione descrittiva e semantica in contesto linked open data (Storti, 2024). Un progetto pionieristico della Biblioteca Vaticana ha messo alla prova l'IA nel riconoscimento dei contenuti iconografici di alcuni manoscritti vaticani, le cui miniature sono state classificate semanticamente tramite annotazioni automatiche in IIIF per realizzare percorsi tematici.[2] Un'altra possibile applicazione dall'IA ai cataloghi è connessa agli ottimi risultati raggiunti nella traduzione automatica, che potrebbero consentire di implementare l'interfaccia di ricerca rendendola consultabile in lingue e alfabeti che non siano ancora non integrati strutturalmente nelle piattaforme.
Poco esplorato rimane invece l'uso dell'IA in altre attività della professione bibliotecaria, in particolare per quello che riguarda il reference ed eventuali nuove funzioni per aiutare gli utenti nell'esplorazione dei cataloghi. Un primo tentativo di affiancare, se non di sostituire, le funzionalità dei cataloghi tramite l'applicazione di tecnologie recenti è rappresentato dai chatbot. Un 'bot' (forma derivata e abbreviata da 'robot') è un programma in grado di compiere velocemente azioni ripetitive; un chatbot, nello specifico, come suggerisce il nome è un programma che può rispondere alle domande poste da un interlocutore tramite una chat collocata in sezioni appositamente ricavate all'interno di una pagina web (Fontanin, 2022). Questa funzionalità, più spesso integrata in siti di natura commerciale o di fornitura di servizi, ha visto una sperimentazione anche nei siti web di musei e biblioteche (Castellucci & Gomelino, 2021). Il primo chatbot progettato per fornire informazioni su una biblioteca e le sue risorse è stato Kingbot, sviluppato a partire dal 2018 dal gruppo per l'intelligenza artificiale della San Jose State University. La prima versione non era basata su IA generativa ma su algoritmi di NLP e machine learning; nel 2024 è stata rilasciata la versione sperimentale (beta) Kingbot GPT, realizzata tramite python, con l'intento di rendere l'esperienza utente più interattiva.[3] Sul sito della biblioteca è chiaramente specificato che Kingbot GPT non è uno strumento di ricerca e che non sostituisce le competenze dei bibliotecari, potendo solo gestire domande semplici e fornire informazioni di base, per esempio sugli orari d'apertura della biblioteca e sulle risorse quando il personale non possa rispondere.[4] La chat è concepita come un canale d'accesso aggiuntivo a pagine presenti nel sito della biblioteca o a record bibliografici del catalogo: alla fine della conversazione, il bot propone solitamente un link a pagine con informazioni più complete e dettagliate, rappresentando di fatto un'alternativa all'uso di una maschera di ricerca tradizionale.
In Europa, il primo chatbot sembra essere quello lanciato il 28 maggio 2024 dall'Associazione delle biblioteche pubbliche di Berlino (Verbund der Öffentlichen Bibliotheken Berlins – VÖBB) per ampliare l'accesso alle risorse di oltre 80 biblioteche aderenti.[5] Anche in questo caso, il chatbot non si prefigge di esautorare dal proprio ruolo i professionisti dell'informazione né di sostituire le funzionalità del catalogo, ma piuttosto di costituire un canale complementare rispetto all'interfaccia di ricerca consentendo di

[1] https://pro.europeana.eu/page/ai4culture-platform (cons. 6/2/2026).
[2] Il progetto, denominato AI4Illuminations, è stato presentato da Paola Manoni nell'ambito del 4° Convegno nazionale del Centro di Eccellenza DTC Lazio organizzato in collaborazione con CHANGES *Tecnologie e patrimonio culturale: nuove competenze e professioni* tenutosi il 30 novembre 2023 alla Sapienza; il video integrale dell'intervento è disponibile al link https://www.youtube.com/live/CSyi3jxFlQg?si=KWqL4crGyromgtmh&t=1766 (cons. 6/2/2026).
[3] https://library.sjsu.edu/librarychatbot/kingbot (cons. 6/2/2026).
[4] "Kingbot is intended to provide basic information about the library and library resources. Kingbot is not a research tool or a replacement for librarian and library staff expertise": cfr. https://libapps.sjsu.edu/kingbot/ (cons. 6/2/2026).
[5] https://www.voebb.de/aDISWeb/app?service=direct/0/Home/$DirectLink&noRedir&sp=SPROD00&sp=SWI01000406 (cons. 6/2/2026); il codice, disponibile su GitHub, può essere eventualmente sperimentato e adattato alle proprie esigenze da altre istituzioni: https://github.com/voebb-dev/voebb-chatbot/ (cons. 6/2/2026).

ottenere informazioni preliminari in modo diretto e interattivo. A differenza di quanto accade nella maggior parte dei siti commerciali, la conversazione non viene avviata automaticamente: per aprire la finestra di dialogo è necessario cliccare sull'icona 'Galaxie-Bibliothekarin' che compare in basso a destra nella pagina (Fig. 1); come nel caso di Kingbot GPT, ove opportuno il bot reindirizza ai record presenti nel catalogo. Ancora da indagare sono il livello di soddisfazione degli utenti, le modalità di gestione di dati disomogenei e talvolta incompleti, i costi infrastrutturali, anche in termini ambientali, e le implicazioni etiche; ma si fa strada l'idea che alcune delle funzioni del catalogo ritenute ormai fondamentali per rispondere nel modo migliore alle necessità degli utenti, come il suggerimento di titoli simili a quelli cercati, potrebbero forse essere svolte in modo più efficace dai Large Language Models (LLM).

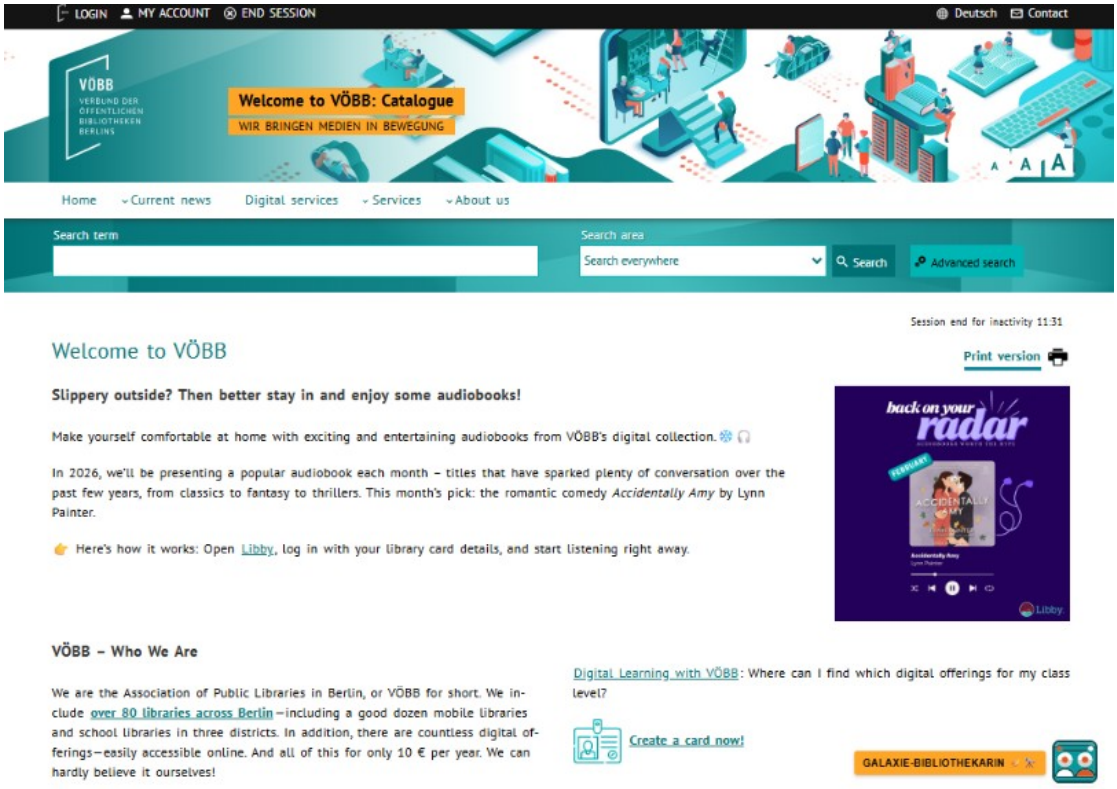


Figura 1. Home page del sito di VÖBB

3. IL MODELLO ALPHABETICA E IL 'REFERENCE OCCULTATO'

L'esperienza di Alphabetica prende forma nel 2017 da un'idea di Simonetta Buttò, allora direttrice dell'ICCU, insieme ad Alberto Petrucciani; viene poi affidata a Luigi Cerullo che elabora la progettazione teorica e ne segue successivamente la realizzazione (Cerullo, 2022). Il primo obiettivo del progetto era quello di integrare le basi dati bibliografiche nazionali gestite dall'ICCU, fino a quel momento tra di loro completamente indipendenti, e di renderle interrogabili tramite un punto d'accesso unico o sistema di ricerca integrato (SRI) dando vita contestualmente a un nuovo portale che fosse facilmente navigabile anche da un utente con scarsa dimestichezza con i cataloghi di biblioteca (Buttò, 2020; Buttò, 2022). Parallelamente agli sforzi per l'armonizzazione dei cataloghi esistenti, il progetto ha perseguito dunque lo scopo di creare uno strumento non soltanto nuovo ma basato su presupposti completamente innovativi, che fosse al tempo stesso attendibile ma sufficientemente attraente da rappresentare una valida alternativa ai motori di ricerca generalisti. In questa visione strategica è stata fondamentale l'adozione di una barra di ricerca semplice, a campo unico, per ricreare visivamente l'ambiente di ricerca familiare adottato dai più comuni motori di ricerca e definita appunto 'Google-like' (Bianchini, 2017; Fig. 2), senza prevedere una maschera avanzata e rimandando il raffinamento dei risultati a una fase successiva in cui questi sono presentati in 'box' (Castro, 2022b).



Figura 2. Home page di Alphabetica

L'elaborazione grafica di Alphabetica è stata particolarmente curata per costruire un'interfaccia accattivante e dinamica, lontana dall'austerità tipica dei cataloghi e volutamente dissimile anche dall'ecosistema gemello ICCU, strutturato invece sulla grafica comune dei portali che lo compongono, ovvero SBN, EDIT16 e Manus Online (Bruni & Ravelli, 2022). Questa apparenza colorata cela indici complessi elaborati sulle competenze specialistiche della biblioteconomia. Nell'era di internet e dei motori di ricerca, il reference rappresenta il cuore del lavoro dei professionisti dell'informazione nel fornire una mediazione mirata e affidabile tra le risorse e gli utenti (Ranganathan, 2010; Bianchini, 2017b).
L'originalità di Alphabetica consiste nell'aver messo a punto una forma particolare di reference digitale che è stato definito 'occultato', articolato in suggerimenti e percorsi di navigazione elaborati su criteri biblioteconomici da professionisti dell'informazione piuttosto che sulla profilazione degli utenti nella consueta logica dei motori di ricerca e dei siti commerciali (Castro, 2022a; Castro, 2022b).
Il 'reference occultato' si articola prima di tutto su thesauri selezionati in fase di progettazione della prospettazione dei risultati articolata in box; oltre ai box sono presenti altri strumenti per guidare l'utente nella scoperta di risorse inaspettate: il radar dei protagonisti, che restituisce graficamente la produzione editoriale di un autore e i collegamenti con altri personaggi, luoghi e periodi; e la mappa concettuale, che consente di espandere il campo di ricerca attraverso le correlazioni di concetti basate sul Nuovo Soggettario (Castro, 2022a; Castro, 2022b). Questa scelta risponde a una preoccupazione etica: mentre i motori di ricerca commerciali si servono della profilazione per inquadrare l'utente in una 'bolla informativa' di campagne pubblicitarie volte a massimizzare i profitti, le biblioteche mantengono un ruolo informativo e sociale neutrale (Petrucciani, 2022). Per questo, pur proponendo percorsi di navigazione articolati, le ricerche su Alphabetica presentano a ogni utente i medesimi risultati (Castro, 2022b).
I risultati dei focus group condotti in collaborazione con il laboratorio BIBLAB della Sapienza di Roma hanno confermato l'efficacia di questo approccio (Battaggia, 2022). Un docente ha sottolineato come la struttura di Alphabetica valorizzi l'interdisciplinarità e la multimedialità insita nel catalogo, suggerendo risorse di diversa natura (testi, video, manifesti e altro) che possono risultare utili per scopi didattici. Il commento di un utente che ha assimilato l'esperienza di navigazione al servizio di reference in biblioteca, "quando il bibliotecario ti porta anche cose che non cercavi", sintetizza perfettamente gli obiettivi che il progetto per la realizzazione di Alphabetica si proponeva e ne dimostra il successo (Castro, 2022a).

4. LE RISORSE ACCESSIBILI IN ALPHABETICA

Durante la fase di progettazione di Alphabetica, in particolare in seguito all'emanazione nel 2019 della direttiva UE 2019/882 o legge europea sull'accessibilità (European Accessibility Act – EAA), entrata pienamente in vigore il 28 giugno 2025 per l'adeguamento a standard rigorosi per prodotti e servizi digitali in favore delle persone con disabilità, è emerso un problema oggettivo riguardo la reperibilità nel catalogo di risorse accessibili di diversa natura: risorse in Braille, libri a caratteri ingranditi, audiolibri, libri in CAA (Comunicazione Aumentativa Alternativa, in cui alle parole espresse tramite l'alfabeto corrispondono simboli grafici), *silent book* (albi illustrati in cui il racconto procede esclusivamente attraverso i disegni), libri ad alta leggibilità (ovvero che utilizzano font privi di grazie, spaziature e contrasto più amichevoli) e LIA (e-book accessibili, cioè libri digitali il cui contenuto può essere adattato a diverse esigenze di lettura come l'ingrandimento dei caratteri, la modifica dei colori e dei contrasti per il testo e lo sfondo e altri accorgimenti su titoli e indici). Test condotti negli OPAC di vari sistemi bibliotecari hanno evidenziato la difficoltà di far emergere queste caratteristiche in quanto gli standard bibliografici attualmente non prevedono codici sufficienti per indicare i formati accessibili, come invece sarebbe auspicabile nel rispetto sia delle disabilità fisiche o cognitive, sia del valore inclusivo in contesti multiculturali. In UNIMARC, il formato di scambio bibliografico adottato anche nel protocollo SBN, il tag 106 permette di codificare solo i valori 'Braille' o 'A caratteri grandi' tra le tipologie di risorse accessibili. Mentre il problema di come esprimere adeguatamente nelle bibliografie nazionali e nei cataloghi collettivi le informazioni sull'accessibilità e altre questioni relative ai formati accessibili sono oggetto di studio nel gruppo di lavoro IFLA Accessibility Metadata and Accessible Formats,[6] per ovviare a questa lacuna l'ICCU ha sviluppato algoritmi specifici per recuperare le informazioni da campi diversi: tipo record, forma del contenuto, parole chiave, note bibliografiche e anche la corrispondenza con editori e collane specializzate nel settore. Nella home page di Alphabetica, in corrispondenza della ricerca nel canale "Libri" è stato inserito un filtro per limitare la ricerca alle sole risorse accessibili (Fig. 3). Contemporaneamente, nel nuovo software gestionale per la catalogazione SBNCloud sono state create nuove funzionalità per permettere di indicare in fase di catalogazione la tipologia di risorsa accessibile, garantendo una prospettiva di sostenibilità a lungo termine al di là degli algoritmi di recupero retrospettivi.

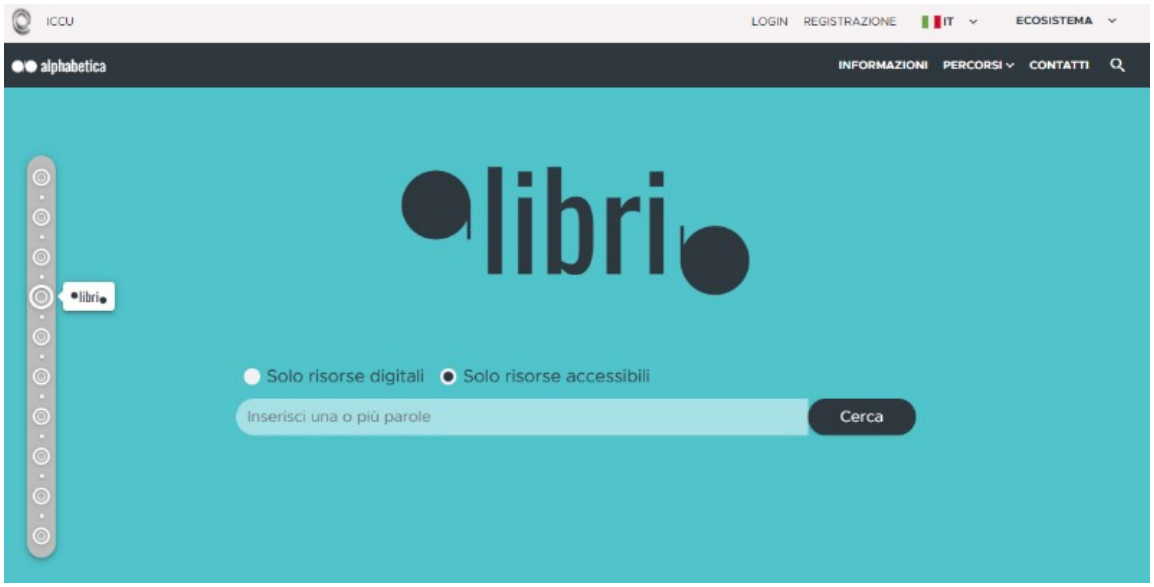


Figura 3. Il canale "Libri" nella home page di Alphabetica

5. PROSPETTIVE: L'INTEGRAZIONE DELL'IA E LE SFIDE APERTE

Nonostante l'impegno costante delle biblioteche nel promuovere e facilitare l'accesso a un patrimonio culturale condiviso, la questione della reperibilità delle risorse accessibili nei cataloghi e tramite i

[6] https://www.ifla.org/g/bibliography/accessibility-metadata-amp-accessible-formats-in-nbs-working-group/ (cons. 6/2/2026).

cataloghi non è stata finora oggetto di azioni o linee guida specifiche. Nel 2025 l'ICCU è risultato vincitore di un finanziamento destinato a istituti e luoghi della cultura per la partecipazione a iniziative di co-creazione di soluzioni innovative per la valorizzazione del patrimonio culturale per lo scenario 1: "Accessibilità e inclusione" nell'ambito del Piano Nazionale di Ripresa e Resilienza (PNRR) Missione M1 – componente C3 "Turismo e cultura 4.0", investimento 1.1 "Strategie e piattaforme digitali per il patrimonio culturale", sub-investimento 1.1.11 "Piattaforma di co-creazione e crowdsourcing". Il progetto, denominato "Il catalogo accessibile", prevede la creazione di un nuovo strumento per rendere direttamente fruibili le risorse accessibili dotate di certificazione per l'uso gratuito o a condizioni agevolate. Grazie alle potenzialità dell'Indice SBN e dell'interoperabilità tra diversi formati e standard, il progetto propone di sfruttare l'IA per il riconoscimento delle risorse accessibili presenti in banche dati selezionate per mettere a disposizione dell'utente le informazioni per l'accesso o le risorse stesse, nel caso in cui queste siano disponibili online, in un nuovo catalogo integrato con Alphabetica, di cui costituirà una banca dati afferente e collegata in fase di consultazione tramite il box "Libri accessibili". Recentemente ha fatto la sua comparsa in Alphabetica anche Alphy, un chatbot basato su IA in via di perfezionamento (Fig. 4) che può contribuire ad abbattere le barriere cognitive, sebbene presenti ancora alcune criticità riguardanti la neutralità e la pluralità dell'informazione fornita all'utente che dovranno essere affrontate nel rispetto dei principi etici che sono i pilastri del progetto. L'IA, tramite sistemi di indicizzazione automatica, potrebbe ulteriormente arricchire la mappa concettuale con risorse che attualmente non sono comprese, come manoscritti e libri a stampa antichi: in linea con l'approccio collaborativo adottato in fase di progettazione e con la visione espressa da S. R. Ranganathan nella sua quinta legge della biblioteconomia (Ranganathan, 2010b), il portale si configura da ogni punto di vista come "un organismo in crescita" (Castro, 2022a), aperto nei confronti delle esigenze della comunità SBN e soprattutto dei suoi utenti.

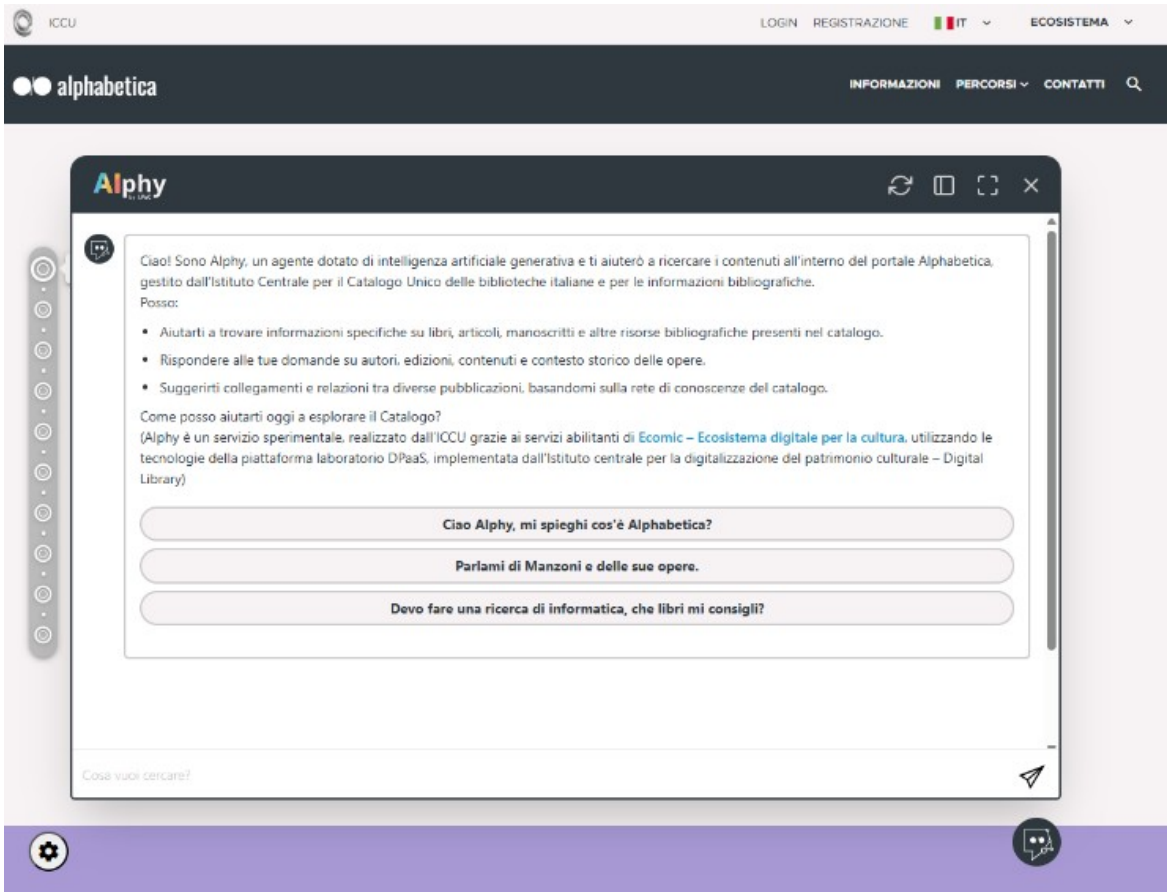


Figura 4. Alphy

BIBLIOGRAFIA

Bardelli, A., & Verga, F. (2019). The Discovery tool is a growing organism. JLIS.It, 10(3), 97–124. https://doi.org/10.4403/jlis.it-12575

Battaggia, M. (2022). Alphabetica: Aspettative, opinioni e ricadute attraverso una ricerca esplorativa. DigItalia, 17(1), 96–105.

Bianchini, C. (2017a). «Funziona come Google, vero?». Prima indagine sull'interazione utente-catalogo nella biblioteca del Dipartimento di musicologia e beni culturali (Cremona) dell'Università di Pavia. AIB studi, 57(1). https://doi.org/10.2426/aibstudi-11557

Bianchini, C. (2017b). La biblioteca è reference: Definizione, modelli, prospettive del servizio di reference. Biblioteche oggi Trends, 3(1). https://doi.org/10.3302/2421-3810-201701-004-1

Bruni, F., & Ravelli, E. (2022). EDIT16: Nuovi strumenti e nuove prospettive per la conoscenza e la valorizzazione del libro italiano del Cinquecento. DigItalia, 17(1), 66–75.

Buttò, S. (2020). Alphabetica, il nuovo portale per la ricerca integrata: Un salto di qualità per le biblioteche italiane. DigItalia, 15(2), 9–15. https://doi.org/10.36181/digitalia-00010

Buttò, S. (2022). Aperto, flessibile, integrato: Tre parole chiave per l'Ecosistema dei servizi bibliografici nazionali. DigItalia, 17(1), 11–17.

Castellucci, P., & Gomelino, E. (2021). Chatbot: Un giorno, al museo. DigItalia, 16(2), 9–24. https://doi.org/10.36181/digitalia-00034

Castro, E. (2022a). Alphabetica e il reference digitale: Suggerimenti e mappe concettuali per gli utenti del portale. AIB studi, 62(3), 559–570. https://doi.org/10.2426/aibstudi-12055

Castro, E. (2022b). Alphabetica: Uno strumento poliedrico di accesso ai servizi bibliografici nazionali. DigItalia, 17(1), 39–47.

Cerullo, L. (2022). L'Ecosistema digitale dei servizi bibliografici nazionali: Una panoramica. DigItalia, 17(1), 31–38.

Fontanin, M. (2022). Dalle fake news all'infodemia. Editrice Bibliografica. https://doi.org/10.53134/9788893575003

Fusetti, C., & Tagliani, E. (2019). Analizzare per innovare: Uno studio sull'usabilità dei cataloghi Discovery NG con strumenti di web analytics. Biblioteche oggi Trends, 5 (1). https://doi.org/10.3302/2421-3810-201901-026-1

Manzoni, L. (2023). L'indicizzazione semantica in automatico o semiautomatico nelle principali biblioteche europee. Biblioteche oggi Trends, 57. https://doi.org/10.3302/2421-3810-202301-057-1

Petrucciani, A. (2022). L' utopia della biblioteca e la missione del bibliotecario. AIB studi, 61(3), 597–602. https://doi.org/10.2426/aibstudi-13399

Ranganathan, S. R. (2010a). Il servizio di reference. Le Lettere.

Ranganathan, S. R. (2010b). Le cinque leggi della biblioteconomia. Le Lettere.

Roncaglia, G. (2025). Verso il reference generativo? Biblioteche oggi Trends, 11(1), 63–73. https://www.doi.org/10.3302/2421-3810-202501-063-1

Roncaglia, G. (2026). Cataloguing, Metadata, and Generative AI. Early Experiences and Future Perspectives. JLIS.It, 17(1), 106–127. https://doi.org/10.36253/jlis.it-693

Storti, C. (2024). I requisiti per l'addestramento degli strumenti di Intelligenza Artificiale e il deposito legale delle risorse digitali: Una riflessione sul contesto normativo italiano e il ruolo delle biblioteche. DigItalia, 19(1), 23–35. https://doi.org/10.36181/digitalia-00092

Grafoteca e la costruzione digitale della memoria manoscritta: curatela, metadatazione e reti di risorse

AMALIA CARRANO
INDEPENDENT RESEARCHER – GRAFOTECA
INFO@GRAFOTECA.COM

ABSTRACT (ITALIANO)

La scrittura a mano, intesa come pratica culturale e oggetto di studio, è oggi attraversata da una progressiva marginalizzazione nei contesti digitali; parallelamente, le risorse dedicate alla cultura grafologica, intesa come tradizione di studi e pratiche sull'analisi della grafia, risultano spesso frammentate tra collezioni fisiche e ambienti eterogenei. Tale dispersione rende complessa la reperibilità delle fonti, la loro contestualizzazione e il riuso informato in ambito di ricerca e di valorizzazione culturale. Il contributo presenta Grafoteca, una biblioteca digitale curatoriale dedicata alla manoscrittura e alla cultura grafologica, basata su attività di selezione, curatela e descrizione. Il progetto raccoglie e mette in relazione testi, studi e materiali documentari, costruendo un ambiente informativo coerente che favorisce il recupero, l'esplorazione e il riuso. Accanto all'archivio descrittivo, la piattaforma integra una rete di collegamenti a portali e collezioni digitali esterne, configurandosi come strumento di orientamento tra biblioteche, archivi e istituzioni culturali nazionali e internazionali impegnate nella conservazione e valorizzazione dei patrimoni manoscritti. Il sistema è realizzato su Omeka Classic e adotta lo standard Dublin Core come schema descrittivo di base, integrato da pratiche di metadatazione arricchita e curatela digitale. L'architettura informativa si fonda su percorsi tematici e strategie di mediazione culturale volte a facilitare la navigazione tra contenuti e contesti. Il contributo illustra le scelte progettuali e metodologiche adottate e discute prospettive di sviluppo verso ecosistemi più interoperabili e modalità di esplorazione evolute. Grafoteca si configura come modello di biblioteca digitale tematica utile sia alla valorizzazione della manoscrittura sia alla diffusione interdisciplinare di un patrimonio specialistico.
Parole chiave: biblioteche digitali; curatela digitale; metadatazione; scrittura manoscritta; grafologia

ABSTRACT (ENGLISH)

Grafoteca and the digital construction of manuscript memory: curation, metadata and resource networks.
Handwriting, as both a cultural practice and an object of study, is currently experiencing progressive marginalization within digital environments. At the same time, materials related to graphological culture, understood as a tradition of studies and practices focused on the analysis of handwritten texts, are often fragmented across physical collections and heterogeneous digital platforms. This dispersion makes source retrieval, contextualization, and informed reuse for research and cultural heritage enhancement purposes more difficult.
This paper presents Grafoteca, a curated digital library dedicated to handwriting and graphological culture, based on systematic activities of selection, curation, and metadata creation. The project gathers and connects texts, studies, and documentary materials, building a coherent information environment that facilitates retrieval, exploration, and reuse. Alongside its descriptive archive, the platform integrates a network of links to external digital portals and collections, functioning as an orientation hub connecting national and international libraries, archives, and cultural institutions engaged in the preservation and enhancement of handwritten heritage.
The system is built on the Omeka Classic platform and adopts the Dublin Core standard as its core metadata schema, enriched through enhanced descriptive practices and digital curation strategies. Its information architecture is based on thematic exploration paths and cultural mediation strategies designed to facilitate navigation across resources and contexts. The paper outlines the adopted design and methodological choices and discusses future developments toward more interoperable information ecosystems and advanced exploration models. In this framework, Grafoteca is conceived as a methodological model for a thematic digital library supporting both the valorization of handwriting and the interdisciplinary dissemination of specialized heritage.
Keywords: digital libraries; digital curation; metadata; handwriting; graphology

1. INTRODUZIONE

La scrittura a mano ha accompagnato la storia culturale dell'umanità come mezzo di espressione del pensiero, di trasmissione della conoscenza e di costruzione della memoria individuale e collettiva. Anche nell'era digitale, continua a rappresentare una pratica significativa, legata non soltanto alla comunicazione, ma anche alla formazione dell'identità, alla dimensione cognitiva e alla conservazione della traccia personale e documentaria. La progressiva riduzione del suo utilizzo non è priva di conseguenze: ricerche neuroscientifiche e psicologiche hanno evidenziato come la scrittura manuale coinvolga processi complessi, con ricadute sulla memoria, sul linguaggio e sull'organizzazione del pensiero.
Parallelamente, la cultura grafologica, a partire dalle prime sistematizzazioni ottocentesche di Jean-Hippolyte Michon, ha sviluppato nel corso del Novecento un articolato insieme di testi, riviste, studi, metodologie e scuole interpretative. Tale patrimonio costituisce oggi una risorsa disciplinare rilevante, ma spesso frammentata e di difficile reperibilità. Le fonti grafologiche risultano infatti distribuite tra pubblicazioni associative, archivi di istituti di formazione, collezioni private e piattaforme digitali eterogenee per struttura e livelli di accesso, in ambito nazionale e internazionale.
In questo scenario, le Digital Humanities offrono strumenti e metodologie per la costruzione di ambienti digitali strutturati, capaci di aggregare, descrivere e valorizzare patrimoni culturali dispersi, favorendo nuove modalità di accesso, esplorazione e riuso delle fonti. In tale contesto si colloca Grafoteca, progetto di biblioteca digitale dedicato alla valorizzazione della scrittura a mano e della cultura grafologica. Il contributo presenta il modello concettuale e operativo del progetto, inteso come ambiente curatoriale digitale, illustrandone le scelte metodologiche, l'architettura informativa e lo stato attuale di implementazione, nonché le prospettive di sviluppo verso un ecosistema informativo interoperabile.

2. CONTESTO E LAVORI CORRELATI

Negli ultimi anni, il campo delle Digital Humanities ha favorito la costruzione di biblioteche digitali, archivi virtuali ed edizioni digitali come strumenti privilegiati per la valorizzazione, l'accesso e la preservazione del patrimonio culturale (Ciotti, 2023; Fiormonte et al., 2010; Tomasi, 2022). La digitalizzazione di fondi documentari e la creazione di ambienti informativi dedicati hanno contribuito allo sviluppo di nuovi modelli di ricerca basati sulla descrizione strutturata delle risorse, sull'interoperabilità dei dati e sulla costruzione di percorsi tematici di esplorazione (Ridi, 2004; Salvatori, 2017).
Parallelamente, si è sviluppata una crescente attenzione verso la scrittura a mano non soltanto come mezzo di comunicazione, ma anche come pratica culturale significativa sul piano dei processi mentali. Ricerche neuroscientifiche e psicologiche indicano che la produzione manoscritta coinvolge processi cognitivi complessi legati alla memoria, al linguaggio, alla coordinazione motoria e all'apprendimento (Berninger & Richards, 2010; James & Engelhardt, 2012). Dal punto di vista culturale, tali evidenze si accompagnano a un dibattito nazionale e internazionale sul valore della scrittura manuale. In Italia sono state avviate iniziative volte al riconoscimento della scrittura corsiva come patrimonio culturale immateriale, riflettendo una crescente attenzione alla scrittura a mano non solo come competenza tecnica, ma come pratica connessa all'identità culturale, alla memoria collettiva e alla trasmissione intergenerazionale dei saperi (Manifesto per la candidatura della scrittura manuale corsiva, 2023; UNESCO, 2003).
In questo quadro, la manoscrittura è riconosciuta come pratica che favorisce l'elaborazione profonda delle informazioni e l'articolazione del pensiero, distinguendosi dalle modalità di scrittura digitale e **rafforzando la percezione della sua rilevanza culturale ed educativa** (Scientific American, 2014).
Nel panorama dei progetti digitali, diverse iniziative si concentrano sulla digitalizzazione e descrizione dei manoscritti storici, sviluppando collezioni digitali e sistemi avanzati di accesso alle fonti. Le edizioni digitali scientifiche costituiscono oggi uno dei principali modelli di valorizzazione delle fonti culturali in ambiente digitale (Del Grosso et al., 2019; Tomasi, 2013). Accanto alle iniziative infrastrutturali, emergono anche progetti culturali che promuovono la scrittura a mano come pratica espressiva e comunicativa. Un esempio significativo è *Digiti. Rivista manoscritta* (e-ISSN 3035-2843), pubblicazione semestrale del Dipartimento di Lettere e Filosofia dell'Università di Trento, nata da un progetto didattico dedicato allo sviluppo delle potenzialità comunicative della scrittura manuale e realizzata con il coinvolgimento di studenti, dottorandi e docenti. La rivista propone la scrittura manoscritta come medium

comunicativo alternativo, valorizzando la varietà di scritture, lingue e sistemi grafici come forme di espressione culturale (Digiti, n.d.).
Rimangono invece limitate le infrastrutture digitali dedicate specificamente alla cultura grafologica e allo studio della scrittura a mano contemporanea. Le risorse grafologiche risultano oggi disponibili attraverso canali eterogenei, tra cui riviste associative, siti di istituti, archivi digitali e collezioni esterne, con livelli diversi di accesso e descrizione, rendendo complessa una ricognizione unitaria delle fonti.
Nel contesto europeo, la disponibilità di testi grafologici digitalizzati presenta livelli differenti. In Francia, infrastrutture nazionali di digitalizzazione come Gallica (Bibliothèque nationale de France) consentono un'ampia consultazione di opere storiche digitalizzate, incluse numerose pubblicazioni riconducibili alla tradizione grafologica. Nel contesto italiano, la digitalizzazione sistematica di testi grafologici, in particolare delle opere del Novecento e della produzione specialistica contemporanea, risulta ancora frammentaria e affidata prevalentemente a iniziative associative o archivi privati. Questa asimmetria evidenzia l'opportunità di sviluppare ambienti digitali strutturati dedicati alla ricognizione, descrizione e valorizzazione delle fonti grafologiche, in dialogo con i modelli descrittivi e le pratiche di curatela proprie delle biblioteche digitali e di altri ambienti informativi.

3. ARCHITETTURA INFORMATIVA E MODELLO DI DESCRIZIONE

Grafoteca[1] è realizzata su piattaforma Omeka Classic, scelta per la sua ampia diffusione nei progetti di biblioteca digitale e per la capacità di supportare la gestione di collezioni digitali e la creazione di percorsi espositivi online (Hardesty, 2014; Salarelli, 2016; Omeka Team, 2023).
Il progetto fa riferimento al set di metadati Dublin Core, impiegato per garantire una rappresentazione standardizzata delle risorse e favorire l'interoperabilità e il riuso dei dati (Dublin Core Metadata Initiative, 2020; Tomasi, 2022). Il modello informativo si fonda su pratiche di metadatazione arricchita e curatela digitale, in coerenza con i modelli di organizzazione della conoscenza e data curation sviluppati nell'ambito delle biblioteche digitali (Ciotti, 2023; Ridi, 2004). Questo approccio favorisce non solo la reperibilità delle fonti, ma anche la loro contestualizzazione e la costruzione di percorsi di navigazione tematici e trasversali.
Le risorse sono organizzate in collezioni tematiche che riflettono i principali ambiti della disciplina e le sue diverse applicazioni. Attraverso le funzionalità espositive di Omeka Classic, queste si articolano in percorsi narrativi e mostre digitali, favorendo una fruizione guidata e un'esplorazione interdisciplinare dei contenuti.
Accanto all'archivio catalografico, la sezione "Esplora con i link" opera come aggregatore tematico di portali e raccolte digitali dedicate alla scrittura a mano, alla grafologia e alla documentazione manoscritta. I collegamenti rimandano a biblioteche nazionali, archivi storici, istituzioni accademiche, musei, fondi speciali e associazioni professionali, italiane e internazionali, nonché a reti accademiche e iniziative culturali dedicate alla diffusione della ricerca e della manoscrittura, permettendo l'accesso a risorse digitalizzate, repertori bibliografici e mostre digitali. Questa rete svolge una funzione di orientamento informativo, culturale e disciplinare, rendendo accessibili contenuti già esistenti ma distribuiti su piattaforme eterogenee. Grafoteca è inoltre integrata con un'interfaccia web basata su WordPress, utilizzata esclusivamente come homepage pubblica e come strumento di mediazione comunicativa del progetto. Tale soluzione consente una presentazione accessibile anche a utenti non specialisti, mantenendo separata l'infrastruttura descrittiva dalla comunicazione web.
Nel suo insieme, il modello adottato combina descrizione strutturata, organizzazione tematica e curatela digitale, configurando Grafoteca come ambiente informativo coerente per la ricognizione e la valorizzazione delle risorse legate alla scrittura manoscritta e alla cultura grafologica.

4. STATO ATTUALE DI IMPLEMENTAZIONE

Grafoteca è operativa come biblioteca digitale e raccoglie, in forma descrittiva e curatoriale, risorse bibliografiche e documentarie dedicate alla cultura grafologica in ambito nazionale e internazionale. Le collezioni sono organizzate secondo criteri tematici e disciplinari e consentono la consultazione di schede descrittive corredate da dati bibliografici essenziali e collegamenti a piattaforme digitali esterne, in larga parte estere, che rendono disponibili i contenuti originali.

[1] https://www.grafoteca.com (cons. 18/04/2026)

L’archivio costituisce il nucleo informativo dell’iniziativa e permette una prima ricognizione strutturata e normalizzata delle fonti grafologiche, spesso disperse tra riviste associative, pubblicazioni di istituti di formazione, cataloghi bibliotecari e archivi digitali eterogenei per struttura e modalità di accesso. Le schede non ospitano direttamente i testi digitalizzati, anche per vincoli legati al diritto d’autore e alla protezione dei dati, ma ne documentano l’esistenza, la collocazione e, ove possibile, le modalità di consultazione online tramite collegamenti a biblioteche digitali e piattaforme esterne, secondo le condizioni di accesso previste dai rispettivi siti. Svolgono così una funzione di orientamento e mediazione tra utenti e fonti.
La sezione “Esplora con i link” opera come aggregatore tematico di portali e collezioni digitali dedicati alla scrittura a mano, alla grafologia e alla documentazione manoscritta. I collegamenti rimandano a biblioteche nazionali, archivi storici, istituzioni accademiche, musei, fondi speciali e associazioni professionali, italiane e internazionali, consentendo l’esplorazione di materiali digitalizzati, repertori bibliografici, schede di manoscritti e mostre digitali. Questa rete svolge una funzione di accesso guidato e di orientamento informativo verso contenuti già disponibili online ma distribuiti su ambienti differenti.
Attraverso le funzionalità espositive di Omeka Classic, alcune collezioni sono già articolate in percorsi tematici che propongono modalità di fruizione narrativa e favoriscono l’esplorazione trasversale dei contenuti. Un esempio è la mostra digitale dedicata al rapporto tra scrittura, alterazione grafica e perizia documentaria nel contesto storico-giudiziario del Novecento. Tali strumenti consentono una **mediazione e una contestualizzazione culturale** della scrittura manoscritta e, parallelamente, un approfondimento specialistico in ambito grafologico.
L’attuale livello di sviluppo configura una prima infrastruttura operativa, orientata alla progressiva crescita delle collezioni, al rafforzamento dei collegamenti con archivi, biblioteche digitali esterne e associazioni internazionali, nonché all’eventuale attivazione di collaborazioni con contesti di ricerca universitari, e allo sviluppo di strumenti più avanzati di esplorazione delle risorse. In prospettiva, sono allo studio possibili accordi con autori, editori e istituti grafologici per favorire la digitalizzazione autorizzata di testi contemporanei e la loro progressiva inclusione nell’ambiente descrittivo della piattaforma.

5. OBIETTIVI DEL PROGETTO

Grafoteca si configura come progetto radicato nel contesto italiano e orientato al dialogo nazionale e internazionale, lungo due direttrici principali, distinte ma complementari: il supporto alla ricerca grafologica e la valorizzazione culturale della manoscrittura. La loro integrazione costituisce il nucleo metodologico e orienta le scelte progettuali, descrittive e di curatela.
Il primo obiettivo riguarda il supporto alla ricerca grafologica attraverso la costruzione di un ambiente digitale strutturato per la ricognizione, la catalogazione e l’organizzazione delle fonti disciplinari. Grafoteca nasce per rispondere alla frammentazione che caratterizza le risorse grafologiche nel contesto italiano, spesso disperse tra testi difficilmente reperibili, riviste specialistiche, atti di convegni, fondi documentari e pubblicazioni associative non indicizzate nei cataloghi generalisti, mantenendo al contempo aperture e confronti con la produzione scientifica internazionale. Tale dispersione pone un problema di accesso, descrizione e organizzazione della conoscenza che il progetto affronta mediante la realizzazione di un’infrastruttura basata su modelli bibliografici standardizzati e pratiche di metadatazione arricchita, in grado di individuare e censire le principali risorse grafologiche esistenti e di documentarne la collocazione e le modalità di consultazione.
Accanto alla funzione repertoriale, l’iniziativa mira a strutturare la conoscenza disciplinare grafologica attraverso schede dedicate alle scuole storiche, agli autori di riferimento, ai concetti teorici e alle metodologie di analisi, configurandosi come ambiente informativo a supporto della ricerca e della modellazione concettuale del dominio grafologico.
Il secondo obiettivo concerne la valorizzazione della scrittura a mano come patrimonio culturale, intesa non soltanto come oggetto di studio specialistico, ma come pratica significativa nella costruzione della memoria, dell’identità e nella trasmissione culturale. In questa prospettiva, Grafoteca non si configura come infrastruttura di digitalizzazione o sede primaria di pubblicazione dei materiali, ambito già presidiato da portali e aggregatori culturali di rilievo internazionale, ma agisce in modo complementare, mettendo a disposizione strumenti di orientamento e contestualizzazione in un ecosistema di risorse distribuite.
Questa attività si realizza attraverso una rete di collegamenti strutturati verso biblioteche, archivi, musei e associazioni grafologiche nazionali e internazionali, raccolti nella sezione “*Luoghi della Manoscrittura”*,

concepita come aggregatore tematico di portali e collezioni digitali dedicati alla scrittura a mano e alla cultura grafologica. La sezione consente un accesso guidato a risorse conservate presso istituzioni culturali, piattaforme digitali e progetti di ricerca, senza duplicarne i contenuti. Grafoteca svolge così un ruolo di raccordo e contestualizzazione, favorendo l'individuazione e l'esplorazione di materiali distribuiti in ambienti istituzionali e digitali eterogenei.

6. CONCLUSIONI

Il contributo evidenzia il ruolo di Grafoteca come ambiente informativo digitale dedicato alla valorizzazione della scrittura manoscritta e della cultura grafologica, attraverso l'integrazione di pratiche di curatela, descrizione strutturata e costruzione di percorsi tematici di esplorazione. Mediante l'adozione di standard descrittivi consolidati e piattaforme per la pubblicazione di collezioni digitali, l'iniziativa rende accessibile un patrimonio documentario e bibliografico spesso disperso tra archivi, pubblicazioni associative e collezioni eterogenee.
La combinazione tra archivio descrittivo, rete di collegamenti a risorse esterne, percorsi tematici e mostre digitali a carattere curatoriale offre modalità diversificate di accesso, esplorazione e contestualizzazione delle fonti.
Le scelte architetturali e metodologiche adottate pongono le basi per una progressiva evoluzione dell'ambiente informativo, orientata all'ampliamento delle collezioni, al rafforzamento delle relazioni con archivi e biblioteche digitali nazionali e internazionali, e allo sviluppo di strumenti più avanzati di interrogazione e analisi delle risorse.
In questa prospettiva, Grafoteca si configura come modello di ambiente digitale specialistico capace di contribuire alla costruzione di una memoria organizzata della scrittura manoscritta e della cultura grafologica, favorendo il dialogo tra ricerca umanistica, curatela digitale e modelli avanzati di accesso e valorizzazione del patrimonio culturale.

Dichiarazione sull'uso di strumenti di IA:

Sono stati utilizzati strumenti di intelligenza artificiale generativa unicamente per supporto redazionale (revisione linguistica e stilistica). Contenuti, struttura e scelte scientifiche sono di esclusiva responsabilità dell'autore.

BIBLIOGRAFIA

Ciotti, F. (2023). Digital humanities. Metodi, strumenti, saperi. Carocci.

Del Grosso, A. M., Capizzi, E., Cristofaro, S., De Luca, M. R., Giovannetti, E., Marchi, S., Seminara, G., & Spampinato, D. (2019). Bellini's correspondence: A digital scholarly edition for a multimedia museum. Umanistica Digitale, 3(7), 23–47. https://doi.org/10.6092/issn.2532-8816/9162

Digiti. (n.d.). Rivista manoscritta. https://teseo.unitn.it/digiti

Dublin Core Metadata Initiative. (2020). Dublin Core metadata element set, version 1.1: Reference description. https://www.dublincore.org/specifications/dublin-core/dces/

Fiormonte, D., Numerico, T., & Tomasi, F. (2010). L'umanista digitale. Il Mulino.

Hardesty, J. (2014). Exhibiting library collections online: Omeka in context. New Library World, 115(3/4), 75–86. https://doi.org/10.1108/NLW-01-2014-0013

James, K. H., & Engelhardt, L. (2012). The effects of handwriting experience on functional brain development in pre-literate children. Trends in Neuroscience and Education, 1(1), 32–42. https://doi.org/10.1016/j.tine.2012.08.001

Meschini, F. (2018). Reti, memoria e narrazione: Archivi e biblioteche digitali tra ricostruzione e racconto. Sette Città.

Omeka Team. (2023). Omeka Classic user manual. Roy Rosenzweig Center for History and New Media. https://omeka.org/classic/docs/

Ridi, R. (2004). La biblioteca digitale: definizioni, ingredienti e problematiche. Bollettino AIB, 44(3), 273–344.

Salarelli, A. (2016). Management of small digital collections with Omeka: The MoRE experience. Bibliothecae, 5(2), 177–200. https://doi.org/10.6092/issn.2283-9364/6393

Salvatori, E. (2017). Storia digitale e pubblica: Lo storico tra i "nuovi creatori" di storia. In P. Bertella Farnetti, L. Bertucelli, & A. Botti (Eds.), Public history. Problemi e pratiche. Mimesis.

Scientific American. (2014). Why writing by hand is better for memory and learning. https://www.scientificamerican.com/article/why-writing-by-hand-is-better-for-memory-and-learning/

Tomasi, F. (2013). Edizioni digitali come nuovo modello di dati di autorità concettuale. JLIS.it, 4(2), 21–44. https://doi.org/10.4403/jlis.it-8808

Tomasi, F. (2022). Organizzare la conoscenza: Digital humanities e web semantico. Editrice Bibliografica.

UNESCO. (2003). Convention for the safeguarding of the intangible cultural heritage. UNESCO. https://ich.unesco.org/en/convention

Valorizzazione digitale del patrimonio culturale: Public-Private Partnership e partecipazione sociale

ANNALISA CILIBERTI
ACCADEMIA DI BELLE ARTI DI BARI, ITALY –
A.CILIBERTI@ACCADEMIABARI.IT

ABSTRACT (ITALIANO)

Il contributo indaga le potenzialità della valorizzazione digitale applicata al patrimonio culturale rurale, con specifico riferimento alle masserie storiche della Murgia pugliese, bene prevalentemente privato e a bassa visibilità turistica. Muovendo da un quadro teorico che interpreta il patrimonio come risorsa condivisa della memoria collettiva, lo studio propone un modello sperimentale fondato su tre componenti operative interdipendenti: un archivio digitale partecipato, produzioni artistiche site-specific e percorsi esperienziali georeferenziati. Il modello è coordinato da una governance tripartita che coinvolge il soggetto privato proprietario, l'Accademia di Belle Arti di Bari e le comunità locali, configurandosi come una Public-Private Partnership orientata a obiettivi culturali condivisi. L'analisi di casi comparativi offre riferimenti metodologici per contesti analoghi. L'obiettivo è attivare forme di valorizzazione sostenibili e partecipative, capaci di connettere memoria locale, produzione artistica contemporanea e fruizione pubblica, senza ricorrere a interventi invasivi sul patrimonio materiale.
Parole chiave: patrimonio culturale rurale; valorizzazione digitale; Public-Private Partnership; partecipazione comunitaria; masserie murgiane.

ABSTRACT (ENGLISH)

Digital promotion of cultural heritage: public-private partnerships and community involvement
This paper investigates the potential of digital enhancement applied to rural cultural heritage, with a specific focus on the historical *masserie* (fortified farmhouses) of the Apulian Murgia—a heritage category that is predominantly privately owned and characterized by low tourist visibility.
Building upon a theoretical framework that interprets heritage as a shared resource of collective memory, the study proposes an experimental model based on three interdependent operational components: a participatory digital archive, site-specific artistic productions, and georeferenced experiential itineraries. The model is coordinated by a tripartite governance structure involving private owners, Accademia di Belle Arti di Bari, and local communities, taking the form of a Public-Private Partnership (PPP) oriented toward shared cultural objectives.
A comparative case analysis provides methodological benchmarks for similar contexts. The ultimate goal is to activate sustainable and participatory forms of enhancement capable of connecting local memory, contemporary artistic production, and public engagement, without resorting to invasive interventions on the physical heritage.
Keywords: rural cultural heritage; digital enhancement; Public-Private Partnership; community participation; Murgian masserie.

1. INTRODUZIONE

La valorizzazione del patrimonio culturale attraverso le tecnologie digitali rappresenta oggi un tema di crescente rilevanza nel dibattito accademico[1], configurandosi come una sfida cruciale per coniugare fruizione partecipativa, tutela dei valori culturali e sviluppo sostenibile[2].

[1] Contributi presentati al convegno *Patrimonio culturale al futuro: sostenibilità sociale, innovazione tecnologica, trasformazione digitale*, Università La Sapienza, 2026; vedi Fondazione CHANGES, *Mappe e scenari per il futuro del patrimonio culturale*, abstract book 2026.
[2] Questo contributo si colloca nell'ambito del dottorato in Valorizzazione e rigenerazione creativa dei beni culturali — istituito dall'Accademia di Belle Arti di Bari, cofinanziato dalla holding Renewia S.p.A., società proprietaria di masserie storiche nel territorio murgiano.

Le masserie storiche del territorio pugliese — centri nevralgici di produzione e accoglienza nonché testimonianze identitarie del contesto rurale — e in particolare quelle della Murgia, costituiscono un patrimonio di proprietà prevalentemente privata che la letteratura ha censito e classificato nella sua distribuzione territoriale e nelle sue caratteristiche tipologiche (Ruggiero, Parlavecchia, Dal Sasso, 2019), ma che raramente è stato oggetto di strategie digitali consapevoli o di modelli di partenariato orientati alla fruizione pubblica. La Murgia custodisce un patrimonio architettonico denso e stratificato non sufficientemente valorizzato nei circuiti culturali e turistici regionali.
Dalle potenzialità inespresse di questo territorio scaturisce una domanda: in che modo una strategia digitale responsabile può favorire una valorizzazione autentica del patrimonio culturale rurale, promuovendo pratiche turistiche etiche e sostenibili, forme di public engagement e co-produzione culturale?
Come dimostrano casi analoghi nel Mezzogiorno, l'attivazione di questo potenziale richiede approcci critici e contestualizzati (Marsili, 2024), all'interno di un quadro teorico che interpreta tale patrimonio come una risorsa condivisa, il cui valore trascende il possesso giuridico per divenire elemento costitutivo della memoria sociale e dell'identità collettiva (Dümcke e Gnedovsky, 2013). In questa prospettiva, le Public-Private Partnerships (PPP) — intese come forme di collaborazione strutturata tra soggetti pubblici, attori privati profit e non-profit e comunità locali, basate sulla condivisione di risorse, rischi e responsabilità decisionali verso obiettivi culturali comuni — emergono come un modello potenzialmente virtuoso per integrare istanze economiche e sociali nella gestione del patrimonio. Questi partenariati possono generare esiti positivi e duraturi se sostenuti da meccanismi di governance trasparenti, equi e orientati al bene comune (Pignatti, 2022).

2. METODOLOGIA DI RICERCA

La ricerca prende le mosse dall'analisi di alcuni casi studio, indagati in una prospettiva design-oriented, al fine di sviluppare una proposta progettuale sperimentale applicabile al territorio murgiano. I casi studio sono stati selezionati secondo tre criteri: (1) presenza di una componente digitale applicata alla valorizzazione del patrimonio culturale; (2) adozione di un modello di governance condivisa tra soggetti pubblici, privati e comunità locali; (3) comparabilità con il contesto murgiano in termini di tipologia patrimoniale o condizione di marginalità turistica. I casi sono assunti come dispositivi euristici e letti attraverso una griglia interpretativa articolata su dimensioni di governance, strumenti digitali impiegati, grado di partecipazione comunitaria e livelli di accessibilità pubblica. Tale approccio consente di evidenziare pattern ricorrenti, criticità operative e condizioni di trasferibilità delle strategie osservate.
Ad eccezione di contributi teorici fondativi, l'arco temporale della revisione della letteratura è limitato agli ultimi dieci anni al fine di garantire l'attualità dei riferimenti analizzati.
Il Borgo Storico Seghetti Panichi — finanziato nell'ambito del PNRR come uno dei sette Giardini Storici della Regione Marche — è selezionato come caso di patrimonio privato valorizzato attraverso partenariato pubblico-privato (PPP) in contesto rurale. Il caso consente di analizzare l'interazione tra governance multilivello e strumenti digitali avanzati, con particolare riferimento a rappresentazione 3D ed esperienze virtuali AI-driven (Clini et al., 2025).
Su scala europea, due progetti Horizon 2020 offrono riferimenti di campo utili a contestualizzare la proposta. Il progetto CrossCult (2016–2019) ha sviluppato tecnologie di narrazione digitale distribuita applicate a reti di siti culturali minori, mostrando come strumenti digitali possano connettere patrimoni eterogenei in percorsi narrativi integrati e sostenere la partecipazione attiva dei visitatori (Kontiza et al., 2020). Il progetto rurAllure (2021–2023) ha promosso siti rurali e musei minori nelle vicinanze di percorsi di pellegrinaggio europei attraverso piattaforme digitali e contenuti georeferenziati, dimostrando la praticabilità di approcci digitali applicati al patrimonio rurale disperso e a bassa visibilità turistica (European Commission, 2023). Entrambi i progetti sono assunti come riferimenti di campo che attestano la vitalità di approcci digitali orientati alla valorizzazione di patrimoni minori in contesti periferici.

3. DIGITALIZZAZIONE, SOSTENIBILITÀ E FRUIZIONE PARTECIPATIVA

Una strategia digitale efficace dovrebbe poter integrare criteri di accessibilità, definizione dei pubblici e obiettivi culturali senza limitarsi alla scelta degli strumenti tecnologici. Come evidenzia Maietti (2023), strumenti digitali che possano definirsi responsabili dovrebbero orientarsi verso forme di accesso socialmente sostenibili, capaci di valorizzare la qualità dell'esperienza culturale e la partecipazione delle comunità; nel contempo dovrebbero operare in sinergia con strategie di sviluppo turistico e sostenibile,

riconnettendo i siti minori a reti territoriali più ampie e favorendo la cooperazione tra istituzioni, comunità locali e industrie culturali.
Un orientamento fondativo è offerto dalla Convenzione di Faro del Consiglio d'Europa (2005), che definisce il patrimonio culturale come «a group of resources inherited from the past which people identify, independently of ownership, as a reflection and expression of their constantly evolving values, beliefs, knowledge and traditions». In coerenza con le linee guida UNESCO e con i Sustainable Development Goals, tale prospettiva attribuisce al patrimonio una funzione di coesione sociale e sviluppo umano.
Numerosi patrimoni minori, caratterizzati da risorse limitate e da una visibilità insufficiente ad attivare dinamiche di fruizione e riconoscimento, restano ai margini dei processi di valorizzazione culturale (Kontiza et al., 2020). In tali condizioni, il patrimonio tende a perdere la propria funzione di risorsa condivisa, rimanendo isolato rispetto alle pratiche di accesso e interpretazione pubblica.
A partire da questi orientamenti, la proposta qui sviluppata prevede piattaforme di archiviazione partecipativa per la produzione condivisa della memoria culturale e sistemi di georeferenziazione audio per reinserire tale memoria nel paesaggio fisico, attivando esperienze simultaneamente narrative e situate. Affinché questi strumenti producano contenuti effettivamente radicati nel territorio, è tuttavia necessario introdurre un ulteriore elemento capace di connettere la produzione digitale alla specificità del contesto locale. Per generare contenuti autenticamente radicati nel territorio, è necessario implementare dispositivi capaci di ancorare la produzione digitale alle peculiarità identitarie dei contesti locali.
Le residenze artistiche site-specific possono rispondere a questa esigenza: sono infatti pratiche consolidate nella rigenerazione culturale, in cui gli artisti operano in relazione diretta con i luoghi e le comunità, contribuendo alla produzione di contenuti ancorati alla dimensione storica e sociale del territorio (Candy & Edmonds, 2011).
Le memorie raccolte possono diventare materia per la pratica artistica, che restituisce al territorio contenuti fruibili attraverso dispositivi narrativi georeferenziati. È all'interno di questa integrazione tra archiviazione, produzione culturale e fruizione situata che viene avanzata la proposta progettuale sviluppata nel capitolo successivo.

4. IPOTESI PROGETTUALE

Il progetto propone un modello sperimentale di valorizzazione digitale del patrimonio culturale rurale, concepito per contesti caratterizzati da proprietà privata, dispersione territoriale e limitata integrazione nei circuiti della fruizione culturale. Tale modello si fonda sull'integrazione di tre componenti operative interdipendenti — archiviazione partecipata, produzione artistica site-specific e visita esperienziale georeferenziata — coordinate all'interno di un assetto di governance pubblico-privato-comunitario. La logica di funzionamento è di tipo processuale e circolare: la produzione di contenuti da parte delle comunità alimenta un archivio digitale strutturato; tale archivio costituisce la base per pratiche di produzione artistica contemporanea; i contenuti generati vengono restituiti al territorio attraverso dispositivi di fruizione situata, capaci di attivare nuove forme di relazione con il patrimonio innescando ulteriori processi partecipativi.
L'obiettivo è attivare forme di valorizzazione culturale che coniughino accessibilità pubblica, partecipazione comunitaria e sostenibilità, evitando interventi invasivi sul patrimonio materiale e promuovendo modalità di visita fondate sulla costruzione di significati condivisi.
La proposta progettuale si articola in una sequenza operativa integrata. La prima fase riguarda la raccolta e organizzazione di contenuti culturali attraverso un archivio digitale partecipato, inteso come dispositivo di raccolta, organizzazione e validazione delle memorie legate alle masserie e al territorio murgiano. Il sistema digitale adottato si basa su Omeka S, configurato come piattaforma per la gestione dell'archivio, integrato con il plugin Collecting per la raccolta partecipata dei contenuti. I materiali — testimonianze orali, fotografie, documenti e descrizioni di pratiche e saperi locali — vengono acquisiti attraverso modalità di contribuzione aperta e successivamente sottoposti a un processo di selezione e validazione.
L'organizzazione dei dati avviene attraverso lo standard Dublin Core, che consente una descrizione coerente e interoperabile delle risorse, mentre l'esportazione tramite protocollo OAI-PMH permette l'integrazione con aggregatori culturali esterni su scala regionale e nazionale. Tale configurazione garantisce non solo l'accessibilità pubblica dei contenuti, ma anche la loro integrazione in ecosistemi informativi più ampi. L'archivio non svolge una funzione meramente conservativa, ma costituisce la base

generativa dell'intero modello, rendendo la comunità locale soggetto attivo nella costruzione del patrimonio narrativo.
La seconda fase è costituita dalla rielaborazione e trasformazione dei dati raccolti in una produzione artistica radicata nelle specificità storiche e sociali del territorio. Questa delicata operazione è affidata ad artisti selezionati dall'Accademia di Belle Arti di Bari, che all'interno di residenze artistiche site-specific, si pongono in relazione diretta con il contesto territoriale e con le comunità locali. L'integrazione di attività laboratoriali e momenti di confronto pubblico contribuisce ad ampliare la partecipazione e a rafforzare il ruolo delle masserie come luoghi attivi di produzione culturale.
La terza fase prevede la restituzione dei contenuti al pubblico mediante percorsi esperienziali georeferenziati, che connettono i diversi siti in un sistema narrativo diffuso e unitario. Questo processo definisce una filiera coerente in cui ciascun elemento svolge una funzione specifica e contribuisce al funzionamento complessivo del modello. I contenuti prodotti nelle residenze vengono associati a specifiche coordinate spaziali e resi accessibili attraverso la piattaforma Echoes, che consente la creazione di itinerari sonori basati su geolocalizzazione. Il sistema attiva automaticamente i contenuti in relazione alla posizione del visitatore, configurando una modalità di fruizione situata in cui il movimento nello spazio fisico attiva livelli interpretativi sonori e narrativi. La possibilità di rendere i contenuti fruibili anche offline risponde alle condizioni infrastrutturali dei contesti rurali.
Il progetto si fonda su un assetto di governance che coinvolge il soggetto proprietario, l'istituzione accademica e le comunità locali, basato sulla separazione e sull'equilibrio delle competenze. Il soggetto privato detiene responsabilità in ambito infrastrutturale ed economico, assicurando l'accessibilità ai siti e la sostenibilità finanziaria del progetto; l'istituzione accademica esercita una funzione di garanzia scientifica e metodologica, attraverso la definizione dei criteri di selezione dei contenuti, la validazione delle produzioni artistiche e la supervisione dell'intero processo; le comunità locali partecipano come soggetti produttivi nella fase di raccolta e come interlocutori attivi nella fase di produzione culturale. La formalizzazione del partenariato avviene attraverso un accordo quadro che stabilisce in modo esplicito i domini decisionali, i criteri di selezione e validazione, le modalità di gestione dei diritti sui contenuti prodotti e i meccanismi di risoluzione dei conflitti. In tale assetto, la governance assume una funzione determinante nel funzionamento del modello, in quanto definisce le condizioni operative per un equilibrio tra interesse pubblico, sostenibilità economica e partecipazione sociale.
Le PPP nel dominio del patrimonio culturale si articolano prevalentemente secondo tre modelli ricorrenti: (a) partenariati orientati alla concessione gestionale, in cui il soggetto privato assume la gestione di beni pubblici verso obiettivi di efficienza ed efficacia economica (Pignatti, 2022; Žuvela et al., 2023); (b) partnership di ricerca a coordinamento pubblico-accademico, come nel caso dei consorzi Horizon 2020 — tra cui CrossCult e rurAllure — in cui reti transnazionali di istituzioni sviluppano strumenti digitali per patrimoni minori, con finanziamento e regia pubblica come condizione strutturale (Kontiza et al., 2020; Lopez Salas, 2021); (c) partenariati di valorizzazione puntuale a patrimonio privato, orientati alla rappresentazione avanzata di singoli siti attraverso tecnologie digitali, come nel caso di Seghetti Panichi, attivati con comunità locali che assumono il ruolo di destinatarie di esperienze progettate esternamente, non di co-produttrici dei contenuti (Clini et al., 2025).
Il soggetto attivante è il privato proprietario, che negozia volontariamente l'apertura del proprio patrimonio verso la sfera pubblica: una configurazione che la letteratura sistematica sui PPP culturali ha fin qui scarsamente tematizzato, privilegiando casi in cui è l'istituzione pubblica a detenere l'iniziativa e il controllo del processo (Žuvela et al., 2023; Macdonald & Cheong, 2014). Il modello integra in un unico processo circolare componenti che nei casi analizzati appaiono disgiunte: l'archiviazione partecipata, la produzione artistica site-specific e la fruizione georeferenziata sono fasi interdipendenti di un'unica filiera culturale. Il contesto applicativo — patrimonio rurale privato diffuso, a bassa visibilità turistica e privo di infrastrutture dedicate — rappresenta un campo ancora scarsamente esplorato dalla ricerca sui PPP culturali, che tende a privilegiare patrimoni urbani, monumentali o già inseriti in circuiti istituzionali consolidati.
Questa interazione — in cui l'interesse privato funge da punto di attivazione di un processo orientato al valore pubblico — si situa all'interno di ciò che Boniotti (2023) definisce public-private-people partnership (P4): una forma di partenariato in cui la dimensione della partecipazione comunitaria è costitutiva del modello, non accessoria.
L'efficacia del modello potrà essere valutata mediante indicatori riferiti alla partecipazione della comunità, alla produzione e qualità dei contenuti, ai livelli di accessibilità pubblica e alla capacità di attivare nel tempo processi di valorizzazione culturale sostenibile.

La proposta, sviluppata per contesti caratterizzati da patrimonio rurale diffuso, proprietà privata e bassa visibilità turistica si basando sull'integrazione tra strumenti digitali, pratiche artistiche e partecipazione comunitaria, può essere applicata anche a territori con caratteristiche analoghe.

5. CONCLUSIONI

Il contributo ha elaborato un quadro concettuale e progettuale entro cui la valorizzazione digitale del patrimonio culturale rurale viene reinterpretata come scelta di campo rispetto al modello di relazione tra patrimonio, comunità e pubblico. La domanda di ricerca non riceve qui una risposta empiricamente validata, ma una proposta strutturata di condizioni operative la cui tenuta dovrà essere verificata nella fase applicativa.
Rispetto al quadro teorico di riferimento, il modello elaborato si situa in continuità con l'interpretazione del patrimonio come risorsa condivisa proposta dalla Convenzione di Faro, ma ne estende le implicazioni operative verso un contesto specifico: quello del patrimonio rurale privato, per il quale il valore culturale trascende il possesso giuridico per divenire elemento della memoria collettiva. In questo scarto tra possesso e appartenenza culturale risiede la tensione costitutiva del modello, e anche la sua sfida principale: costruire forme di accesso pubblico e partecipazione comunitaria che non erodano la disponibilità del soggetto privato a rendere accessibile il proprio patrimonio. La governance tripartita proposta governa questa tensione attraverso la separazione esplicita dei ruoli e la formalizzazione degli accordi.
I casi analizzati convergono nell'indicare che la componente digitale produce valore duraturo quando è integrata in architetture di governance che riconoscono alle comunità un ruolo attivo e non meramente ricettivo. Questo orientamento ha guidato l'intera architettura del modello proposto, in cui l'infrastruttura digitale basata su Omeka S, le residenze site-specific e i percorsi georeferenziati tramite Echoes sono componenti di un sistema in cui la produzione culturale e la fruizione pubblica si alimentano reciprocamente, senza che l'una si esaurisca nell'altra.
È tuttavia necessario riconoscere i limiti che caratterizzano questa proposta nella sua fase attuale. Il modello è elaborato a livello progettuale e non ha ancora ricevuto una validazione empirica: gli indicatori di efficacia individuati — partecipazione comunitaria, qualità dei contenuti, accessibilità pubblica, sostenibilità nel tempo — restano al momento strumenti di valutazione attesi e non misurati. Analogamente, le condizioni di trasferibilità del modello ad altri contesti analoghi, pur plausibili sul piano logico, richiedono verifiche applicative che esulano dall'ambito di questo contributo. Un ulteriore nodo critico riguarda la dipendenza del modello dalla continuità del partenariato: la sostenibilità del processo nel lungo periodo presuppone un allineamento stabile tra le istanze di redditività del privato, il rigore metodologico dell'accademia e i bisogni identitari della comunità.
Le fasi successive della ricerca si orienteranno verso la sperimentazione applicata del modello su uno o più siti del patrimonio Renewia, con l'obiettivo di verificare la tenuta delle ipotesi progettuali nel confronto con le complessità operative del contesto reale. In parallelo, sarà necessario approfondire la dimensione della misurazione dell'impatto culturale e sociale, integrando strumenti di valutazione qualitativa e quantitativa capaci di restituire la complessità dei processi partecipativi previsti dal modello. Più in generale, il contributo invita a considerare il patrimonio rurale privato non come un settore residuale della politica culturale, ma come un laboratorio in cui sperimentare forme innovative di partenariato e di costruzione condivisa del significato culturale — un ambito in cui la ricerca accademica può svolgere un ruolo di mediazione tanto metodologica quanto etica.

BIBLIOGRAFIA

Boniotti, C. (2023). The public–private–people partnership (P4) for cultural heritage management purposes. Journal of Cultural Heritage Management and Sustainable Development, 13(1), 1–14. https://doi.org/10.1108/JCHMSD-12-2020-0186

Candy, L., Edmonds, E. (Eds.). (2011). *Interacting: Art, Research and the Creative Practitioner*. Quest Publishing.

Clini, P., Angeloni, R., D'Alessio, M., Coppetta, L., & Galli, I. (2025). Digital representation and AI-driven virtual experience for historic houses. The case study of Borgo Storico Seghetti Panichi. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLVIII-M-9-2025, 337–344.

Council of Europe. (2005). Framework Convention on the Value of Cultural Heritage for Society (Faro Convention).

Dümcke, C., Gnedovsky, M. (2013). The social and economic value of cultural heritage: Literature review. EENC.

Kontiza, K., Antoniou, A., Daif, A., Reboreda-Morillo, S., Bassani, M., González-Soutelo, S., Lykourentzou, I., Jones, C. E., Padfield, J., & López-Nores, M. (2020). On how technology-powered storytelling can contribute to cultural heritage sustainability across multiple venues — Evidence from the CrossCult H2020 Project. Sustainability, 12(4), 1666. https://doi.org/10.3390/su12041666.

Macdonald, S., Cheong, C. (2014). The Role of Public-Private Partnerships and the Third Sector in Conserving Heritage Buildings, Sites, and Historic Urban Areas. Los Angeles, CA: Getty Conservation Institute. http://hdl.handle.net/10020/gci_pubs/public_private_partnerships

López Salas, E. (2021). A collection of narrative practices on cultural heritage with innovative technologies and creative strategies. Open research Europe, *1*, 130. https://doi.org/10.12688/openreseurope.14178.1

Maietti, F. (2023). Heritage Enhancement through Digital Tools for Sustainable Fruition—A Conceptual Framework. Sustainability, 15(15), 11799. https://doi.org/10.3390/su151511799

Marsili G. (2024). Digital strategies for enhancing cultural heritage: the Villa del Casale of Piazza Armerina Project, from legacy data to digital ecosystem, Archeologia e Calcolatori, 35(2), 445-454. https://doi.org/10.19282/ac.35.2.2024.46

Pignatti, M. (2022). I modelli di partenariato pubblico-privato nella gestione e valorizzazione dei beni culturali come strumento per la creazione di ecosistemi innovativi e di sviluppo economico e sociale. DPCE Online, *1*, 1–24. https://doi.org/10.57660/dpceonline.2022.1563

Ruggiero, G., Parlavecchia, M., & Dal Sasso, P. (2019). Typological characterisation and territorial distribution of traditional rural buildings in the Apulian territory (Italy). Journal of Cultural Heritage, 39, 278-287. https://doi.org/10.1016/j.culher.2019.04.010

Žuvela, A., Šveb Dragija, M., & Jelinčić, D. A. (2023). Partnerships in Heritage Governance and Management: Review Study of Public–Civil, Public–Private and Public–Private–Community Partnerships. Heritage, 6(10), 6862-6880. https://doi.org/10.3390/heritage6100358

I Vangeli digitali e accessibili, per un'editoria religiosa universale e senza barriere

GIULIA D'ARCANGELO

DOTTORATO IN SCIENZE DEL PATRIMONIO CULTURALE,
CURRICULUM DI EDITORIA E INNOVAZIONE
UNIVERSITÀ DEGLI STUDI DI ROMA "TOR VERGATA"-
IL PORTICO GRUPPO EDITORIALE BOLOGNA - EDB EDIZIONI DEHONIANE BOLOGNA, ITALY
GIULIA-DARCANGELO@VIRGILIO.IT

ABSTRACT (ITALIANO)

Il progetto sviluppa un modello EPUB 3 dei Vangeli, accessibile e navigabile, che valorizza la complessità testuale del testo biblico e ne garantisce la fruizione anche a persone con disabilità visive. Fondato su marcatura semantica coerente e organizzazione rigorosa del contenuto, il modello rappresenta un riferimento editoriale replicabile e adattabile a edizioni bibliche in altre lingue e contesti culturali.
Parole chiave: EPUB 3; Vangeli; accessibilità; tag semantici; replicabilità

ABSTRACT (ENGLISH)

Digital and Accessible Gospels: Towards a Universal and Barrier-free Religious Publishing
The project develops an accessible, navigable EPUB 3 model of the Gospels, enhancing the textual complexity of the Bible while ensuring accessibility for users with visual impairments. Based on coherent semantic markup and rigorous content structuring, the model serves as a replicable editorial reference adaptable to Bible editions in other languages and cultural contexts.
Keywords: EPUB 3; Gospels; accessibility; semantic tags; replicability

1. IL PROGETTO

La natura e la struttura complesse di un testo come quello biblico non hanno fatto sì che, nel processo attuale di digitalizzazione del cartaceo, venisse escluso, anzi ne hanno garantito una posizione privilegiata e di forte interesse. Nell'ambito del percorso dottorale in Editoria e Innovazione, curriculum inserito nel Dottorato in Scienze del Patrimonio Culturale, si è delineata l'urgenza di focalizzarsi sulla comunicabilità digitale e sull'accessibilità di una fetta di editoria talvolta marginalizzata: quella religiosa e spirituale. Con una rinnovata consapevolezza della potenza dello strumento "eBook", ma anche e soprattutto dell'elevato numero di persone che convivono con disabilità visive,[1] più o meno invalidanti, ha preso avvio un processo volto a realizzare un modello pienamente navigabile e accessibile della Bibbia, in particolare della sezione dei Vangeli nel Nuovo Testamento. I Vangeli in questione sono stati inizialmente tratti dall'edizione graficamente rinnovata del 2024 de *La Bibbia di Gerusalemme*, edita da EDB del gruppo editoriale Il Portico, i cui testi riproducono fedelmente la versione italiana e liturgica curata dalla CEI, Conferenza

[1] World Health Organization, *Disability and health*, 7 March 2023, https://www.who.int/news-room/fact-sheets/detail/disability-and-health.

episcopale italiana, elaborata a partire dagli originali in ebraico, aramaico e greco.[2] In seguito, si è scelto di adottare la nuova versione di questi, tratta da *Evangeli e Salmi,*[3] anticipazione della rinnovata edizione de *La Bibbia di Gerusalemme*, la cui pubblicazione non ha ancora data definita.

2. IL MODELLO ACCESSIBILE

La configurazione del modello si sviluppa in modo da riprodurre, in parte, la struttura originaria dei testi evangelici, senza comprometterne in maniera indebita l'impianto, ma con lo scopo di agevolarne la lettura e la consultazione; una mirata progettazione dei codici e l'utilizzo di tag semantici sono pensati per garantire la compatibilità con i principali strumenti di tecnologia assistiva, quali gli *screen reader* o i software in grado di ingrandire lo schermo. L'architettura canonica di una pagina del testo biblico è la seguente (vd. Fig.1):



Gal 6,2; Gc 1,25 **La nuova giustizia superiore all'antica**

Lv 19,15s; Rm 10,3; Fil 3,9 **20**Io vi dico infatti: se la vostra giustizia non supererà quella degli scribi e dei farisei, non entrerete nel regno dei cieli.

Gen 9,6 Tg; Es 20,13; Dt 5,17 **21**Avete inteso che fu detto agli antichi: *Non ucciderai*; chi avrà ucciso dovrà essere sottoposto al giudizio. **22**Ma io vi dico: chiunque si adira con il proprio fratello dovrà essere sottoposto al giudizio. Chi poi dice al fratello: "Stupido", dovrà essere sottoposto al sinedrio; e chi gli dice: "Pazzo", sarà destinato al fuoco della Geènna.

Sir 10,6; Ef 4,26; Gc 1,19-20 Dt 17,8-13

1Gv 3,15; Gc 3,6; Mt 3,12+ Mc 11,25 **23**Se dunque tu presenti la tua offerta all'altare e lì ti ricordi che

5,15 Nell'antichità il *moggio* era un mobiletto a tre o quattro piedi. Qui si tratterebbe dunque di nascondere la *lampada* sotto questo mobiletto, un po' come sotto il letto di Mc 4,21p, non di spegnerla coprendola con un moggio moderno.

5,17 Gesù non viene né a distruggere la Legge (Dt 4,8+) e tutta l'economia antica, né a consacrarla come intangibile, ma a darle, con il suo comportamento, forma nuova e definitiva, dove si realizza nella pienezza ciò verso cui la Legge stessa era avviata. Ciò si applica in particolare alla «giustizia» (v 20, cf. 3,15; Lv 19,15; Rm 1,16+), «giustizia perfetta» (v 48), di cui le sentenze 40p), diviene il comandamento nuovo e inesauribile di Gesù (Gv 13,34) e compie tutta la legge (Rm 13,8-10; Gal 5,14; cf. Col 3,14+).

5,18 Introducendo con «Amen» (= *In verità*, cf. Sal 41,14+ e Rm 1,25+) alcune sue parole, Gesù ne sottolinea l'autorità (6,2.5.16, ecc.; Gv 1,51, ecc.). La parola ebraica che significava in origine la stabilità si è sviluppata in due direzioni: quella di verità e quella di fedeltà. Qui significa semplicemente «in verità». – *un solo iota o un solo trattino della Legge*: lett.: «non uno iota, non un piccolo tratto»; BJ traduce: «non una i, non un punto sull'».

Figura 1. Pagina della Bibbia

A prima vista, anche per un lettore normovedente potrebbe risultare complesso districarsi tra i versetti numerati, i rimandi, l'apparato di note e di commenti, e le abbreviazioni e i simboli che accompagnano il testo biblico. Trasporre, pertanto, i Vangeli in una versione digitale accessibile rappresenta un potenziamento funzionale necessario e atteso: garantirà, infatti, a un bacino sempre maggiore di "lettori", in senso lato, di fruire del testo. L'agevolazione riguarda anche persone in età avanzata, la cui capacità visiva sia ridotta o danneggiata, o quanti abbiano un temporaneo impedimento (si pensi a un incidente, a una operazione e così via). La possibilità di approcciarsi in maniera naturale e con un'assistenza guidata è l'obiettivo del progetto.
L'impianto, sviluppato in collaborazione con la casa editrice EDB e già concretamente avviato, è tuttora in fase di definizione e perfezionamento, e si articola secondo lo schema che segue:

- I titoli e i titoletti sono definiti tramite i tag semantici <h1>, <h2>, <h3> e <h4>, rispettando la gerarchia originaria del testo e assicurando al fruitore di navigare con sicurezza tra un capitolo/sezione e l'altro/a;

[2] *La Bibbia di Gerusalemme* (2024). Edizioni Dehoniane Bologna (EDB). ISBN 9788810821466.

[3] Bigoni, L., Mambelli, A., & Mariotti, G. (Eds.). (2025). *Evangeli e Salmi. Edizione illustrata. La Bibbia di Gerusalemme* (pp. 1–569). Edizioni Dehoniane Bologna (EDB). ISBN 978-8810038987.

- I versetti sono marcati tramite elementi <p> con classe "versetto", ogni versetto è corredato da un numero distinguibile e marcato in <sup> ma non linkabile. Si farà eccezione per quei numeri a cui si leghi una nota a piè di pagina: in questo caso, esso va in <sup> contenuto in <a>, con i seguenti attributi: href="Apparato_note_Mt.xhtml#notaN", epub:type="noteref", id="notaN";
- I testi delle note, contenute nel file xhtml dedicato, sono marcate dal tag semantico <aside>, con rispettivi attributi, tra cui quello accessibile role="doc-backlink". Quest'ultimo è contenuto nel modulo WAI-ARIA Digital Publishing (DPUB-ARIA)[4] e rappresenta un link necessario per spostarsi dal contenuto secondario (nota) al punto esatto del testo da cui è stato richiamato.
- I rimandi sono indicati nel corpo del testo dal simbolo del rombo (✦), racchiuso nei tag e <a>, e sono caratterizzati dall'attributo del backlink nel file xhtml che contiene l'insieme dei rimandi.

Si ritiene necessario soffermarsi sulla questione delle note e dei rimandi, nodi critici e nevralgici del progetto:

- Note→ come anticipato, il numero di versetto a cui è legata una nota viene reso linkabile nel corpo del testo, grazie ai tag <a> e <sup>. In una versione precedente si era ipotizzato di assegnare agli attributi id delle note la numerazione corrispondente a quella dell'edizione cartacea (1,18 – 1,19 – 2,1 – 2,15 e via dicendo); tuttavia, poiché la sintassi dell'attributo id non prevede l'impiego della virgola, si è adottata una numerazione progressiva standard (vedi nota5 nell'esempio), mantenendo il numero del cartaceo (1,18) nel contenuto testuale dell'elemento <a>:

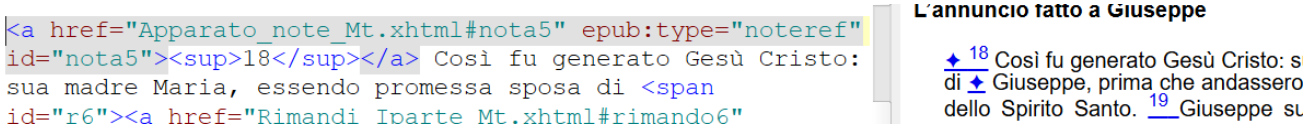


Figura 2. Sezione del corpo di testo del Vangelo digitale

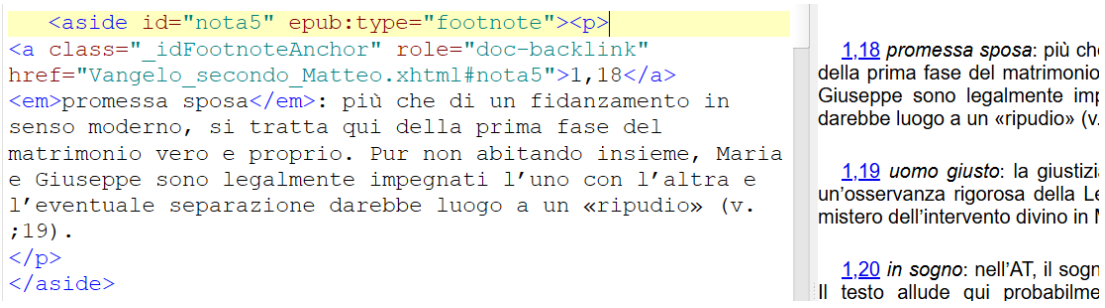


Figura 3. Sezione dell'apparato di note del Vangelo digitale

[4] Digital Publishing WAI-ARIA Module 1.1, https://www.w3.org/TR/dpub-aria-1.1/.

- Rimandi (riferimenti marginali)→ il formato EPUB, per sua natura *reflowable*, non consente di mantenere nel digitale la stessa struttura della versione cartacea del Vangelo. Il testo, infatti, si adatta dinamicamente al dispositivo e alle preferenze di lettura, rendendo instabile l'ancoraggio dei numerosi rimandi ai rispettivi versetti. Pertanto, in sostituzione di una disposizione fissa poco funzionale, si adottano dei collegamenti interni, gestiti tramite l'utilizzo di un simbolo: il rombo nero, reso accessibile tramite degli attributi con un forte valore semantico (vedi fig. 4). Anche solo sfiorando il simbolo, comparirà un tooltip contestuale con le informazioni del rimando in questione.

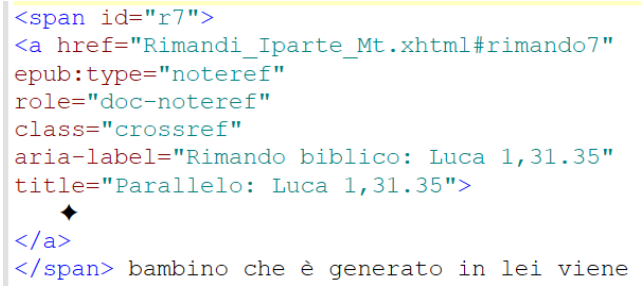


Figura 4. Attributi accessibili

Nel link di rimando, nel codice nel file xhtml dei riferimenti marginali, il contenitore contiene lo stesso id, il contenuto testuale del rimando (per esempio Lc 1,31.35) e il backlink (freccetta) che riporta esattamente al simbolo del rombo e quindi alla posizione iniziale (vedi fig. 5).

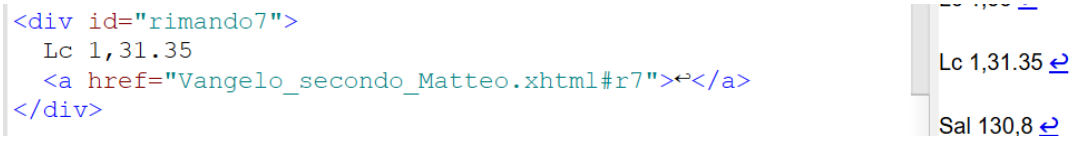


Figura 5. Riferimenti marginali

In una seconda fase si ipotizza di raggiungere un grado di completezza di navigabilità rendendo linkabile anche il riferimento stesso (vedi a destra della fig. 5). In altre parole: se le parole del versetto «Genealogia di Gesù» sono riscontrabili nel rimando Lc 3,23-28, cliccando sul simbolo del rombo, non solo andremmo sul rimando Lc 3,23-28, ma avremmo la possibilità di muoverci da Lc 3,23-28 ai versetti specifici in questione nel Vangelo di Luca, «Genealogia di Gesù – Gesù quando cominciava il suo…». In questa prima fase del progetto dottorale, tuttavia, si lavorerà solo sui quattro vangeli, pertanto qualora il rimando fosse a tutti gli altri libri, i versetti non sarebbero presenti quindi né rintracciabili né linkabili.

Nella versione cartacea di *Evangeli e Salmi*, i testi sono arricchiti da illustrazioni tratte da fonti varie, nella maggior parte dei casi si tratta di evangelari conservati nelle biblioteche o nei musei più prestigiosi del mondo, come il Metropolitan Museum di New York o la Bayerische Staatsbibliothek di Monaco. Nel formato EPUB 3, ogni immagine è preservata e rielaborata attraverso l'associazione a ciascuna di un testo alternativo (alt-text): esso consente ai lettori di schermo e ad altre tecnologie assistive di restituire verbalmente le informazioni veicolate dall'immagine (vedi Fig.6), in assenza della percezione grafica

diretta. Tramite l'adozione di una corretta marcatura semantica, le immagini non saranno elementi accessori, ma componenti strutturali del testo evangelico: il legame tra apparato iconografico e contenuto testuale viene così salvaguardato, svelando ancora una volta l'efficacia e la potenza del formato EPUB 3.[5]

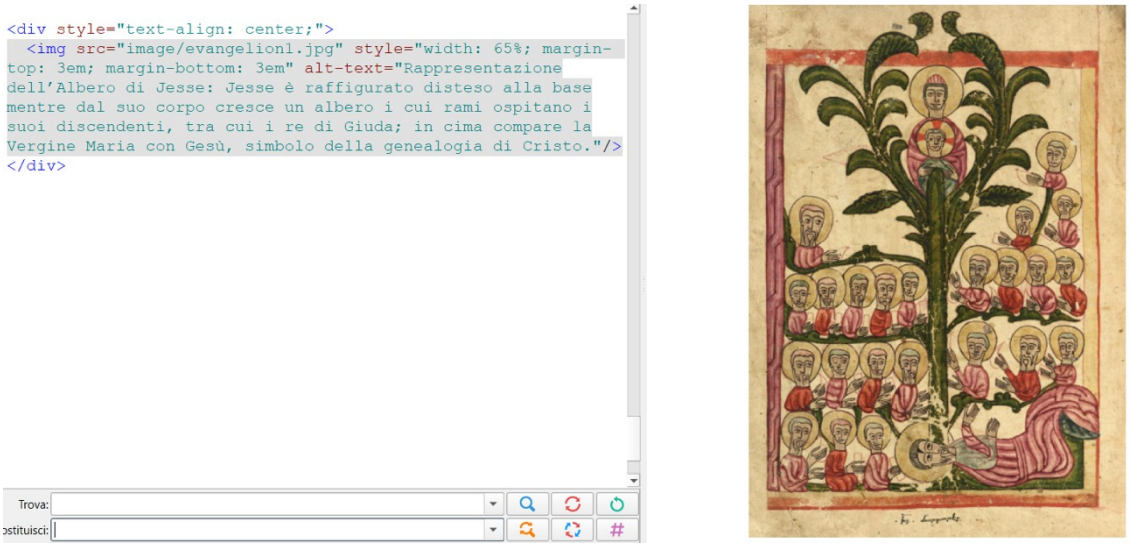


Figura 6. Alt-text dell'immagine

Operare in questa direzione significa partecipare attivamente alle dinamiche europee, rispettando le direttive stabilite dall'EAA 2025,[6] e allineandosi perciò a una visione che riconosce l'accessibilità come valore strutturale di fondo. In quest'ottica, lavorare alla digitalizzazione e all'accessibilità dei prodotti diventa un impegno di cittadinanza attiva, volto a favorire una partecipazione sempre più ampia alla vita culturale e sociale, e a generare un ambiente digitale in cui tutti siano i benvenuti.
Sebbene sia intuibile la complessità del progetto e i tempi lunghi necessari per il suo completamento, si è ritenuto opportuno adottare un'impostazione in formato EPUB, nella sua già citata natura *reflowable*, in quanto rappresenta una soluzione particolarmente efficace: essa, infatti, non si limita a riprodurre un'impaginazione statica, ma consente una continua riorganizzazione, garantendo dinamicità e adattabilità. Un altro sistema accessibile per un utente non vedente, quale il braille, configurerebbe invece un ulteriore oggetto-libro da fruire in modo sequenziale e con limitata libertà di navigazione.
Analogamente, altri formati diffusi per la fruizione digitale, come i formati proprietari, sebbene in grado di integrare dei livelli di interattività, risultano chiusi e vincolati a ecosistemi definiti e controllati da specifiche aziende, oltre a comportare costi elevati. L'EPUB, al contrario, è pensato per essere uno standard aperto, universale, flessibile.

[5] World Wide Web Consortium. (2025, 13 marzo). *EPUB 3 Overview*. W3C. https://www.w3.org/TR/epub-overview-33/.
[6] European Commission. (2025, 27 giugno). *The EU becomes more accessible for all*. Shaping Europe's digital future. https://digital-strategy.ec.europa.eu/en/news/eu-becomes-more-accessible-all/

3. LE PROSPETTIVE FUTURE

Il progetto propone di definire un modello editoriale di riferimento, concepito per essere applicabile a livello internazionale. L'EPUB in questione è fondato su principi di progettazione indipendenti dalla lingua, basati, come spiegato nel paragrafo precedente, su una certa organizzazione del testo, su una marcatura semantica coerente e sulla veste accessibile prevista dalle normative europee. Con queste caratteristiche, il modello sarà pienamente spendibile, adattabile alla realizzazione di edizioni bibliche in lingue diverse, senza comprometterne il contenuto e la qualità della "lettura". Perché la sua diffusione sia ottimale, non sarà esclusa la possibilità di corredare il modello da linee guida.
In questo panorama, il modello tenta di prevedere e di anticipare anche altre esigenze del lettore, mediante strumenti di personalizzazione della lettura: non solo tag semantici per le disabilità visive gravi, ma anche font sans-serif idonei a utenti con dislessia, "pagine" con alto contrasto cromatico per ipovedenti o daltonici, e ancora elementi di tipo multimediale quali audio integrati con marcatori semantici, che offrano la corretta pronuncia di nomi e luoghi biblici complessi.
In questa prospettiva, il modello digitale delineato costituisce un intervento critico nei confronti dell'abilismo che struttura ancora numerose pratiche editoriali digitali. Un lettore normovedente e neurotipico spesso, e implicitamente, produce delle forme di esclusione talvolta invisibili, che relegano la disabilità a una condizione marginale. Ma progettare in questo senso significa riconoscere l'accessibilità come principio fondativo e non come un optional o come un'aggiunta accessoria.

BIBLIOGRAFIA

Bigoni, L., Mambelli, A., & Mariotti, G. (Eds.). (2025). *Evangeli e Salmi. Edizione illustrata. La Bibbia di Gerusalemme* (pp. 1–569). Edizioni Dehoniane Bologna (EDB). ISBN 978-8810038987.

Digital Publishing WAI-ARIA Module 1.1, https://www.w3.org/TR/dpub-aria-1.1/.

European Commission. (2025, 27 giugno). *The EU becomes more accessible for all*. Shaping Europe's digital future. https://digital-strategy.ec.europa.eu/en/news/eu-becomes-more-accessible-all/

La Bibbia di Gerusalemme (2024). Edizioni Dehoniane Bologna (EDB). ISBN 9788810821466.

World Health Organization, *Disability and health*, 7 March 2023, https://www.who.int/news-room/fact-sheets/detail/disability-and-health.

World Wide Web Consortium. (2025, 13 marzo). *EPUB 3 Overview*. W3C. https://www.w3.org/TR/epub-overview-33/.

Implementing a controlled vocabulary for medieval administrative acts: theoretical and methodological considerations

CHIARA DE BASTIANI

UNIVERSITÀ CA' FOSCARI VENEZIA, ITALY –
CHIARA.DEBASTIANI@UNIVE.IT

ABSTRACT (ENGLISH)

This paper presents the theoretical and methodological considerations on the implementation of a controlled vocabulary for medieval administrative acts expressed in imperial diplomas. The controlled vocabulary constitutes an expansion of the ontological module for the bibliographic description of medieval charters related to Germanic emperors, OntoVE WrittenRecords, described in De Bastiani (2023, 2025). It is being conceived to overcome the limitations of the so far implemented bibliographic module. In fact, the bibliographic module currently in use models the acts as events but does not allow for finer grained inquiries on the data. While the controlled vocabulary is still under construction, this paper presents the theoretical and methodological underpinnings for its implementation, starting from the literature on ontologies and the semantic modelling of medieval charters.
Keywords: Linked Open Data; Semantic Web; Ontologies; bibliographic ontology; controlled vocabularies; imperial charters.

ABSTRACT (ITALIANO)

Implementazione di un vocabolario controllato per la descrizione degli atti amministrativi medievali: considerazioni teoriche e metodologiche.
Il contributo presenta le considerazioni teoriche e metodologiche alla base dell'implementazione di un vocabolario controllato per gli atti amministrativi medievali espressi attraverso diplomi imperiali. Il vocabolario costituisce un'espansione del modulo ontologico per la descrizione bibliografica di diplomi imperiali relativi a imperatori germanici, OntoVE_WrittenRecords, descritto in De Bastiani (2023, 2025). La concezione del vocabolario controllato è necessaria per superare i limiti del modulo bibliografico finora implementato. Infatti, il modulo bibliografico attualmente in uso modella gli atti come eventi, ma non permette ricerche più granulari sui dati. Mentre il vocabolario controllato è ancora in fase di costruzione, questo contributo presenta le basi teoriche e metodologiche per la sua implementazione, a partire dalla letteratura sulle ontologie e la modellazione semantica di diplomi imperiali.
Parole chiave: Linked Open Data; web semantico; ontologie; ontologia bibliografica; vocabolari controllati; diplomi imperiali.

1. INTRODUCTION AND RESEARCH BACKGROUND

The present paper is embedded within the OntoVE project, now concluded, which was concerned with the modelling of ontological descriptions for archaeological and bibliographic items of Germanic origin related to the Veneto region (see De Bastiani 2023, 2024, 2025; De Bastiani & Fabbris 2023, 2025).[1] One of the goals of the project was to aggregate metadata about the relevant artefacts and describe them according to interoperable standards, in order to allow for queries into the relationship between Germanic populations and the history of the region. Among the project's outcomes are two ontological modules,[2]

[1] The OntoVE ontology specification can be consulted here: https://w3id.org/ontove/ontology (last accessed 21st April 2026), while the project's UI is visible at this link: https://ontove-ui.chitontove.it/en (last accessed 21st April 2026). The source code for the UI is published here: https://zenodo.org/records/14551333 (last accessed 21st April 2026).

[2] The two ontological modules are now merged into one ontology but could be also used separately (see De Bastiani 2025: 146). The OntoVE_Archeo module can be found here: https://github.com/ChiDeBa/ontove_archeo (last accessed 21st April 2026), while the OntoVE_WrittenRecords module can be found here: https://github.com/ChiDeBa/ontove_writtenrecords (last accessed 21st April 2026).

which reuse the CIDOC-CRM model, v. 7.1.2 (Bekiari et al., 2022) and the ArCo ontologies (Carriero et al., 2019) for the description of archaeological artefacts (OntoVE_Archeo), and LMRoo v. 0.9.6 (Bekiari et al., 2023) and Bibframe[3] for the description of bibliographic records (OntoVE_WrittenRecords). The latter constitutes the starting point for the implementation of a controlled vocabulary and the relative methodological and theoretical reflections presented in this paper. In the following, the main architecture of the OntoVE_WrittenRecords module is summarized, in order to situate the implementation within the project's context. The full description of the OntoVE_WrittenRecords module can be consulted in De Bastiani (2025).

As stated in De Bastiani (2023, 2025), given that the goal of the project was to highlight the history of the Veneto region with respect to its relationship with Germanic populations, among the first bibliographic items collected were imperial charters written after the fall of the Lombard Kingdom of Italy. The modelling of the charters started from the *Digital Monumenta Germaniae Historica* editions,[4] which convey different relevant data about the charters, e.g. the place and date of issuance, and a summary of the main contents, together with an edition of the charters and information about possible interpolations or forgeries. The approach undertaken was that of translating these data into an ontological model that captures both the complexity of the written transmission, as well as the contents of the texts (De Bastiani 2025: 147-148).

The bibliographic ontological module devised within the OntoVE project responds to these requirements by aligning the LRMoo 0.9.6 model to the Bibframe model. As argued in De Bastiani (2025: 152-153), the WEMI (Work, Expression, Manifestation, Item) scheme, employed in LRMoo, is in fact a powerful conceptualization of the different steps in the creation of an intellectual work, and allows modelling the relationship between the content of a work and its different embodiments. The Bibframe model provides finer grained classes that complement the more abstract classes and properties devised within the LRMoo model.[5] These two ontologies allow describing the transmission of a given work and its contents, the latter constituted also by the administrative acts promulgated with the charters. In fact, the alignment between the LRMoo *F1_Work* and *F2_Expression* classes with Bibframe's *Work* class[6] allows conceptualizing the specific acts mentioned in the charters as subjects of them, via the property *Bibframe:Subject*, together with the places and people mentioned (De Bastiani 2025: 153-154). The modelling choices enable therefore the formulation of SPARQL queries, locating for example which areas are referred to in the documents (De Bastiani 2025: 158), and are reflected in the User Interface (UI) devised for the project, which reuses the Sampo model (Rantala et al, 2023). The UI presents for example a *Texts* perspective, where the subjects of a specific document are given and allows refining the search through pre-defined filters (De Bastiani 2025: 159-160).

The modelling of the acts described in the charters follows from the LRMoo model and its inherent alignment with CIDOC-CRM. As shown in De Bastiani (2025: 155), these are instances of *E4_Period*; they constitute however an area for improvement. Whereas these instances are linked to the agents and participants in a given interaction through the CIDOC-CRM properties *E14_carried_out_by* and *E11_had_participant* (De Bastiani 2025: 155), the semantics of the event itself are opaque for further SPARQL processing. In fact, the label assigned to the *E4_Period* individuals is a string in natural language, describing the whole event (for example: "Charlemagne allows Aio to divide his properties among his sons"),[7] but the specific acts are not further broken down in units assigned to a defined taxonomy, which would be more easily retrievable through a SPARQL query and give e.g. quantitative insights into the types of acts involved. Moreover, also behind the different search perspectives and search filters provided in the project's UI are SPARQL queries (see De Bastiani 2025; De Bastiani & Fabbris 2025). This also means that the UI cannot, at the moment, provide a filter function to isolate the

---

[3] https://www.loc.gov/bibframe/ (last accessed 21st April 2026).

[4] https://www.dmgh.de/ (last accessed 21st April 2026).

[5] Note moreover that, besides the *Bibframe:Subject* property mentioned in this section, which connects the content of a work with the events, people and places mentioned, the ontology also foresees specific properties to make explicit the connection between events and the works citing them (cf. De Bastiani 2025: 153 and https://w3id.org/ontove/ontology#citesEvent, last accessed 21st April 2026), thereby also allowing to identify the source for a statement. This would provide more information on the provenance and therefore reliability of the statement, a feature that is also relevant in research on historical documents (Julien 2024: 24).

[6] The Bibframe *Work* class is equated to LRMoo's *F1* OR *F2* classes (De Bastiani 2025: 153).

[7] See https://w3id.org/ontove/ontology#ecb1d6f0-00ed-4ed2-946b-272fa6a394f1 (last accessed 21st April 2026).

types of acts promulgated, since these are not appropriately modelled. To overcome these limits, I propose the modelling of a controlled vocabulary to be integrated in the bibliographic ontological module.

2. ONTOLOGIES AND SEMANTIC MODELS FOR MEDIEVAL CHARTERS

Vogeler (2019: 131) defines charters as "written documents of legal acts in their historical development". According to Vogeler, the notion of historical development is also linked to their transmission history and to the fact that due to different diplomatic practices, charters exist in different versions. Moreover, in diplomatics the concept of charter also represents the legal or administrative fact recorded, carried out or witnessed. Therefore, a charter is a written document that may exist in different versions (included the digital versions as result of several digitization projects, Vogeler 2019: 128) and that represents a set of actions about a legal or administrative fact. Vogeler notices that the FRBR model can help conceptualize the abstract content of a charter and its different embodiments through different expressions and manifestations. Moreover, Vogeler (2019) notices that there have been scholarly initiatives to model the contents of the acts represented in the charters. The *Making of Charlemagne's Europe* project[8] for example proposes a factoid model to capture the prosopographic data contained in the charters. A factoid represents what a document asserts and represents the relationships between points or periods in time, people, and places and these assertions are linked to a particular source (Bradley et al., 2019: 4). In order to capture the acts expressed by the charters, the *Making of Charlemagne's Europe* project modelled transaction factoids.[9] This relational model makes explicit the links between actors and places and conceptualizes the contents of the charters as different kinds of transactions described by a document. As stated by the database project members, however, "interconnectivity" between their model and others is certainly a potential advantage for the scholarly community, but was not a goal striven to within the project, also due to the nature of the facts to be described.[10] This makes the model conceptually appealing, but hard to align to existing ontological models. However, Pasin & Bradley (2015) decomposed their factoid model into expressions of entities and relations that could be mapped to the CIDOC-CRM Model. In fact, the model already foresees the concepts of agent that can participate in an activity, and this relation is made explicit through specific properties; these can be equated to the elements and relations expressed in CIDOC-CRM.

CIDOC-CRM has been moreover adopted by Sippl et al. (2021) to enrich digital scholarly editions of the charters held by the Katharinenspital in Regensburg with a graph-based approach to represent their contents. In their proposal, a Charter is modelled as an instance of *E5_Event*, which is then broken down into instances of an *E7_Activity* involving different *E39_Actor* instances. The activities are labelled with the verbs (such as *verkaufen*, 'to sell') used in the document and extracted through NLP pipelines (Sippl et al. 2021: 190). Since the project uses Neo4j database, they can be retrieved through Neo4J's graph query language.

Finally, also Álvarez-Fidalgo et al. (2025) adopt CIDOC-CRM and merge it with the Wikidata data model to describe the assertions made in medieval charters. Also this project focuses on the contents of the documents, which are reflected in the event-centric CIDOC-CRM model, but merges the description with Wikidata concepts, that represent shorter paths in the representation of genealogical and sociological information (Álvarez-Fidalgo et al., 2025:3). Social roles, social descriptions and social individuals are modelled by Garbacz et al. (2010) in their ontology for Ecclesiastical administrative acts, based upon the DOLCE foundational ontology. This ontology highlights social and political dimension of the acts carried out, but its potential mapping with CIDOC-CRM is not straightforward (Garbacz et al., 2010: 11).

Concluding this section, the models and theoretical considerations briefly summarized here show on the one hand how charters can be conceptualized as documents that exist in several versions, and how the facts they describe can be broken down into ontological classes within an even-centric approach. The proposal by Sippl et al. (2021) sees the charter as an instance of an event. As noted by Vogeler (2019), in fact, a charter can be both understood as a document and as the legal act being described. While the even-centric approach proposed in CIDOC-CRM allows capturing the underlying semantics and

[8] https://charlemagneseurope.ac.uk/ (last accessed 21st April 2026).

[9] https://charlemagneseurope.ac.uk/blog/building-a-charter-database-1-the-factoid-model-and-its-discontents/ (last accessed 21st April 2026).

[10] https://charlemagneseurope.ac.uk/blog/interconnectivity-of-databases-and-charter-data/ (last accessed 21st April 2026).

relationships between agents and events, and what is being predicated about them, Álvarez-Fidalgo et al. (2025)'s modelling highlights also how CIDOC-CRM could be complemented with the use of other data models, such as Wikidata, to further define the social roles involved, an aspect also considered in the SacrOnt ontology by Garbacz et al. (2010). Moreover, Pasin & Bradley (2015) explore the implementation of an ontology about 'feudalism', basing on their experience in several projects exploiting the factoid model, and considering the more powerful description possibilities offered by the Semantic Web. They note that while CIDOC-CRM and the factoid models provide the relevant concepts, the requirements of the single projects, and the different historical periods they referred to, required different extensions to the model. Their approach however shows that a finer-grained classification can be potentially grafted onto the event-centric stock provided by CIDOC-CRM, while preserving richer semantics for the categorization of several acts reported in the documents of interest.

3. CONCEPTUALIZATION OF THE CONTROLLED VOCABULARY

As it emerges from the discussion illustrated in Section 2, the OntoVE_WrittenRecords ontological module already foresees the necessary ingredients to refine the conceptualization of the acts promulgated in the charters. In fact, while the LRMoo classes and properties allow accounting for the different versions of the charters, the Bibframe *Subject* property allows asserting that a specific event is the subject of a charter, as stated in Section 1 and De Bastiani (2023, 2025). Therefore, in our conceptualization, the charter is seen as an abstract work, instantiated in a specific manifestation (or more than one), which is about a legal or administrative act, in line with the literature reviewed in Section 2. Since the LRMoo ontology is already aligned with CIDOC-CRM, the conceptualization proposed allows to exploit CIDOC-CRM's event-centric approach (see also De Bastiani 2023, 2025), thereby also providing the first steps to enrich the bibliographic model with a first set of data relevant for historical research.[11] In fact, as mentioned in Section 1, the current modelling of the events at the present state of the ontology consists in assigning a label to an instance of *E4_Period* which corresponds to a description in natural language of the event. However, as was noted in Section 1, this kind of modelling does not make the type of acts further searchable and would thus not allow to uncover more precise connections between the acts promulgated, the actors involved and their representation in the charters; a strategy similar to the one already available in the OntoVE_WrittenRecords module is represented in Sippl et al. (2021), where the acts are further broken down, and qualified through the verbs depicting the action. While such an approach would not hinder searchability with SPARQL, albeit requiring strategies for the extraction of literals, I argue that a controlled taxonomy would allow isolating relevant acts in a more efficient way. The question now arises as to whether it is preferable to align the bibliographic ontology with an ontology for medieval legal or administrative acts, or whether it is preferable to use CIDOC-CRM's *E55_Type* class as an interface. For the general architecture of the OntoVE project, it was decided to align the reused ontologies to the CIDOC-CRM and LRMoo models. This was done to maximize their interoperability and to permit the inheritance of properties from the finer-grained classes of the aligned ontologies (see De Bastiani 2023, 2024, 2025). However, for the implementation at hand, the underlying semantics are already given by the CIDOC-CRM model, and the intended implementation has the goal of providing for an established terminology to further qualify the instances of an *E4_Period* individual. Therefore, the implementation of the taxonomy is currently being carried out through the *E55_Type* class, which is intended as an interface to model controlled vocabularies (Bekiari et al., 2022: 42-43). The definitions for the controlled vocabulary will be based upon previous works defining the nature of the acts, such as the *Making of Charlemagne's Europe* project. Figure 1 depicts a first proposal of the modelling of the controlled vocabulary with Protégé (Musen, 2015); the vocabulary items constitute the individuals belonging to the class *E55.1_Act_Type* and are described through an rdfs:comment:

[11] One of the reviewers notes in fact that the taxonomy proposed here would offer an improvement of the project's Knowledge Base, but more is needed for advanced historical queries. Given the fact that the modelling started from a bibliographic ontological module, this taxonomy would provide a first step towards a more nuanced description of the contents of the charters.

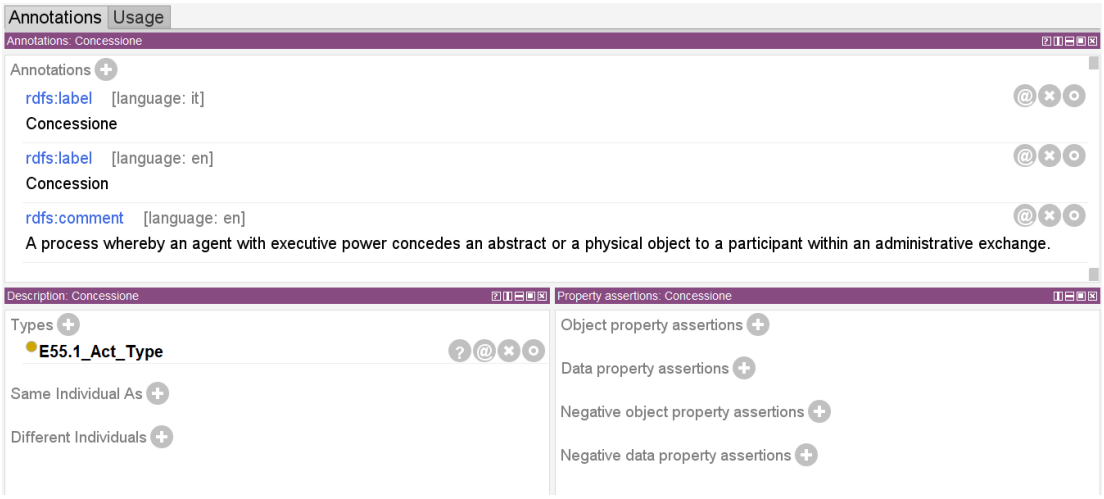


Figure 1. Modelling of the taxonomy, the Concession Act

The application of the property *P2_has_type,* and therefore of the proposed taxonomy*,* to a specific act can be seen in Figure 2:[12]

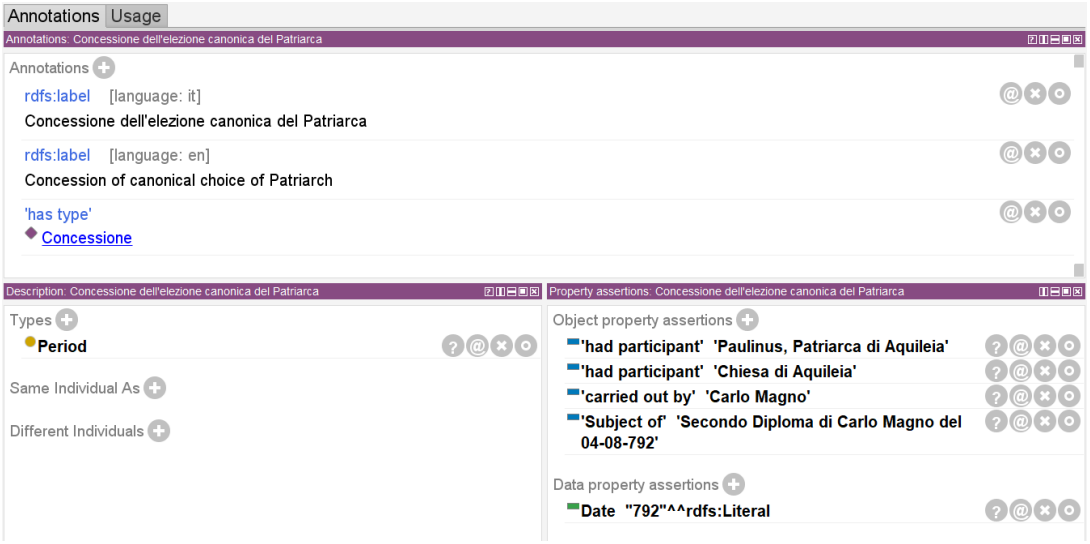


Figure 2. The application of the taxonomy to one specific act

The property would therefore explicitly link the instance of *E4_Period,* which also provides the information about the participants involved and the date of the event, when known, to a formalized description of the type of act promulgated. The proposed taxonomy can be therefore used in SPARQL queries to extract information about the type of acts assigned to an instance of *E4_Period* and would allow to gain gaining quantitative insights into the acts promulgated by the different rulers. To search for specific acts, the rdfs:label assigned can be used within a SPARQL query, since these would provide for stable and searchable terms for the acts. As mentioned above, given that the UI implemented for the project gives back data extracted through SPARQL queries, the implementation of the taxonomy allows adding filtering functions and metadata about the types of acts in the UI. Figures 3 and 4 show how the taxonomy can be

[12] The description of the act under consideration, before the application of the taxonomy proposed here, is available through the OntoVE ontology specification: https://w3id.org/ontove/ontology#1ea5b6a7-9e12-4f52-9679-96e29b7ff8f6 (last accessed 21st April 2026), while the first layer of the conceptualization of the relevant diploma expressing this act can be consulted here: https://w3id.org/ontove/ontology#9beb2f9d-52a9-415f-91b1-e2d2d3df7763 (last accessed 21st April 2026). The reader is referred to De Bastiani (2025) for a description of the concrete application of the modelling established so far, without the taxonomy proposed in this contribution, to a specific charter.

integrated within the project's UI, especially in the "Events" perspective (De Bastiani 2025: 159-160), to refine the search for types of acts.

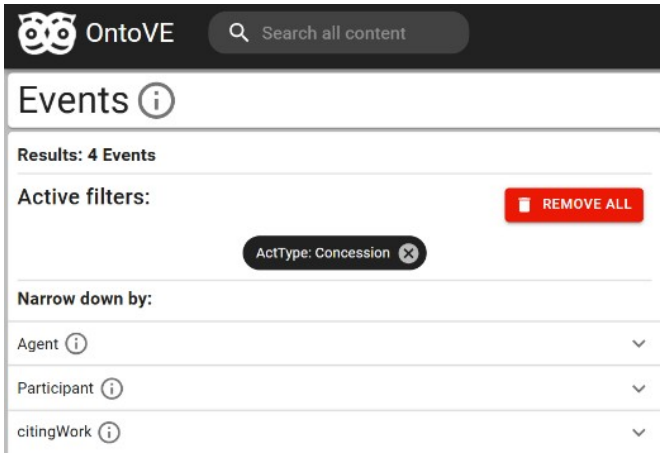


Figure 3. Filter ActType in the UI, Demo Version

Rows per page 10 1-4 of 4

Event	Agent	Participant	citingWork	ActType
Concession of grace ...	Carlo Magno	Aio, il Longobardo	Diploma of Charlemagne from 02-02-799	Concession
Concession of immunity ...	Carlo Magno	Chiesa di Aquileia ...	Diploma of Charlemagne from 04-08-792	Concession
Concession of canonical choice of Patriarch ...	Carlo Magno	Chiesa di Aquileia ...	Secondo Diploma di Carlo Magno del 04-08-792	Concession
Carlo Magno concede a Aio di dividere i suoi territori tra i suoi figli. ...	Carlo Magno	Agislaf, figlio di Aio ...	Diploma of Charlemagne from 07-07-809	Concession

Figure 4. Events belonging to the Concession Act type and related properties, Demo Version

4. CONCLUSIONS AND OUTLOOK

This paper discussed the theoretical and methodological underpinnings behind the implementation of a taxonomy of acts expressed in medieval charters. As argued in Section 3, the implementation expands an ontological module that already foresees the necessary ingredients to qualify a charter as a work conveyed in different manifestations and which is about (a) specific administrative act(s). While CIDOC-CRM provides for the general underlying semantics for the description of the acts, as argued in Section 3, a controlled vocabulary qualifying the different acts involved would allow for finer-grained queries on the data, thereby highlighting in more detail what the relationships between the rulers promulgating the act and the subjects and territories involved were. I argued that such a controlled vocabulary could be integrated within a *E55.1_Act_Type* class. The implementation proposed here would therefore not only represent an improvement of the model and of its related UI but would constitute a first step in a more structured and nuanced representation of the historical facts stated in the charters of interest. Further paths to enrich the expressive power of the conceptualization of events could involve devising a controlled vocabulary for social and political roles, as also argued for by Garbacz et al. (2010) and Álvarez-Fidalgo et al. (2025).

ACKNOWLEDGEMENTS

The OntoVE project was funded by the Italian National Operational Programme on Research and Innovation 2014-2020[13] and the European Social Fund[14] within the Recovery assistance for cohesion and

[13] PON, http://www.ponricerca.gov.it/pon-ricerca/programme/, (last accessed 21st April 2026).

[14] ESF, https://eur-lex.europa.eu/EN/legal-content/glossary/european-social-fund-esf.html, (last accessed 21st April 2026).

the territories of Europe[15]. The project was carried with Chiara Aiola and Alessandro Bertozzi the partner company Net7.[16] I moreover thank the anonymous reviewers of this contribution for their comments.

REFERENCES

Álvarez-Fidalgo, J., Rodriguez-Martin, E., & Labra-Gayo, J. (2025). MCKG: A Community-Driven Knowledge Graph on Medieval Charters, SWJ, https://www.semantic-web-journal.org/content/mckg-community-driven-knowledge-graph-medieval-charters

Bekiari, C., Bruseker, G., Canning, E., Doerr, M., Michon, P., Ore, C.E., Stead, S., & Velios, A. (2022). Definition of the CIDOC Conceptual Reference Model, Version 7.1.2.

Bekiari, C., Doerr, M., Le Boeuf, P., & Riva, P. (2015). FRBR, object-oriented definition and mapping from FRBRER, FRAD and FRSAD, approved by IFLA, Version 2.4.

Bekiari, C., Doerr, M., Le Boeuf, P., & Riva, P. (2023). LRMOO object-oriented definition and mapping from IFLA LRM, Version 0.9.6.

Bradley, J., Broun, D., Rio, A., & Hammond, M. (2019). Exploring a Model for the Semantics of Medieval Legal Charters. International Journal of Humanities and Arts Computing, 13(1–2), 136–154. https://doi.org/10.3366/ijhac.2017.0184

Carriero, V.A., Gangemi, A., Mancinelli, M.L., Marinucci, L., Nuzzolese, A.G., Presutti, V., & Veninata, C. (2019). ArCo: The Italian Cultural Heritage Knowledge Graph, In: Ghidini, C., et al. The Semantic Web – ISWC 2019. ISWC 2019. Lecture Notes in Computer Science(), vol 11779. Springer, Cham. https://doi.org/10.1007/978-3-030-30796-7_3

De Bastiani, C. (2023). Harmonizing and modelling a bibliographic ontology of imperial diplomas within the OntoVE Knowledge Base. In: Carbé, E. et al., La memoria digitale: forme del testo e organizzazione della conoscenza. Atti del XII Convegno Annuale AIUCD, In: Quaderni di Umanistica Digitale, pp. 51-156, https://doi.org/10.6092/unibo%2Famsacta%2F7721

De Bastiani, C. (2024). Connecting and visualizing data: a Knowledge Base of Germanic Cultural Heritage items in Venice', Short paper presentation held during the DH Benelux 24 conference. https://zenodo.org/records/11485003

De Bastiani, C. (2025). Modelling a bibliographic ontology of imperial diplomas: updates from the OntoVE Knowledge Base. In UMANISTICA DIGITALE, vol. 20:2025, pp. 145-164, http://doi.org/10.6092/issn.2532-8816/21245

De Bastiani, C. & Fabbris, G. (2023). Collaboration within a shared digital paradigm: opportunities and outcomes, poster presented at the ADHO2023 conference, https://zenodo.org/records/8198555.

De Bastiani, C. & Fabbris, G. (2025). Making Germanic Cultural Heritage accessible to students: a proposal for a case study, In: Rebora, S., Rospocher, M. & Bazzaco, S., Diversità, Equità e Inclusione: Sfide e Opportunità per l'Informatica Umanistica nell'Era dell'Intelligenza Artificiale, Proceedings del XIV Convegno Annuale AIUCD2025. In: Quaderni di Umanistica Digitale, pp. 485-491, https://doi.org/10.6092/unibo%2Famsacta%2F8380

Garbacz P., Trypuz R., Szady B., Kulicki P., Gradzki P., & Lechniak M.. (2010). Towards a formal ontology for history of church administration. In Frontiers in Artificial Intelligence and Applications. IOS Press. https://doi.org/10.3233/978-1-60750-535-8-345

Julien, O. (2024). Creating and using databases in medieval history: historiography, concepts and practice. Seminários do GIHM, Apr 2024, Porto, Portugal. ⟨halshs-04550456⟩

Musen, M.A. (2015). The Protégé project: A look back and a look forward, in: AI Matters. Association of Computing Machinery Specific Interest Group in Artificial Intelligence, 1(4), June 2015. DOI: 10.1145/2557001.25757003

[15]REACT-EU, https://www.europarl.europa.eu/factsheets/en/sheet/215/recovery-assistance-for-cohesion-and-the-territories-of-europe-react-eu-, (last accessed 21st April 2026).

[16] https://netseven.it/ (last accessed 21st April 2026).

Pasin, M., & Bradley, J. (2015). Factoid-based prosopography and computer ontologies: Towards an integrated approach. Digital Scholarship in the Humanities, 30(1), 86–97. https://doi.org/10.1093/llc/fqt037

Rantala, H., Ahola, A., Ikkala, E., & Hyvönen, E. (2023). How to create easily a data analytic semantic portal on top of a SPARQL endpoint: introducing the configurable Sampo-UI framework. VOILA! 2023 Visualization and Interaction for Ontologies, Linked Data and Knowledge Graphs 2023. CEUR Workshop Proceedings, Vol. 3508, October, 2023. https://ceur-ws.org/Vol-3508/paper3.pdf

Sippl, C., Burghardt, M., & Wolff, C. (2021). Modelling Cross-Document Interdependencies in Medieval Charters of the St. Katharinenspital in Regensburg. In: Graph Data-Models and Semantic Web Technologies in Scholarly Digital Editing, pp. 181-203. Norderstedt: BoD. ISBN 978-3-7543-4369-2

Vogeler, G. (2019). Versioning Charters: On the Multiple Identities of Historical Legal Documents and their Digital Representation. In: Versioning Cultural Objects: Digital Approaches, pp. 127-150. Norderstedt: BoD.

L’Archivio del Centro Conservazione e Restauro “La Venaria Reale”: un caso studio per la diffusione dei dati tecnico-scientifici

Stefania De Blasi[1], Lorena Palmieri[2], Chiara Pipino[3]

[1] Responsabile area Documentazione e Comunicazione Fondazione Centro Conservazione e Restauro “La Venaria Reale”, Italia, stefania.deblasi@ccrvenaria.it
[2] Storico dell'arte specialista, Conservatore Museo San Giacomo di Lu, palmieri.lorena@gmail.com
[3] Archivista, area Comunicazione e Documentazione, Fondazione Centro Conservazione e Restauro “La Venaria Reale”, Italia, chiara.pipino@ccrvenaria.it

ABSTRACT (ITALIANO)

Il contributo presenta il quadro metodologico e i primi risultati di un tirocinio di ricerca condotto sull’Archivio del Centro Conservazione e Restauro “La Venaria Reale” con un focus particolare sugli interventi di diagnostica eseguiti durante il restauro di dipinti su tela.
Il progetto formativo si inserisce nel contesto del Master di II livello *Esperti nelle attività di valutazione e di tutela del patrimonio culturale* (2023-2025) dell’Università di Roma Tre e prende le mosse dalla recente attività di catalogazione e pubblicazione online degli archivi del Centro, avviata nell’ambito del bando PNRR 2023. Il contributo intende mostrare come la documentazione diagnostica, spesso utilizzata esclusivamente in ambito tecnico-scientifico, possa costituire una risorsa significativa per la costruzione di forme di accesso e divulgazione del sapere. In questo senso, l’Archivio Restauri viene analizzato come caso di studio per esplorare le potenzialità di riuso e disseminazione dei dati scientifici in un’ottica di apertura, inclusività e condivisione della conoscenza.
Parole chiave: archivi digitali; conservazione; diagnostica per il restauro; public engagement.

ABSTRACT (ENGLISH)

Title: *Disseminating Technical and Scientific Data through Restoration Archives: the Case of the Centro Conservazione e Restauro “La Venaria Reale”*
This paper outlines the methodological framework and presents the preliminary results of a research internship carried out on the archive of the Centro Conservazione e Restauro “La Venaria Reale”, with a particular focus on diagnostic investigations conducted during the conservation of canvas paintings.
The project forms part of the second-level Master’s programme *Esperti nelle attività di valutazione e di tutela del patrimonio culturale* (2023–2025) at Roma Tre University and builds on the recent cataloguing and online publication of the Centre’s archival collections, developed within the framework of the 2023 PNRR funding programme. The paper argues that diagnostic documentation—traditionally confined to technical and scientific contexts—can constitute a valuable resource for widening access to knowledge and supporting its dissemination. From this perspective, the Archivio Restauri is examined as a case study to explore the potential reuse and dissemination of scientific data within a framework of openness, inclusivity, and shared knowledge.
Keywords: digital archive; conservation; diagnostic documentation; public engagement.

1. INTRODUZIONE

Il contributo si colloca nel dibattito sulle *Digital Humanities* applicate al patrimonio culturale, proponendo una riflessione sul ruolo degli archivi digitali come infrastrutture di conoscenza e come potenziali strumenti di *public engagement* (Ciotti, 2023). In particolare, il progetto di ricerca ha individuato come caso studio il Fondo Archivio Restauri del Centro Conservazione e Restauro “La Venaria Reale” (CCR), che raccoglie la documentazione relativa agli interventi di conservazione e restauro effettuati dal 2005 ad oggi su differenti tipologie di opere (dipinti su tela e tavola, arredi e sculture lignee, materiali lapidei, dipinti murali, opere in metallo, ceramica e vetro, materiale cartaceo, opere di arte contemporanea, manufatti tessili). L’eterogeneità della documentazione conservata (relazioni di restauro, schede tecniche di rilevamento, mappature grafiche, relazioni delle analisi scientifiche, condition report, relazione di

manutenzione ordinaria e straordinaria, documentazione fotografica delle fasi di restauro e delle campagne di diagnostica[1]) offre una molteplicità di punti di vista ed evidenzia il rapporto tra discipline umanistiche e scientifiche.
In questo contesto si è inserito il tirocinio formativo del master di II livello dell'Università di Roma Tre, volto alla formazione di professionisti nell'ambito della valutazione e nella tutela dei beni culturali, in relazione a musei e comunità, mercati, traffici illeciti e aree di crisi.
La ricerca è stata condotta con un approccio storico-artistico e con una particolare attenzione all'ambito interdisciplinare afferente alla *technical art history* (Cardinali, 2025)[2].

2. IL PROGETTO DI RICERCA E LA METODOLOGIA DI ANALISI DEI DATI

Il progetto di ricerca attualmente in corso prevede lo studio e l'analisi critica di parte del materiale documentario conservato presso l'Archivio del Centro Conservazione e Restauro "La Venaria Reale"[3].
La ricerca è stata avviata con l'individuazione di una base dati significativa e di particolare interesse per le evidenze documentarie presenti all'interno del Fondo Archivio Restauri e per gli ambiti di studio di chi ha condotto la raccolta dati.
La prima fase ha previsto la definizione di criteri per selezionare il materiale tra gli oltre 3.200 interventi di restauro condotti dal Centro e schedati su un *database web based* (*software* open source Pimcore). La scelta è stata impostata sui seguenti parametri: tipologia di opera, periodo di produzione, territorialità degli enti proprietari, artisti di interesse, presenza di analisi scientifiche, periodo dei restauri. Il lavoro si è quindi concentrato su un campione di circa 200 dipinti su tela, spesso attribuiti ad artisti di rilievo - Paolo Caliari (il Veronese), il Tintoretto, Rubens, Van Dyck -, databili fra il XVI e il XIX secolo e di proprietà di istituti culturali e museali del Piemonte, sui quali sono state effettuate specifiche campagne di diagnostica tra il 2005 e il 2022 (Biondi et alii, 2013: 201-209; De Blasi, Cardinali 2014: 75-79).
La prospettiva con la quale si propone di leggere i dati è duplice: da un lato una lettura critica e quantitativa delle tipologie di indagini diagnostiche documentate; dall'altro una riflessione sulle modalità di rappresentazione e valorizzazione di tali dati all'interno di un archivio digitale e dei luoghi della cultura per diffondere conoscenze altamente scientifiche attraverso approcci inediti e di partecipazione attiva (Pizzirusso, 2020).
Questo in funzione della divulgazione di dati tecnici volta ad alimentare il confronto interdisciplinare fra gli specialisti del settore culturale (storici dell'arte, restauratori, diagnosti, ricercatori) e ad ampliare le prospettive di conoscenza nei contesti di divulgazione pubblica (Costanzi Cobau, 2004). Infatti, ad oggi sono ancora poche le realtà culturali che riescono a rendere strutturale l'ampia diffusione di contenuti e informazioni altamente specialistiche. Un caso particolarmente interessante e preso come riferimento è quello del Rijksmuseum di Amsterdam, che grazie al progetto *Operation Night*[4] ha messo a disposizione del pubblico i risultati del restauro e delle indagini scientifiche condotte sul dipinto la *Ronda di notte* di Rembrandt (Gabrieli et alii, 2021).
Sulla base dei dati ad oggi informatizzati da parte dell'Archivio del Centro, risulta un campione di circa 500 restauri di opere su tela (escluse le opere di arte contemporanea e di collezione privata, di cui la documentazione è soggetta a restrizioni di consultazione). Sul totale degli interventi informatizzati più

[1] La documentazione digitale comprende: fotografie in luce visibile diffusa, radente e trasmessa, infrarosso ravvicinato in bianco e nero, in falso colore, luminescenza ultravioletta, ultravioletto riflesso, ultravioletto in falso colore, luminescenza nell'infrarosso indotta da luce visibile, luminescenza nell'infrarosso indotta da LED Red, luminescenza nell'infrarosso indotta da LED Green, luminescenza nell'infrarosso indotta da LED Blue, reflectance transformation imaging, radiografie digitale.

[2] La *Technical Art History* è un campo di studi interdisciplinare che indaga la storia dell'arte attraverso l'analisi dei materiali, delle tecniche esecutive e dei processi di produzione delle opere, avvalendosi in modo integrato di strumenti storico-artistici, conservativi e scientifici. Le indagini diagnostiche e la documentazione di restauro costituiscono fonti primarie per la ricostruzione della storia tecnica e materiale delle opere d'arte.

[3] https://www.centrorestaurovenaria.it/archivio-digitale-ccr/

[4] https://www.rijksmuseum.nl/en/whats-on/exhibitions/operation-night-watch. Tra le realtà italiane invece si segnalano i risultati pubblicati sul sito della Pinacoteca di Brera (https://pinacotecabrera.org/collezioni/conservazione-e-restauro/), dove trova spazio anche una sintetica descrizione degli interventi di diagnostica. Mentre tra gli istituti analoghi si segnala la presenza di materiale online per l'Opificio delle Pietre dure di Firenze (https://opificiodellepietredure.cultura.gov.it/) e l'Istituto Centrale per il Restauro di Roma (http://www.icr.beniculturali.it/).

della metà presenta oltre alla documentazione di restauro anche dati scientifici (immagini digitali della campagna diagnostica multispettrale, relazioni tecniche delle analisi condotte).
Questa prima analisi dei dati relativi agli interventi di diagnostica su un campione mirato di opere aiuta a quantificare con precisione la presenza di questa tipologia di documentazione all'interno dell'archivio. Inoltre, dall'analisi sulla documentazione condotta finora è stato riscontrato che le campagne di diagnostica sono state realizzate ovviamente in funzione delle necessità conoscitive funzionali al restauro e che i risultati sono stati utilizzati per lo più all'interno di pubblicazioni di settore (cataloghi di mostre tematiche o monografiche) o nell'ambito di conferenze di approfondimento. In alcuni casi di esibizioni temporanee i dati delle indagini scientifiche sono stati tradotti in narrazioni accessibili ai non specialisti, andando così a costituire un patrimonio informativo di grande valore che, quando valorizzato, ha rappresentato un elemento altamente apprezzato. Mettere in mostra elementi non visibili delle opere d'arte permette di presentare indicazioni utili per una comprensione complessiva dell'opera, del contesto artistico di produzione e che rivelano informazioni relative al *modus operandi*, alle tecniche esecutive e ai materiali impiegati, avvicinando quindi il pubblico a una conoscenza del dato materico delle opere.
Un caso eccezionale, in particolare rivolto alla diagnostica di dipinti su tela e tavola, e recentemente messo in atto è la mostra "Art from Inside. Capolavori svelati tra arte e scienza" a Palazzo Reale di Milano (9 ottobre 2025-6gennaio 2026) che ha registrato un importante successo di pubblico e di consensi sia dal mondo scientifico che dai mass media. La mostra si configurava nella presentazione al pubblico, con un allestimento particolarmente coinvolgente, delle analisi non invasive e multibanda effettuate su alcuni dei principali capolavori della pittura italiana, senza esporre alcun dipinto, ma con una narrazione accessibile della lettura delle indagini diagnostiche su tecnica, materiali, pentimenti e storia delle opere.[5]

3. POSSIBILI OUTPUT DEL PROGETTO

In questa prospettiva l'Archivio del Centro Conservazione e Restauro "La Venaria Reale" si configura come un osservatorio privilegiato sulle pratiche di diagnostica per il restauro e come possibile laboratorio per sperimentare nuove forme di accesso e condivisione dei dati scientifici nel contesto degli archivi digitali.
Il presente progetto si presenta quindi come un caso di attività di *public engagement*, inteso come messa a disposizione dell'archivio per attività di formazione e ricerca nell'ottica di una rielaborazione del materiale documentario e di una sua più ampia diffusione.
In questa sezione si desidera quindi delineare i possibili *output* individuati sulla base dello stato attuale della ricerca.

a. Analisi critica dei risultati presenti all'interno dell'archivio digitale del Centro Restauri in vista di una loro implementazione.
A distanza di un anno e mezzo dall'inizio dell'informatizzazione del Fondo Archivio Restauri e dalla sua parziale messa a disposizione *online* l'analisi dei casi studio selezionati ha permesso di evidenziare alcune criticità legate alla scarsa chiarezza dedicata ai dati sulla diagnostica per il restauro. Il tirocinio si propone quindi l'obiettivo di fornire una definizione più dettagliata della tipologia di analisi effettuata per ogni opera selezionata con l'ottica di arricchire la qualità dei dati già messi a disposizione. In questo modo si desidera favorire una diffusione più ampia e inclusiva dei risultati scientifici con l'obiettivo di coinvolgere pubblici eterogenei nella produzione e interpretazione dei dati e nella generazione di nuove forme di collaborazione (De Biase, 2014). Nello specifico durante l'attività di tirocinio in corso vengono segnalate le singole tipologie di analisi scientifiche con eventuali aggiornamenti dei contenuti presenti nel *database*. Inoltre, viene prestata particolare attenzione all'uniformazione del linguaggio e del lessico scientifico, prendendo come riferimento il *Glossario accessibile dei termini della conservazione* pubblicato dal Centro con l'obiettivo di superare le barriere linguistiche, culturali e cognitive (De Blasi, 2025).[6]

[5] La mostra è stata promossa dal Comune di Milano-Cultura e prodotta da Palazzo Reale e Fondazione Bracco, in collaborazione con 24 ORE Cultura-Gruppo 24 ORE, con la consulenza scientifica del *team* coordinato da Isabella Castiglioni, Professoressa Ordinaria di Fisica Applicata presso l'Università degli Studi di Milano-Bicocca, Direttore scientifico Centro Diagnostico Italiano-CDI, e dello storico dell'arte Stefano Zuffi.

[6] Il *Glossario accessibile dei termini della Conservazione* è l'esito di un lavoro di gruppo sviluppato con i partecipanti dello Young Professionals Forum 2024, volto all'individuazione dei termini più ricorrenti nell'ambito della conservazione e alla loro traduzione in un linguaggio facilitato, adatto alla comprensione per un pubblico di non specialisti e ispirato ai principi dell'*easy to read*. Glossario in italiano e inglese:

La segnalazione, volta all'implementazione dei dati, è eseguita tramite l'apposita funzionalità di "Annotazione" presente sulla piattaforma digitale, che permette di mettere in evidenza le schede e di riportare in un commento le correzioni e le integrazioni da effettuare. Viene così messa a disposizione del progetto una modalità di revisione dei contenuti in un'ottica di archiviazione partecipata e collaborativa, che potrebbe essere implementata in futuro.

b. Dalla teoria alla pratica: strumenti per il *public engagement*
La fase preliminare prevede l'analisi e lo studio delle *best practices* individuate su una selezione di realtà museali, archivi e istituti di cultura che hanno diffuso risultati di restauro e diagnostica in chiave divulgativa. Questa fase di confronto permetterà di mettere in evidenza le metodologie più utilizzate e di valutarne l'efficacia per lo sviluppo di strumenti operativi in futuro. Inoltre, è previsto il confronto su questi temi con i professionisti che lavorano all'interno del CCR (restauratori, diagnosti, esperti di didattica).
A partire da queste evidenze verranno individuati i punti di interesse e gli strumenti per superare la barriera tra dato tecnico e pubblico non specialista. Alla base di questa riflessione, infatti, è la convinzione che le campagne di restauro e diagnostica sono momenti di studio e ricerca particolare sulle opere, che permettono di restituire informazioni dettagliate in merito all'individuazione dei materiali, alla conoscenza delle tecniche esecutive, al processo creativo, al riconoscimento tra opera originale e copia.
Per passare dalle informazioni dell'archivio digitale alle esperienze museali ed educative sono state individuate due direttrici operative:

- l'archivio fuori dall'archivio: inserire la possibilità all'interno dei percorsi di visita museali e dei percorsi di didattica degli enti proprietari delle opere di collegarsi direttamente all'archivio *online* del Centro per visualizzare il restauro e le indagini scientifiche. In questo modo, con i dati già disponibili sul web sarebbe facilitata la comprensione dell'opera e l'accesso a conoscenze tecnico-scientifiche. Questo tipo di pratica potrebbe essere adottata anche da altri istituti che si occupano di conservazione e restauro, consentendo nuove connessioni tra centri di ricerca e sedi museali. Attualmente questa prassi è in atto presso la Palazzina di Caccia di Stupinigi e presso la Reggia di Venaria ed è integrata nel percorso di visita. Questo sistema, che consiste nell'applicazione di QR code collegati al catalogo digitale del CCR e posizionati accanto alle didascalie delle opere, è minimamente impattante e facilmente realizzabile.

- nuovi narrazioni digitali: a partire dallo studio delle *best practices* selezionate individuazione di almeno un format divulgativo (gaming, percorsi virtuali, mappe concettuali e grafici, podcast, video di approfondimento) particolarmente adatto a tradurre la complessità della diagnostica e del restauro grazie all'uso delle nuove tecnologie.

4. CONCLUSIONI

A distanza di qualche mese dell'inizio del tirocinio è possibile effettuare un primo bilancio delle attività e misurare le effettive ricadute del progetto.
La ricerca si presenta come primo caso di sperimentazione sull'Archivio del Centro di attività di *public engagement* e archiviazione partecipata inteso ancora in senso stretto. Infatti, il coinvolgimento di soggetti terzi, per ora, non è da intendersi nei confronti di un pubblico allargato, ma verso un pubblico di studiosi interessati ad utilizzare i materiali dell'archivio secondo nuove prospettive di ricerca.
Allo stato attuale del lavoro è quindi possibile evidenziare le azioni intraprese durante il progetto e individuarne i limiti. Tra le attività in fase di conclusione c'è l'implementazione dei dati relativi alla diagnostica attraverso il sistema di "annotazioni" presente sul software in uso. Grazie al focus specifico della ricerca di tirocinio è stato possibile ottenere dati più puntuali relativi alla diagnostica applicata ai beni culturali. Questo permetterà di facilitare la ricerca di questo tipo di documentazione ai futuri utenti che accederanno all'archivio digitale, previa registrazione.
Un limite del progetto consiste nel fatto che non propone o sperimenta effettivamente una nuova modalità di diffusione dei contenuti scientifici, ma prevede solamente lo studio di una selezione di casi messi in rete da istituti di cultura e ricerca ed enti museali. I motivi di questa scelta sono diversi, oltre al

centrorestaurovenaria.it/wp-content/uploads/2025/10/Glossario-atti-YPF-ITA-ENG.pdf; glossario in arabo: centrorestaurovenaria.it/wp-content/uploads/2026/04/Glossario-termini-conservazione_ARABO.pdf

numero di opere indagato e alle tempistiche del progetto, la ragione principale risiede nel fatto che il tirocinio ha privilegiato uno studio quantitativo che desse conto della percentuale di presenza di documentazione diagnostica su un campione selezionato di opere. Lo scopo principale è stato quindi quello di implementare la precisione dei dati all'interno del database attraverso un'azione partecipata, finalizzata a mostrare la rilevanza dei dati scientifici in relazione agli obiettivi dei singoli interventi di restauro e alla conoscenza più dettagliata delle opere che sarà poi più funzionale alla selezione di opere e elementi contenutistici utili alla divulgazione scientifica e al coinvolgimento del pubblico.

BIBLIOGRAFIA

Biondi, E., Cavaleri, T., Cardinali, E., Gatti, A., Giovagnoli, A. (2013). Un dipinto del Tintoretto al Centro Conservazione e restauro "La Venaria Reale": la diagnostica al servizio dell'arte, in Lo stato dell'arte 11, Atti XI Congresso nazionale IGIIC, Bologna 10-12 ottobre 2013, 201-209.

Cardinali, M. (2025). *Dalla Diagnostica Artistica alla Technical Art History. Nascita di una metodologia di studio della storia dell'arte (1874-1938)*, Cores.

https://www.cores-project.it/wp-content/uploads/Volume_TAH_Anteprima.pdf

Ciotti, F. (2023), Digital humanities. Metodi, strumenti, saperi. Carrocci.

Costanzi Cobau, A. (2004), Aperto per restauro: quando un cantiere diventa ambiente museale. In *Lo stato dell'arte. Conservazione e restauro. Confronto di esperienze, Atti del II congresso Nazionale IGIIC* (pp. 1-9). https://www.academia.edu/2056004/Aperto_per_restauro_comunicare_la_conservazione

De Biase, F. (2014), *I pubblici della cultura. Audience development, audience engagement*, Franco Angeli Editore.

De Blasi, S., Cardinali, M. (2014). L'analisi e il restauro delle Allegorie di Pallanza, in Quattro Veronese venuti da lontano. Le Allegorie ritrovate, Milano, Officina Libraria, 75-79.

De Blasi, S. (2025). Glossario Accessibile dei Termini della Conservazione, Venaria Reale. https://www.centrorestaurovenaria.it/wp-content/uploads/2025/10/Glossario-atti-YPF-ITA-ENG.pdf

Gabrieli, F., Delaney, J. K., Erdmann, R. G., Gonzalez, V., van Loon, A., Smulders, P., Berkeveld, R., van Langh, R., & Keune, K. (2021). Reflectance Imaging Spectroscopy (RIS) for *Operation Night Watch*: Challenges and Achievements of Imaging Rembrandt's Masterpiece in the Glass Chamber at the Rijksmuseum. *Sensors (Basel, Switzerland)*, *21*(20), 6855. https://doi.org/10.3390/s21206855

Pizzirusso, I. (2020), Digital è public? L'esempio di alcune banche dati a contenuto storico. Novecento.org, 14. DOI: 10.12977/nov351.

Strumenti digitali per i percorsi di visita della Palazzina di Caccia di Stupinigi: Residenza Sabauda e Museo dell'Ammobiliamento

STEFANIA DE BLASI

DOCUMENTAZIONE E COMUNICAZIONE FONDAZIONE CENTRO CONSERVAZIONE E RESTAURO "LA VENARIA REALE" – COLLEZIONI ARTISTICHE DELLA FONDAZIONE ORDINE MAURIZIANO, ITALIA-
stefania.deblasi@ccrvenaria.it

ABSTRACT (ITALIANO)

La Palazzina di Caccia di Stupinigi, padiglione di caccia alle porte di Torino progettato nel 1729, ha una duplice funzione e missione: inserita nel sito seriale UNESCO delle Residenze Sabaude, rappresenta l'immagine della vita di corte nel Settecento, e al contempo è dal 1926 il primo Museo dell'Ammobiliamento italiano. In occasione dei cento anni del Museo, la Fondazione Ordine Mauriziano (FOM), proprietaria del Bene, ha avviato una riflessione sui percorsi di visita grazie ad una massiva campagna di digitalizzazione degli inventari del XIX e XX secolo della Real Casa, schede e immagini storiche delle collezioni. L'obiettivo è rendere accessibile e comprensibile ad un pubblico universale le due anime del luogo e distinguere, nell'articolato susseguirsi di stanze, cosa sia riferibile alla storia della Residenza e cosa invece sia rappresentativo del Museo. La creazione di una piattaforma web based "Collezioni in Ordine", in grado di gestire dati provenienti da sistemi pregressi di catalogazione on line e locali, inventari storici e schede per la conservazione programmata, ha consentito di creare un catalogo delle collezioni consultabile sul sito web della FOM, affrontando molte criticità.
I dati accessibili sono stati base per definire un percorso di visita che presenta, attraverso QR code e pannelli fotografici, approfondimenti su opere, sulla storia e sul restauro o sulle destinazioni d'uso delle sale nei diversi periodi. È anche disponibile un'esperienza immersiva attraverso la realtà aumentata che mette in scena, attraverso QR su totem tematici, diversi momenti di vita della residenza e dei suoi protagonisti.
Infine, la connessione tra "Collezioni in Ordine" e gli archivi digitali del Centro Conservazione e Restauro "La Venaria Reale" (CCR) ha consentito l'accrescimento delle informazioni sul restauro e la loro traduzione in italiano e inglese facilitato, grazie al progetto Humanities in Conservation, sull' accessibilità universale ai dati scientifici.
Parole chiave: public engagement nei musei; realtà aumentata; catalogo digitale delle collezioni e degli archivi; API; easy-to-read

ABSTRACT (ENGLISH)

Digital tools for different visitor experiences inside the Stupinigi Hunting Lodge: Savoy Residence and Furniture Museum
The Stupinigi Hunting Lodge, a hunting pavilion on the outskirts of Turin designed by Filippo Juvarra in 1729, has a dual function and mission: included in the UNESCO serial site of the Savoy Residences, it represents the courtly image of court life in the 18th century, and at the same time, since 1926, it has been the first Italian Furniture Museum. To mark the museum's centenary, the Fondazione Ordine Mauriziano (FOM), which owns the property, has launched a review of the visitor experience thanks to a massive campaign to digitise the 19th- and 20th-century inventories of the Royal House, historical records and images of the collections. The aim is to make the two souls of the place accessible and understandable to a universal audience and to distinguish, in the articulated succession of rooms, what is related to the history of the Residence and what is representative of the Museum. The creation of a web-based platform called 'Collezioni in Ordine', capable of managing data from previous online and local cataloguing systems, historical inventories and records for planned conservation, has made it possible to create a catalogue of collections that can be consulted on the FOM website, resolving many critical issues. The accessible data was used as a basis for defining a visitor route that uses QR codes and photographic panels to provide in-depth information on the works, their history and restoration, and the uses of the rooms in different historical periods. An immersive experience is also available through augmented

reality, which uses QR codes on themed totems to bring to life different moments in the history of the residence and its protagonists.
Finally, the connection between "Collezioni in Ordine" and the digital archives of the "La Venaria Reale" Conservation and Restoration Centre (CCR) has made it possible to increase the amount of information available about the restoration and translate it into easy-to-read Italian and English, thanks to the Humanities in Conservation project aimed at universal accessibility to scientific data.
Keywords: museums public engagement; augmented reality; collections and inventories digital catalogue; API; easy-to-read

1. INTRODUZIONE

In occasione del centenario della creazione del primo Museo dell'Ammobiliamento in Italia, realizzato nel 1926 all'interno della Palazzina di Caccia di Stupinigi, residenza storica alle porte di Torino, l'Ente proprietario del Bene si è posto l'obiettivo di riflettere sulla natura e sulle origini del museo che rimane un *unicum* nel panorama culturale italiano (Gabrielli, 2014). La necessità principale è di offrire nuovi strumenti agli oltre duecentomila visitatori annui del sito museale una chiara chiave di lettura degli ambienti e delle collezioni, creando momenti di approfondimento, esperienziali e di apprendimento culturale facilmente raggiungibili ai differenti pubblici per età, cultura e provenienza linguistica.
Punto fondamentale di partenza è la rappresentazione di un sito museale con due anime distinte: il museo del mobile italiano del Settecento e la residenza sabauda dalla lunga e articolata storia. La criticità principale consiste nella sovrapposizione delle due anime che nel tempo si sono fuse conferendo al luogo un'unica immagine e lettura di ambienti aulici riccamente arredati senza distinzioni e con pochi mezzi di discernimento e di coinvolgimento del pubblico che conduce una visita non guidata.
La sfida consiste nel trasferimento delle conoscenze emerse negli oltre trent'anni di ricerche, campagne di catalogazione, digitalizzazione e restauri ad un pubblico generico e specialistico che sceglie di percorrere autonomamente le sale del museo. La difficoltà è rappresentata dalla capacità di traduzione di elementi specialistici in registri di comunicazione accessibili e coinvolgenti.
Nulla si sarebbe potuto avviare senza la realizzazione di una piattaforma documentale che potesse collezionare i dati provenienti dai diversi sistemi di catalogazione e inventariazione, attraverso un riversamento di dati che si è confrontato con la progettazione di una struttura in grado di accogliere numerose strutture di dati eterogenee e senza la disponibilità di open data o API delle piattaforme di partenza. La progettazione della struttura "Collezioni in Ordine" ha tenuto conto dell'esigenza non solo di mettere in relazione diversi data set ma anche di restituire dati e strutture attraverso API che possano in futuro essere la porta di accesso per confluire in reti Linked Open Data non ancora disponibili o nella Digital Library del Ministero della Cultura, in fase di costruzione.

2. COLLEZIONI IN ORDINE

La Fondazione Ordine Mauriziano (FOM) è erede del vastissimo patrimonio culturale del soppresso Ordine Mauriziano, ordine cavalleresco e ospedaliero legato al casato sabaudo, fondato nel 1572 e trasformato nel 2004 in Fondazione pubblica demaniale con finalità di tutela e valorizzazione dei Beni Culturali ad essa conferiti. Il patrimonio consiste in diciotto Beni tra siti UNESCO, abbazie medioevali, basiliche e chiese barocche e monumenti di epoca romana dislocati tra il Piemonte e la Sardegna.[1] La Palazzina di Caccia di Stupinigi è uno di questi Beni che oltre ad essere residenza storica, inserita nel sito seriale UNESCO delle Residenze Sabaude, ospita dal 1926 l'unico Museo dell'Ammobiliamento esistente in Italia.
Presupposto per consentire una corretta gestione di un patrimonio vasto e di grande valore culturale e scientifico è la conoscenza delle consistenze, attraverso censimenti, inventari, campagne di documentazione e catalogazione. Inoltre, elemento imprescindibile per qualsiasi piano di valorizzazione e

[1] I Beni di proprietà della Fondazione Ordine Mauriziano sono: la Palazzina di Caccia di Stupinigi, la chiesa parrocchiale di Stupinigi, il Castelvecchio di Stupinigi, la Precettoria di Ranverso e l'Ospedaletto di Ranverso, le chiese della Madonna dei Boschi di Buttigliera Alta e di San Maurizio di Mongiove di Chivasso, il Priorato e la chiesa di San Martino a Torre Pellice in provincia di Torino, l'Ospedale Mauriziano Umberto I, l'Archivio Storico Mauriziano e la Basilica Mauriziana dei Santi Maurizio e Lazzaro a Torino, la chiesa di Santa Caterina e la Torre Rossa ad Asti, l'Abbazia di Staffarda e le cappelle di Saccabonello e Lagnasco in provincia di Cuneo e la Basilica Mauriziana di Santa Croce a Cagliari.

sviluppo scientifico è la messa a disposizione dei dati connotanti delle opere (contenuti) e quindi la possibilità di promuovere campagne di analisi e studio. Se manca il primo elemento, ovvero la ricognizione e l'inventariazione con una fotografia dello stato di fatto, diventa estremamente arduo e dispendioso il secondo elemento, quale base per possibili programmi di valorizzazione e public engagement.
Per gestire un patrimonio storico artistico museale e bene pubblico, come lo straordinario Patrimonio della FOM, è stato necessario puntare su una razionalizzazione degli innumerevoli progetti che nel tempo hanno sondato, censito, analizzato e studiato i beni dai molti punti di vista possibili, partendo dai progetti ministeriali, fondamento primario per la tutela.
Fin dalla legislazione preunitaria italiana, i beni mauriziani, e in particolar modo Sant'Antonio di Ranverso, sono stati coinvolti dalle campagne di censimento ministeriali e di studio, finalizzate a conoscere per tutelare e preservare. Ne è seguita una preziosissima mole di attività e studi che non era più possibile lasciare scollegati. Solo mettendo a sistema le numerose campagne di catalogazione e fotografiche con un attuale aggiornamento inventariale e con l'innovazione tecnologica di cui oggi disponiamo è stato possibile costituire uno strumento utile, non solo alla conoscenza delle opere, ma anche alla loro corretta conservazione, gestione e valorizzazione, verso nuovi programmi di studi e mostre sulle collezioni e sui percorsi museali dei beni. Nel corso degli anni i beni FOM sono stati catalogati, schedati e fotografati in occasione di campagne ministeriali, campagne di restauro e progetti universitari su sistemi e formati differenti, digitali e non, e con finalità non sempre coerenti. Dagli anni Novanta ad oggi, inoltre, una grande mole di attività di conservazione e restauro hanno interessato i beni mauriziani, *in primis* la Palazzina di Stupinigi, producendo moltissima documentazione, solo in parte digitale. Una puntuale ricognizione e la costituzione di un archivio della storia conservativa dei beni si sono poste come base per lo sviluppo di nuove linee di ricerca e applicazione. Le catalogazioni ministeriali dei beni FOM si presentano incomplete e con ultimo aggiornamento al 2006. La recente ricognizione ha rintracciato sistemi, che in parte si sovrappongono per presenza di dati ripetuti presenti: sulla piattaforma online Sigecweb,[2] dove risultano on line accessibili 3455 opere, ossia meno della metà dei beni mauriziani;[3] sul software di catalogazione locale della Soprintendenza piemontese ArtView[4] (in uso fino al 2006), che raccoglie 4780 schede OA, in schedari cartacei che raccolgono campagne di catalogazione preinventariale datate 1994-2006 e che in parte si sovrappongono con le campagne digitali e in parte costituiscono l'unica fonte di schedatura con fotografie analogiche in bianco e nero per un totale di 810 schede; sul software Filemaker, dove risultavano schedati su database correlati i diciotto inventari storici e i testimoniali di stato della Palazzina di Caccia di Stupinigi dal 1739 al 1926, conservati in diversi fondi archivistici dell'Archivio di Stato di Torino, dell'Archivio Centrale dello Stato di Roma e degli Archives Nationales de France.
Nel corso del 2023-24 è stata progettata e realizzata una piattaforma documentaria web based in grado di gestire le diverse esigenze di archiviazione, implementazione e aggiornamento dei dati relativi alle opere al fine di garantire il monitoraggio dei beni, anche dal punto di vista conservativo e di aggiornamento scientifico oltre che di archivio storico. Lo strumento, progettato nell'ambito della collaborazione tra FOM e l'Area documentazione del Centro Conservazione e Restauro "La Venaria Reale", realizzato da Lyme Agency di Novara[5] e con un contributo della Fondazione Compagnia di San Paolo, è stato pensato per rispondere alle norme di catalogazione dei beni storico artistici ministeriali, ma anche per rispondere alle esigenze operative dell'ente che gestisce serie complesse di collezioni.
La piattaforma web, denominata specificatamente "Collezioni in Ordine", è funzionale alla fruizione on line delle schede delle singole opere conservate in tutti i beni FOM ed è progettata secondo tracciati e lemmari ICCD e senza escludere in futuro la messa a disposizione dei dati su sistemi linked open data. La piattaforma dispone di un backend di collegamento e ordinamento dei diversi sistemi di schedatura delle

[2] https://www.catalogo.beniculturali.it/ (cons. 07/02/2026. Sigecweb raccoglie la digitalizzazione parziale di schede cartacee e il riversamento delle schede informatizzate con i software che si sono susseguiti in dotazione all'ente schedatore di competenza, l'allora Soprintendenza per i Beni Stoci Artistici e Demoetnoantropologici del Piemonte fino al 2007 (ArtPast e Sigec).
[3] Le schede riguardano mirate campagne di catalogazione sulla Palazzina di Stupinigi (percorso di visita e depositi), la chiesa parrocchiale di Stupinigi, la Precettoria di Ranverso, l'Abbazia di Staffarda, l'Ospedale Mauriziano, la Basilica Mauriziana, Chiesa di Santa Caterina di Alessandria, Chiesa di San Martino di Torre Pellice.
[4] Software sviluppato da TAI sas di Milano, si veda https://www.taisas.com/products.html (cons. 07/02/2026).
[5] https://lymeagency.com/portfolio/collezioni-in-ordine/ (cons. 07/02/2026).

opere avendo recepito il materiale presente su Sigecweb, Artview e Filemaker. Per le opere della Palazzina di Caccia di Stupinigi è presente, inoltre, la connessione tra ciascuna con la propria catena inventariale (presente sulle schedature ICCD) e quindi con gli inventari storici informatizzati (su Filemaker). Per ogni opera è stata predisposta la possibilità di aggregare fascicoli documentari (già in formato digitale o che verranno via via digitalizzati), fotografie e video. Costituisce quindi anche un repository per documentazioni di diversa natura e formato: campagne fotografiche, video, documenti. È inoltre correlato un database relazionale per la digitalizzazione degli inventari storici, idoneo alla referenziazione delle opere all'interno del documento per garantire e agevolare la ricostruzione delle catene inventariali e per favorire gli studi di provenienza (es. collezioni di pertinenza del Museo dell'Ammobiliamento di Stupinigi provenienti da altre residenze italiane). È stata compiuta una scelta dei campi e delle occorrenze da rendere visibili nel front end della piattaforma, che risiede direttamente sul sito web della FOM, e che ad oggi comprende la pubblicazione di una selezione di schede di catalogo relative ai tre principali beni aperti al pubblico: Palazzina di Caccia di Stupinigi, Precettoria di Sant'Antonio di Ranverso e Abbazia di Staffarda.[6]

3. DALLA CATALOGAZIONE ALLA VALORIZZAZIONE

Attualmente il sistema è in uso e in continuo aggiornamento, la struttura consente dal back end di aggregare dati correnti, aggiornamenti e dati storici. In occasione del centenario del Museo dell'Ammobiliamento della Palazzina di Caccia di Stupinigi si è avviata la prima vera messa a valore dei dati provenienti dal sistema di catalogazione e gestione. Si è definito per uno degli appartamenti della Residenza, l'Appartamento di Levante, un doppio percorso di visita che racconti il Museo e la vita di corte, in particolare la vita dell'Appartamento nell'epoca della regina Margherita di Savoia, ultima abitante di Stupinigi prima della retrocessione dell'edificio al demanio che ne decretò la conversione in museo (Failla, 2018; De Blasi, 2025, aprile). In alcune sale dell'appartamento sono stati posizionati dei pannelli descrittivi e fotografici con la possibilità di approfondire la destinazione d'uso della sala attraverso una gallery di immagini storiche che dal back end sono state rese visibili su una landing page del sito, oscurata dalla navigazione in chiaro. Nelle sale stesse i cartellini delle opere distinguono se l'arredo o il dipinto sia parte delle collezioni di museo o derivi dai tempi della residenza storica. I pannelli di sala mostrano riportano QR code che puntano al front end del catalogo per approfondirne la storia di alcune opere e fornire maggiori elementi al visitatore. In altri casi, in cui l'attenzione del visitatore deve essere posta sulla storia di un'opera della collezione museale che il restauro ha consentito di approfondire, il QR code posto sul cartellino o pannello di sala punta al link relativo ai dati di restauro presenti in back end, connesso a sua volta all'archivio digitale dei restauri del CCR "La Venaria Reale".
Grazie a una ventennale collaborazione tra la FOM e il CCR, l'archivio digitale dei restauri, liberamente fruibile sul sito, raccoglie oltre 250 interventi di restauro effettuati sulle opere della Palazzina di Caccia di Stupinigi, con circa 3000 immagini relative al restauro e alle analisi scientifiche condotte principalmente su arredi, dipinti, manufatti tessili e metallici. L'archivio digitale dei restauri è stato realizzato secondo criteri di accessibilità offrendo un abstract descrittivo di ciascun intervento in linguaggio easy-to-read italiano e inglese, nell'ambito del progetto grazie al progetto *Humanities in Conservation*, sull'accessibilità universale ai dati scientifici[7](De Blasi, Guerzoni, Pipino, 2025). Il sitoweb su cui l'utente approda consente inoltre di approfondire i significati del lessico specifico della conservazione attraverso la messa a disposizione di un "Glossario accessibile ai termini della conservazione", pubblicazione open access in italiano, inglese e anche in arabo (De Blasi, 2025, settembre)[8].
Infine, nell'ultimo anno il percorso della Palazzina si è arricchito anche di quattro stazioni di realtà aumentata all'interno del monumentale Salone Centrale che consentono di seguire una breve piéce teatrale messa in scena da rievocatori storici che rappresentano quattro salienti momenti di vita della storia della Palazzina di Caccia di Stupinigi in diverse epoche storiche: il cantiere di Filippo Juvarra, la stagione delle cacce alla fine del Settecento, la visita di Napoleone nel 1805 e la vita della piccola corte

[6] https://catalogo-opere.ordinemauriziano.it/ (cons. 07/02/2026).

[7] Finanziato con fondi PNRR NextGeneration UE (2023-2024).

[8] https://www.centrorestaurovenaria.it/archivio-digitale-ccr/ (cons. 07/02/2026).
https://www.centrorestaurovenaria.it/wp-content/uploads/2025/10/Glossario-atti-YPF-ITA-ENG.pdf (cons 07/02/2026).

della Regina Margherita nel 1909[9]. Un'esperienza immersiva attivabile attraverso QR code posti su pannelli esplicativi che dall'inquadratura con la telecamera del proprio smartphone, in uno specifico scorcio del salone, consente al visitatore di rivivere momenti storici e comprendere meglio l'importanza di una residenza suburbana dedicata alla caccia nella storia della corte sabauda.

BIBLIOGRAFIA

De Blasi, S. (2025, aprile). Becoming Stupinigi. La mostra del 1922, gli allestimenti d'ambientazione e le premesse per il Museo di Arte, Storia e Ammobiliamento. Failla, M.B. & Quagliaroli, S. Officina 1922. Una mostra alle origini della fortuna del Barocco. Quaderni di Ricerca, 6, Fondazione 1563 per l'Arte e la Cultura della Compagnia di San Paolo. Genova: Sagep Editori. 89-103. https://www.fondazione1563.it/wp-content/uploads/2025/06/Fondazione1563-Officina-1922-2025.pdf

De Blasi, S. (a cura di) (2025, settembre). Glossario accessibile dei termini della conservazione. Accessible glossary of conservation terms. Centro Conservazione e Restauro "La Venaria Reale", Roma: Cores. https://www.centrorestaurovenaria.it/wp-content/uploads/2025/10/Glossario-atti-YPF-ITA-ENG.pdf

https://www.centrorestaurovenaria.it/wp-content/uploads/2026/04/Glossario-termini-conservazione_ARABO.pdf

De Blasi, S., Guerzoni, E., & Pipino, C. (2025). Accessibilità e inclusione per la documentazione del restauro: gli archivi del Centro Conservazione e Restauro "La Venaria Reale". Rebora, S., Rospocher, M. & Bazzaco, S. (2025) Diversità, Equità e Inclusione: Sfide e Opportunità per l'Informatica Umanistica nell'Era dell'Intelligenza Artificiale, Proceedings del XIV Convegno Annuale AIUCD2025. Verona: AIUCD. 66-71.

Failla, M.B. (2018). Ambientazioni e "gusto modernissimo". Musei a Torino negli anni tra le due guerre. Le voci del museo, 39, Firenze: Edifir.

Gabrielli, E. (a cura di) (2014). La Palazzina di Caccia di Stupinigi. Firenze: Leo S. Olschki Editore.

[9] Progetto di Marta Fusi, realizzato da No real interactive srl con Motion Pixel, ARCA Studio, testi di Alessia Giorda e Davide Borra, interpretati da 50 rievocatori in costumi storici dell'Associazione Le Vie del Tempo, regia di Davide Borra. https://www.ordinemauriziano.it/palazzina-di-caccia-di-stupinigi/visita-stupinigi/ (cons 07/02/2026).

Strategie di protezione, governance e sostenibilità degli archivi orali

Rosaria De luca[1], Elvira Mercatanti[2], Monica Monachini[2]
[1] CNR Istituto di Informatica e Telematica (CNT-IIT), Italia - rosaria.deluca@iit.cnr.it
[2] CNR Istituto di Linguistica Computazionale "A. Zampolli" (CNR-ILC), Italia – nome.cognome@ilc.cnr.it

ABSTRACT (ITALIANO)

Il contributo presenta il modello di governance sviluppato nell'ambito del progetto PRIN 2022 ROADS per la gestione sostenibile degli archivi orali, risorse caratterizzate da rilevanti criticità sul piano normativo ed etico. Al centro della riflessione si colloca il dato vocale, inteso come oggetto ibrido che funziona simultaneamente come documento storico, fonte per la ricerca scientifica e dato biometrico, richiedendo un bilanciamento tra apertura dei dati e protezione dei diritti.
Per affrontare tale complessità, ROADS integra i principi FAIR con l'approccio di Privacy by Design e con il principio di accountability del GDPR, delineando un sistema di accesso controllato, basato su protocolli di ricerca diligente per la gestione dei materiali d'archivio (legacy) e meccanismi di autenticazione federata. Tale struttura protegge le fonti e i diritti fondamentali, garantendo al contempo la solidità giuridica degli archivi.
Il modello è messo a confronto con il progetto Voci dall'Inferno, che affronta criticità analoghe nella gestione di un archivio digitale di testimonianze orali di sopravvissuti ai Lager. L'analisi del caso di studio mostra come ROADS possa supportare la transizione da archivi di progetto a risorse FAIR, sostenibili e riusabili, nel rispetto della protezione dei dati personali e della specificità del contesto storico documentato.
Parole chiave: archivi orali; dati sensibili; principi FAIR; privacy; governance dei dati

ABSTRACT (ENGLISH)

Protection, governance and sustainability strategies for oral archives.
This paper presents the governance model developed within the PRIN 2022 ROADS project for the sustainable management of oral archives, resources characterized by significant legal and ethical challenges. At the core of the framework lies voice data, understood as a hybrid object that simultaneously functions as a historical document, a source for scientific research, and a biometric datum, requiring a balance between data openness and the protection of individual rights.
To address this complexity, ROADS integrates the FAIR principles with a Privacy by Design approach and the GDPR accountability principle, outlining a controlled access system that incorporates due diligence protocols for the management of legacy archival materials and federated authentication mechanisms for data access. This structure ensures the protection of sources and fundamental rights, while guaranteeing the legal robustness of the archives.
This approach is compared with the Voci dall'Inferno project, which faces similar challenges in managing a digital archive of oral testimonies by survivors of Nazi concentration camps. The case study shows how the ROADS model can support the transition from project-based archives to FAIR, sustainable, and reusable resources, while ensuring compliance with data protection requirements and respect for the specificity of the documented historical context.
Keywords: oral archives; sensitive data; FAIR principles; privacy; data governance

1. INTRODUZIONE

Gli archivi orali rappresentano un patrimonio fondamentale di memoria collettiva e una risorsa centrale per le Digital Humanities, in particolare per lo studio della diversità linguistica, delle pratiche culturali e delle dinamiche sociali. Essi rivestono un ruolo cruciale in ambiti quali la linguistica, la storia orale e l'etnomusicologia, anche grazie alla crescente disponibilità di strumenti digitali per l'analisi, la conservazione e la condivisione delle fonti sonore (De Dominicis, 2002; Sornicola et al., 2019).

Le testimonianze orali presentano una natura complessa, rendendo gli archivi orali un terreno privilegiato per l'incontro tra pratiche di ricerca umanistica, infrastrutture digitali e quadri normativi, introducendo criticità rilevanti nella gestione, nell'accesso e nel riuso dei dati. La loro gestione richiede pertanto un approccio interdisciplinare che integri competenze umanistiche, archivistiche, tecnologiche e giuridiche, capace di coniugare apertura e fruizione scientifica con la tutela dei diritti degli interessati. In questo contesto, l'applicazione dei principi FAIR (Findable, Accessible, Interoperable, Reusable) non può essere ridotta a un requisito tecnico, ma implica una riflessione sulle condizioni giuridiche ed etiche dell'utilizzo dei dati, in particolare alla luce del Regolamento Generale sulla Protezione dei Dati (GDPR).
Nonostante il crescente interesse per il patrimonio orale, in Italia la digitalizzazione e l'accessibilità degli archivi risultano ancora limitate da criticità quali la frammentazione delle pratiche, la conservazione locale e le incertezze normative. L'adozione dei principi FAIR nei dati orali richiede quindi un ripensamento delle responsabilità giuridiche ed etiche della ricerca, oltre che delle soluzioni tecniche (Calamai & Frontini, 2018).
In questo contesto si colloca il progetto ROADS[1] che propone un modello FAIR-by-design per la gestione degli archivi orali, in cui gli aspetti tecnici, giuridici ed etici sono integrati lungo l'intero ciclo di vita dei dati. Il progetto affronta due tipologie di fonti: archivi orali storici preesistenti - raccolti in epoche precedenti all'attuale quadro normativo - e interviste raccolte ex novo nell'ambito della ricerca. Questa duplice prospettiva consente di sviluppare strategie differenziate di governance, rendendo operativa la FAIRness anche in presenza di vincoli normativi eterogenei.
Il contributo mette in dialogo il modello ROADS con il progetto Voci dall'Inferno, un'iniziativa scientifica che ha visto la creazione di un archivio digitale di testimonianze non letterarie di sopravvissuti ai Lager. L'obiettivo è mostrare come un modello di governance strutturato possa supportare la valorizzazione responsabile di archivi orali complessi, consentendo la transizione da archivi di progetto a risorse FAIR, sostenibili e riutilizzabili.
L'articolo è organizzato come segue: la sezione 2 introduce il progetto ROADS; la sezione 3 illustra il modello di governance e l'inquadramento giuridico; la sezione 4 descrive il progetto Voci dall'Inferno; la sezione 5 discute l'applicazione al caso di studio; la sezione 6 presenta le conclusioni.

2. IL PROGETTO ROADS: UN MODELLO PER I DATI ORALI

Il progetto ROADS nasce con l'obiettivo di superare la frammentazione del patrimonio orale italiano attraverso lo sviluppo di strumenti per il recupero, la conservazione, la descrizione e il riuso degli archivi orali, integrando esigenze scientifiche e vincoli normativi (Abete et al., 2025). A queste attività si affiancano iniziative di formazione e disseminazione finalizzate allo sviluppo di competenze per la gestione e per la valorizzazione delle fonti orali come risorse di ricerca accessibili, sostenibili e riutilizzabili in contesti scientifici (Abete et al., 2026b; Monachini & Mercatanti, 2026).
ROADS adotta un'impostazione in cui la FAIRness costituisce un principio guida per l'intero ciclo di vita dei dati, orientando le scelte metodologiche, infrastrutturali e giuridiche. In questo quadro, il riconoscimento della natura ibrida del dato orale rappresenta un elemento centrale del modello e implica una gestione differenziata dei livelli di accesso e di riuso.
Il progetto integra il trattamento e l'analisi di archivi orali storici preesistenti con attività di acquisizione diretta di dati. In particolare, sono state condotte interviste a ricercatori e responsabili di raccolte, finalizzate a ricostruire la genesi degli archivi, la loro consistenza, i criteri di organizzazione e lo stato di conservazione.
L'archivio pilota selezionato è quello della storica Gabriella Gribaudi, già docente di Storia contemporanea all'Università degli Studi di Napoli Federico II, nonché fondatrice e prima presidente dell'Associazione Italiana di Storia Orale. Si tratta di un archivio di ricerca esemplare, basato su indagini sul campo condotte a partire dal 1974 sulla storia sociale dell'Italia meridionale nel periodo della Seconda Guerra Mondiale (Gribaudi, 1980, 1990, 2005, 2016, 2023). La raccolta è composta da 148 nastri audio, di cui 91 audiocassette, 15 minicassette e 42 Digital Audio Tapes (DAT). Il numero complessivo degli intervistati è di circa 189 persone, 108 uomini e 81 donne, nate tra il 1909 il 1945. L'archivio comprende sia testimoni diretti degli eventi narrati sia individui che, allora bambini, ne hanno restituito il ricordo attraverso i racconti di nonni e genitori.

[1] https://csc.dei.unipd.it/roads-project/index.html

3. LA TUTELA DI DATI PERSONALI NEL PROGETTO ROADS: INQUADRAMENTO GIURIDICO E GOVERNANCE

La gestione dei dati personali nel progetto ROADS ha richiesto l'espletamento di specifici obblighi normativi, volti a garantire la piena legittimità dell'attività di ricerca (Abete et al., 2026a).
Data la natura multicentrica e interdisciplinare della ricerca, e in conformità al principio di Accountability (Responsabilizzazione), la governance è stata formalizzata mediante un Accordo di Contitolarità tra i partner del progetto. Tale accordo definisce in modo puntuale ruoli e responsabilità, individua le basi giuridiche appropriate, fondate sull'esecuzione di un compito di interesse pubblico (ex Art. 6, par. 1, lett. E) del GDPR e Art. 2-ter del Codice Privacy), e designa un Punto di Contatto Unico per l'esercizio dei diritti degli interessati (istanze ex Art. 15-22 GDPR). Inoltre, è stata individuata nell'infrastruttura di ricerca CLARIN-IT il custode tecnologico e garante della conservazione a lungo termine (Mercatanti et al., 2026c).
L'attività di ricerca si articola lungo due direttrici: la valorizzazione di archivi orali storici e la raccolta ex novo di testimonianze. Tale distinzione comporta implicazioni rilevanti sul piano giuridico e organizzativo. Nel caso degli archivi preesistenti - spesso privi di un consenso informato conforme agli standard attuali e caratterizzati dall'impossibilità di acquisirlo, per decesso o irreperibilità degli intervistati - il progetto ha adottato un protocollo formale di "Ricerca Diligente" (Diligent Search), che prevede la pubblicazione di avvisi pubblici sui siti istituzionali per informare gli interessati, o i loro eredi, sulla digitalizzazione e sul riuso dei materiali, garantendo il diritto di opposizione, ai sensi dell'Art. 2-terdecies del Codice Privacy italiano. Il bilanciamento tra l'interesse pubblico alla salvaguardia del patrimonio culturale e il diritto alla riservatezza è stato garantito dalla rimozione dei metadati identificativi dei testimoni storici. Inoltre, l'accessibilità ai file audio è limitata a modalità o versioni parziali, al fine di ridurre al minimo il rischio di un uso illecito o non conforme dei contenuti storici.
Per le interviste raccolte ex novo, ROADS adotta un sistema di consenso stratificato, che distingue tra presa visione obbligatoria dell'informativa e opzioni facoltative relative alla divulgazione e al riuso dei materiali (ad esempio distinguendo tra usi didattici interni, usi da parte di terzi per tesi o attività didattiche e pubblicazione scientifica integrale o parziale). Tale approccio consente agli interessati di esercitare un controllo consapevole sui diversi livelli di utilizzo delle proprie testimonianze, in linea con i principi di trasparenza e correttezza.
In coerenza con quanto esplicitato nell'informativa e attraverso la presa visione dei protocolli FAIR da parte dell'interessato, la sicurezza del trattamento è garantita da un sistema di Accesso Multilivello e Accreditato, volto a bilanciare la protezione dell'individuo con le necessità della comunità scientifica. In concreto, mentre una versione curata è disponibile in Open Access per massimizzare la rintracciabilità (findability), i file audio/video integrali sono depositati nel repository certificato ILC4CLARIN (nodo italiano della federazione europea CLARIN). L'accesso a queste versioni integrali è protetto da protocolli di autenticazione forte e federata (Single Sign-On tramite IDEM/Edugain/Shibboleth), prevenendo pratiche di voice-harvesting indiscriminato e garantendo un audit trail completo per l'accountability del trattamento. Il sistema complessivo prevede dunque la pubblicazione dei metadati per finalità di scoperta e attribuzione scientifica, affiancata da un accesso controllato ai contenuti integrali riservato alla comunità di ricerca.

4. IL PROGETTO VOCI DALL'INFERNO: TRA DATI "SENSIBILI" E VALORE STORICO

Voci dall'Inferno è un progetto di ricerca dell'Università di Pisa, avviato nel 2016, che si articola attorno a due obiettivi principali: a) la digitalizzazione e la codifica in XML-TEI del primo corpus digitale di testimonianze *non letterarie* di sopravvissuti ai Lager; b) l'analisi della presenza di lessico e di immagini dantesche all'interno di tali testimonianze (Mercatanti et al., 2025b).
Le testimonianze non letterarie comprendono un insieme di tipologie testuali che si colloca nello spazio intermedio tra oralità e scrittura, quali interviste, diari, racconti autobiografici, memoriali e lettere. Il progetto si fonda sull'ipotesi che molti sopravvissuti, indipendentemente dal loro livello di istruzione, abbiano fatto ricorso in modo istintivo e naturale, al linguaggio e all'immaginario della Divina Commedia come strumento cognitivo ed espressivo per dare forma a un'esperienza altrimenti indicibile (Calderini & Riccucci, 2020).

Il progetto integra ricerca e didattica, coinvolgendo laureandi dei Corsi di Laurea in Informatica Umanistica e in Italianistica dell'Università di Pisa. Si avvale inoltre del supporto di numerose istituzioni scientifiche e infrastrutture di ricerca nazionali e internazionali, tra cui il CDEC,[2] CLARIN-IT,[3] ed EHRI[4].
La realizzazione del corpus digitale consente sia la conservazione e la tutela di una tipologia di testimonianze particolarmente fragile, sia l'applicazione di strumenti e di metodologie computazionali per svolgere attività di studio, analisi e ricerca su tali materiali (Del Grosso et al., 2026).
Attualmente, il corpus comprende 25 testimonianze appartenenti a 20 individui ed è suddiviso in due macro-categorie: fonti orali e fonti scritte. In totale, sono state trascritte e codificate 18 ore e 35 minuti di registrazioni audio-video e 395 pagine di testimonianze scritte.
Il corpus è estremamente eterogeneo, sia in termini di modalità di produzione delle testimonianze sia in termini di provenienza delle fonti, che variano in base al tipo di esperienza vissuta dai testimoni. Esso include infatti sia resoconti di persone che hanno vissuto l'esperienza dei campi di concentramento, sia testimonianze di chi non fu mai deportato, offrendo una visione ampia e articolata del Lager e del periodo del Nazifascismo. La maggior parte delle testimonianze è inedita e costituisce un patrimonio storico di eccezionale valore che richiede adeguate strategie di tutela e di valorizzazione.
Le due macro-classi in cui si ripartisce il corpus determinano anche differenti esigenze rappresentazionali, funzionali e fruizionali dal punto di vista della codifica XML-TEI. Nonostante le differenze formali legate alla tipologia delle fonti, sono stati definiti obiettivi di codifica comuni a entrambi i modelli, orientati a una rappresentazione il più possibile fedele della fonte primaria, sia nella struttura sia nel contenuto (Mercatanti et al., 2025d). Il corpus è accessibile tramite una applicazione web, attualmente non disponibile al pubblico, che consente di catalogare, consultare, interrogare e analizzare tutte le testimonianze dell'archivio, nonché di censire e di studiare la presenza di lessico dantesco e le modalità attraverso cui esso si manifesta nel corpus (Mercatanti et al., 2025a).
La componente orale riveste un ruolo centrale e particolarmente sensibile del progetto, in ragione della presenza di dati personali e biometrici e della delicatezza del contesto storico documentato. Il corpus dispone di 13 testimonianze orali (3 video e 10 audio), di cui: otto[5] provenienti da una raccolta di interviste inedite realizzate tra il 2005 e il 2006 da Anna Segre e Gloria Pavoncello[6]; una[7] realizzata nel 2020 da Marina Riccucci e Laura Ricotti[8]; quattro provenienti da archivi istituzionali (CDEC)[9] e piattaforme pubbliche (YouTube)[10] .

5. L'APPLICAZIONE DEL MODELLO ROADS A VOCI DALL'INFERNO: SFIDE ETICHE E METODOLOGICHE

L'elevato livello di annotazione e la ricchezza informativa del corpus rendono Voci dall'Inferno una risorsa di grande valore per studi interdisciplinari. Tuttavia, come molti archivi storici, il progetto si confronta con criticità legate alla gestione di dati personali e biometrici sensibili, ulteriormente accentuate dalla delicatezza del contesto documentato. Tali fattori hanno finora limitato la possibilità di accesso, pubblicazione e riuso delle testimonianze, incidendo sulla piena valorizzazione scientifica del corpus.
Una delle principali difficoltà riguarda i materiali legacy, per i quali risulta spesso impossibile acquisire un consenso informato conforme agli standard attuali. Ciò rende complessa l'individuazione di basi giuridiche adeguate, soprattutto in presenza di contenuti altamente sensibili e riferimenti a esperienze traumatiche.
In questo quadro, il modello sviluppato in ROADS offre un riferimento metodologico e operativo per affrontare tali criticità, attraverso strategie differenziate per archivi storici e raccolte ex novo. Nel caso di Voci dall'Inferno, l'adozione del protocollo di "Ricerca Diligente" consente di informare gli interessati o i

[2] https://www.cdec.it/
[3] https://www.clarin-it.it/it
[4] https://www.ehri-project.eu/
[5] Si tratta delle interviste realizzate a Edith Bruck, Nedo Fiano, Ida Marcheria, Liliana Segre (2), Piero Terracina, Shlomo Venezia e Idek Wolfowicz.
[6] Cfr. Segre, A., & Pavoncello, G. (A c. Di). (2009). *Judenrampe. Gli ultimi testimoni*. Elliot.
[7] Si tratta dell'intervista realizzata a Goti Herskovits Bauer.
[8] Cfr. Riccucci, M., & Ricotti, L. (2021). *Il dovere della parola. La Shoah nelle testimonianze di Liliana Segre e di Goti Herskovitz Bauer*. Pacini Editore.
[9] Si tratta dell'intervista ad Arminio Wachsberger.
[10] Le testimonianze video in questione appartengono a Enrico Vanzini, Samuel Modiano e Arminio Wachsberger.

loro eredi in merito alla digitalizzazione e al riuso, garantendo il diritto di opposizione. Allo stesso tempo, la dimensione contenuta del corpus - attualmente composto da 13 testimonianze orali – permette un approccio basato su valutazioni caso per caso, in linea con il principio di accountability promosso dal GDPR. La possibilità di rintracciare alcuni testimoni ancora in vita rappresenta inoltre un elemento favorevole per l'acquisizione di un consenso aggiornato. In questo contesto, il modello di consenso stratificato elaborato in ROADS offre uno strumento efficace per modulare i diversi livelli di accesso, diffusione e riutilizzo.
La natura eterogenea delle fonti introduce ulteriori elementi di complessità, ma anche opportunità. Parte delle testimonianze proviene infatti da archivi istituzionali e piattaforme che rendono già accessibili i contenuti. Pur non implicando automaticamente una piena legittimazione al riuso scientifico, tali casi consentono di fondare il trattamento dei dati su basi più solide per attività di consultazione e analisi in contesti di ricerca controllati.
Un ulteriore elemento chiave riguarda la governance. Il modello ROADS prevede la formalizzazione dei rapporti tra i soggetti coinvolti mediante un Accordo di Contitolarità del trattamento dei dati personali, volto a chiarire ruoli e responsabilità e a garantire una gestione coordinata dei dati sensibili distribuiti tra più enti. L'adozione di tale strumento permetterebbe a Voci dall'Inferno di definire in modo esplicito le responsabilità relative alla conservazione, all'accesso e al riuso delle testimonianze, nonché di stabilire procedure condivise per la gestione delle richieste degli interessati.
Nel complesso, l'applicazione del modello ROADS al caso di studio mostra come un approccio FAIR-by-design, supportato da un solido assetto di governance, possa favorire una valorizzazione responsabile degli archivi orali sensibili, consentendo una progressiva transizione verso risorse accessibili e riusabili nel rispetto dei diritti delle persone e del contesto storico in questione.

6. CONCLUSIONI

Il contributo ha mostrato come la gestione degli archivi orali richieda un approccio integrato, in cui dimensioni archivistiche, tecnologiche, giuridiche ed etiche risultano strettamente interconnesse. La natura complessa del dato orale impone modelli capaci di bilanciare la tutela dei diritti degli interessati con le esigenze di accesso e riuso.
In questo quadro, ROADS dimostra come un approccio integrato lungo l'intero ciclo di vita dei dati possa offrire soluzioni concrete. La definizione delle basi giuridiche, insieme all'adozione di strumenti quali accesso multilivello, consenso stratificato e accordo di contitolarità, consente di rendere operativa la FAIRness anche in presenza di dati sensibili e di archivi storici.
Il confronto con Voci dall'Inferno mette in evidenza il carattere trasferibile del modello, mostrando come criticità ricorrenti – quali la presenza di fonti preesistenti, la frammentazione delle responsabilità e le incertezze relative al consenso - possano essere affrontate attraverso adeguati strumenti giuridici e infrastrutturali. In questa prospettiva, tali archivi possono essere progressivamente ricondotti a forme sostenibili di gestione e riusabili. La sostenibilità degli archivi orali emerge così non come questione tecnica ma come esito di una progettazione responsabile, capace di garantire nel tempo la legittimità giuridica e l'integrità scientifica delle risorse.
In questa prospettiva, ROADS si configura come un modello di riferimento per la valorizzazione del patrimonio orale, contribuendo alla costruzione di ecosistemi di ricerca aperti, eticamente fondati e giuridicamente solidi.

BIBLIOGRAFIA

Abete, G., Calamai, S., Canazza, S., Casellato, A., Mercatanti, E., & Monachini, M. (2026a). *La filiera legale di ROADS. Una proposta FAIR per archivi orali analogici.* Zenodo. https://doi.org/10.5281/zenodo.18411522

Abete, G., Calamai, S., Canazza, S., Casellato, A., Mercatanti, E., Monachini, M., Niri, V., Vecchia, C., Zitelli Conti, G., & Zuccolo, G. (2026b). *Fare didattica con gli archivi orali: la fonte orale al crocevia di discipline e di saperi.* https://h2iosc-training-platform.ilc4clarin.ilc.cnr.it/course/1e8b021e-b036-42da-b9d8-1a7d6b0d4dd2/8c1c1b33-92b1-4dfd-a87b-9ea263548d83/749

Abete, G., Vecchia, C., Calamai, S., Casellato, A., Canazza, S., Mercatanti, E., Monachini, M., Ottaviani, R., Zitelli Conti, G., & Zuccolo, G. (2025). On the lifecycle of Italian oral archives: The ROADS project. *La voce della grammatica. Nuove prospettive sull'interazione tra fonetica e morfologia,*

sintassi, lessico. XXI Convegno Annuale AISV, Urbino. https://drive.google.com/file/d/1kDLfbulztuk_ddkAA2oYzKDn1yuOyqfh/view

Calamai, S., & Frontini, F. (2018). FAIR data principles and their application to speech and oral archives. *Journal of New Music Research*, (47), 339–354. https://doi.org/10.1080/09298215.2018.1473449

Calderini, S., & Riccucci, M. (2020). L'ineffabilità della nefandezza: Dante «per dire» il Lager: Un sondaggio preliminare nelle testimonianze non letterarie. *Italianistica: rivista di letteratura italiana: XLIX, 1, 2020.* https://doi.org/10.19272/202001301011

De Dominicis, A. (A c. Di). (2022). *La voce come bene culturale*. Carocci.

Del Grosso, A. M., Mercatanti, E., Congiu, C., & Riccucci, M. (2026, maggio). Integrating TEI Publication, Guided Exploration, and Vector Databases for Semantic Search in the Voci dall'Inferno Project. *Proceedings of the Second Workshop on Holocaust Testimonies as Language Resources (HTRes) @ LREC-COLING 2026.*

Gribaudi, G. (s.d.). *Combattenti, sbandati, prigionieri. Esperienze e memorie di reduci della seconda guerra mondiale*. Donzelli.

Gribaudi, G. (1980). *Mediatori: Antropologia del potere democristiano nel Mezzogiorno.* Rosenberg & Sellier.

Gribaudi, G. (1990). *A Eboli. Il mondo meridionale in cent'anni di trasformazione.* Marsilio.

Gribaudi, G. (2005). *Guerra totale. Tra bombe alleate e violenze naziste. Napoli e il fronte meridionale 1940-1944.* Bollati Boringhieri.

Gribaudi, G. (A c. Di). (2023). *Terra bruciata. Le stragi naziste sul fronte meridionale*. Guida.

Mercatanti, E., Congiu, C., Del Grosso, A. M., & Riccucci, M. (2026a). Voci dall'Inferno: A Web application to study and analyze the Lager testimonies. In G. del R. Riande, M. Portela, D. Alves, & J. V. Paulino (A c. Di), *Digital Humanities 2025: Book of Abstracts*. ADHO. https://doi.org/10.5281/zenodo.19494802

Mercatanti, E., Del Grosso, A. M., & Riccucci, M. (2025b). Voci dall'Inferno: Dante per esprimere l'indicibile: Un'applicazione digitale per esplorare le testimonianze non letterarie dei sopravvissuti ai Lager. *Umanistica Digitale*, (20), 527–562. https://doi.org/10.6092/issn.2532-8816/21262

Mercatanti, E., Monachini, M., Abete, G., Calamai, S., Canazza, S., Casellato, A., Niri, V., Vecchia, C., Zitelli Conti, G., & Zuccolo, G. (2026c). Balancing FAIR and GDPR: A Governance Framework for Oral Archives. *Proceedings of the Workshop on Legal and Ethical Issues in Human Language Technologies @ LREC-COLING 2026.*

Mercatanti, E., Riccucci, M., & Del Grosso, A. M. (2025d). Voci dall'Inferno: A TEI-Based Digital Archive for finding Dante in Concentration Camp Testimonies. In J. Hałaczkiewicz & J. C. for D. Humanities (A c. Di), *«New Territories». Text Encoding Initiative Conference and Members' Meeting 2025*. Zenodo. https://doi.org/10.5281/zenodo.17312233

Monachini, M., & Mercatanti, E. (2026, marzo). Materiali didattici innovativi per le fonti orali: Percorsi interdisciplinari per la scuola. In *Didacta Italia*.

Sornicola, R., Abete, G., D'Argenio, E., & Vecchia, C. (2019). Raccontare un archivio di fonti orali: Il progetto Voci, parole e testi della Campania [DA]. In D. Piccardi, F. Ardolino, & S. Calamai (A c. Di), *Gli archivi sonori al crocevia tra scienze fonetiche, informatica umanistica e patrimonio digitale* (Vol. 6, pp. 75–93). Officinaventuno. https://doi.org/10.17469/O2106AISV000005

RIFERIMENTI NORMATIVI

Regolamento (UE) 2016/679 (GDPR) del Parlamento europeo e del Consiglio, del 27 aprile 2016, relativo alla protezione delle persone fisiche con riguardo al trattamento dei dati personali, nonché alla libera circolazione di tali dati. Link: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.L_.2018.295.01.0039.01.ENG&toc=OJ:L:2018:295:TOC

Decreto Legislativo 30 giugno 2003, n. 196, "Codice in materia di protezione dei dati personali", come integrato e modificato dal D.Lgs. 10 agosto 2018, n. 101. Link: https://www.garanteprivacy.it/documents/10160/0/Codice+in+materia+di+protezione+dei+dati+personali+%28Testo+coordinato%29

Regole deontologiche per scopi storici, statistici o di ricerca scientifica, pubblicate ai sensi dell'art. 20, comma 4, del d.lgs. 101/2018 (Provvedimento del Garante n. 513 del 19 dicembre 2018). Link: https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/9069661
Regole deontologiche per il trattamento a fini statistici o di ricerca scientifica (Provvedimento del Garante n. 515 del 19 dicembre 2018). Link: https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/9069637

STANDARD E PROTOCOLLI DI GOVERNANCE

Linee guida sulla "Ricerca Diligente" (Diligent Search) per la gestione delle opere orfane e dei materiali d'archivio privi di titolarità certa (recepite in ambito ROADS per la gestione dei materiali legacy).

FAIR Memories: A Workflow for the Preservation and Collaborative Reuse of Analog Oral History Collections

Nike del Quercio[1], Costanza Paolillo[2], Laurent Fintoni[3]

[1] Department of Classical Philology and Italian Studies, University of Bologna, Italy – nike.delquercio2@unibo.it
[2] Department of Classical Philology and Italian Studies, University of Bologna, Italy – costanza.paolillo@gmail.com
[3] Department of Classical Philology and Italian Studies, University of Bologna, Italy – laurent.fintoni2@unibo.it

ABSTRACT (ENGLISH)

FAIR Memories is a Horizon Europe-funded project developing a reusable workflow for the FAIR-ification of legacy oral history collections to allow for the transition of these important but underused collections to current Open Science standards. Making use of a unique analog collection about Italian mass culture, the workflow is being developed jointly by domain experts and digital humanists with particular attention to the preservation of the recorded voice as a core epistemic dimension, standards-based interoperability, semantic enrichment, and compliance with contemporary privacy regulations. This poster presents the initial stages of the project, detailing the collection used, the proposed steps of the workflow, and key theoretical approaches.
Keywords: Oral History; FAIR Data; Semantic Web.

ABSTRACT (ITALIANO)

FAIR Memories: Un flusso di lavoro per la conservazione e il riutilizzo delle collezioni analogiche di storia orale.
FAIR Memories è un progetto finanziato da Horizon Europe che sta sviluppando un flusso di lavoro riutilizzabile per la FAIR-ificazione delle collezioni analogiche di storia orale, consentendo la transizione di queste importanti ma sottoutilizzate collezioni agli attuali standard di Open Science. Avvalendosi di una collezione analogica unica sulla cultura di massa italiana, il flusso di lavoro è sviluppato congiuntamente da esperti del settore e umanisti digitali, con particolare attenzione alla conservazione della voce registrata come dimensione epistemica fondamentale, all'interoperabilità, all'arricchimento semantico, nonché alla conformità con le normative contemporanee in materia di privacy. Questo poster presenta le fasi iniziali del progetto, descrivendo la collezione utilizzata, le fasi del flusso di lavoro e principali approcci teorici.
Parole chiave: Storia Orale; FAIR Data; Web Semantico.

1. INTRODUCTION

While oral sources are recognized as critical scholarly and cultural assets (*Archiving Oral History*, 2019), many oral history collections[1] produced before the widespread adoption of digital standards remain fragile, difficult to access, and ethically complex (Bonomo et al., 2016). Furthermore, the digitization and reuse of analog collections continues to be a persistent challenge (Rogers & Mcallister, 2014), one that sits at the intersection of important current themes such as sustainability, reuse, and ethics (Calamai, 2025). These materials, often recorded on frail analog media and governed by consent frameworks that predate contemporary privacy regulations, have high research value but frequently fall outside current Open Science standards without extensive, additional transformative work. Within the context of the humanities, the preservation and reuse of oral history collections cuts across many fields and practices and is particularly relevant to the digital humanities where a multimodal and multidisciplinary approach is well suited to bringing together the different technological, curatorial, and organizational skills needed to manage and valorize these collections (Calamai & Frontini, 2018). Aligning the digital transformation of oral history collections with the FAIR Data Principles (Wilkinson et al., 2016), while not straightforward,

[1] We should note here the inconsistency of labels used when referring to oral history—*archive*, *corpus*, and *collection* but also *oral*, *speech*, and *audio*—which reflects its fragmentation across disciplines (Calamai & Frontini, 2018). In the case of this paper, we have chosen to use oral history collection though archive or corpus could also be applicable.

can make them more accessible both to researchers but also to the wider public, an important aspect as it relates to the products of research in a social sciences and humanities context.
This poster presents FAIR Memories, a Horizon Europe-funded project developing a reusable workflow for the FAIR-ification of legacy oral history collections, allowing for their transition to current Open Science standards. Using a historically important analog oral history collection about Italian popular culture, the workflow is being developed jointly by domain experts and digital humanists with particular attention to the following: preservation of the recorded voice as a core epistemic dimension; standards-based interoperability; semantic enrichment; and compliance with contemporary privacy regulations. Here we present the initial stages of the project, detailing the collection used, the proposed steps of the workflow, and some of our key theoretical approaches, such as the reframing of FAIR-ification as a process of co-produced knowledge.

2. AN ORAL HISTORY OF ITALIAN MASS CULTURE

To achieve its goals, FAIR Memories makes use of a historically significant but underused collection of oral history interviews on Italian popular culture, recorded in the early 1990s for the project *Cultural Industries, Governments and the Public in Italy, 1938–1954* (Forgacs & Gundle, 2010), which was sponsored by the UK's Economic and Social Research Council. This collection consists of 117 interviews with individuals born between the 1910s and 1930s and conducted in various regions of Italy in 1991 and 1992. The goal of the project was to examine relations between cultural production, consumption, and political power before and after World War II. It remains to this day the most extensive and substantially funded oral history research on mass culture in those years ever conducted in the country, the results of which formed the basis for the book *Mass Culture and Italian Society from Fascism to the Cold War* (Forgacs & Gundle, 2007).
Textual components of this collection (over 4,000 pages of photocopied transcripts) were scanned and deposited at the UK Data Service in 2010, however the promise of the addition to this deposit of the audio material (over 70 cassettes) was never completed. The current digital version of the collection is available as safeguarded data, requiring registration with the UK Data Service, a process that is made easier for academic users but not for the general public[2]. Thus, the collection exists in a fragmented digital state. Textual records are technically preserved, but they are not easily accessible to non-academics and are only composed of low-quality digital copies of physical copies of the original typed transcripts, well below accepted standards for archival records. Meanwhile the recorded voices of everyday people remain tied to obsolete analog media, in this case cassette tapes that are also copies of the original recordings, obscuring prosodic and embodied features (intonation, hesitation, pacing, and involuntary speech acts) that constitute the most representative and specific part of oral testimony (Passerini, 2020; Pessanha, 2022; Viswanath, 2023).

3. FRAGMENTED STATES AND ACCESSIBILITY CHALLENGES

As with many similar collections found across the Global North, and in particular Europe, the fragmented state of the collection we intend to work with reflects the inherent politics of digitization projects (Zaagsma, 2023) and highlights persistent challenges: the technological obsolescence of recording media; the uneven remediation of analog materials; the difficulty in applying broad ethical and legal guidelines to oral history collections characteristically defined by their uniqueness; and the gap between formal data deposit and meaningful public or scholarly access. Tackling these challenges requires collaborative, multidisciplinary work in order to go beyond the simple creation of digital surrogates that, like their analog counterparts, may only end up residing in a space where few are aware of their existence or able to access them. We must also find ways to navigate the current legal and ethical requirements that digitalization entails, especially in the European context where GDPR regulation creates various layers of friction for collections of oral testimonies collected before current privacy regulations, and make these findings accessible to others to grow a shared body of knowledge that can be used to facilitate future work and ease tension between ethics, legality, and meaningful reuse. And we must ensure that such digitization efforts also consider the importance of community restitution as central to the work. That is, any effort to bring such collections into our current digital world should include meaningful efforts to involve the communities that helped create them, so as not to deprive them of

[2] https://ukdataservice.ac.uk/find-data/access-conditions/

meaningful resources (Garellek et al., 2020) but also to underline the importance of understanding restitution in this context as a process that is just as valuable and challenging as recovery, use, and reuse of collections (Ardolino et al., 2025). While restitution can take different forms, we believe that at least with making these collections publicly available offers a starting point towards more meaningful conversations and efforts. It is also important to remember that metadata, not just digitization, constitutes a major point of entry to cultural heritage for many (Zaagsma, 2023) by allowing the digitized cultural heritage to be found in the first place. Efforts such as the workflow we are developing must therefore balance preservation with interoperability, contextualization, and access in order to foster potential reuse whether that be within the context of the European Open Science Cloud (EOSC) or in similar efforts elsewhere, both within and outside of academia. In this sense, the Forgacs & Gundle collection functions not merely as necessary data with which to test and develop the FAIR-oriented remediation strategies advocated by our workflow, but as a paradigmatic example of the limits of early digital transitions in the humanities (Zaagsma, 2024) as well as a needed Italian use case to better understand how to act with such collections by ensuring that the decision-making flow is public and replicable, even if applied to collections that may not always end up fully open (Calamai, 2025). It is our hope that the FAIR Memories project can meaningfully contribute to recent efforts within the Italian context — such as the 2022 PRIN and PRIN-PNRR projects *ROads to Oral Archives Development and Sustainability (ROADS)*[3] and *Space and Time in Oral Archive (STOA)*[4] and the *Tavolo permanente per le Fonti Orali* — that seek to find ways to develop shared guidelines to help align oral history collections with FAIR principles and make it easier for others to navigate the ethical and legal requirements of such work in order to make them accessible.

4. FAIR-IFICATION WORKFLOW

The digitization and FAIR-ification workflow proposed by FAIR Memories is conceived, and will be delivered, within the context of the EOSC. As such it is thought of as a form of open method, a type of research product that is more fluid than traditional ones and is integral to the work of the Social Sciences and Humanities cluster within EOSC via its Social Sciences and Humanities Open Marketplace (SSHOMP)[5]. In the SSHOMP, workflows can be thought of as concrete examples of best research practices within a given community[6], focusing on accessible and reproducible processes rather than products. Our goal is to deliver our workflow in two formats: as a narrative workflow, as defined by Bardot et al. (2024) in the context of the SSHOMP, where it can be integrated with other Open Science tools and resources already included in the marketplace, adding further context to the service; and as an interactive workflow featuring conditional logics and a web-based user interface that can more closely mirror real-life situations and allow others to evaluate their own situation and better understand how to adapt the suggested actions, tools, and reflections to their specific needs and situations.
Our proposed workflow will involve the following principal stages: (a) assessment of the existing analog materials and their digitization (focused on text and audio), including recommendations for documenting the history of the collection and possible revision and alignment of pre-existing digitized assets to current norms, based on our experience in working with existing low-quality scans of transcripts; (b) normalization of textual materials using Text Encoding Initiative guidelines to enhance machine readability while preserving features specific to oral discourse; (c) identification and anonymization of potentially sensitive content using OpenAIRE's Amnesia[7], or similar tools, as an example of how to adapt current solutions to oral data collected under pre-digital ethical frameworks; (d) data modelling options that reuse established international standards, including Dublin Core, CLARIN's Component Metadata Infrastructure, and the Europeana Data Model, with particular attention to existing oral history profiles (Calamai et al., 2022), to facilitate interoperability with the EOSC as well as demonstrate how to align oral history collections with the metadata needs of various Open Science infrastructures; (e) semantic enrichment of textual data using the INCEpTION[8] semantic annotation platform, from basic annotation

[3] https://csc.dei.unipd.it/roads-project/index.html
[4] https://sites.google.com/unisi.it/stoaproject/home
[5] https://marketplace.sshopencloud.eu/
[6] https://marketplace.sshopencloud.eu/about/data-population#types-of-content
[7] https://amnesia.openaire.eu/
[8] https://inception-project.github.io/

layers (people, places, dates) to more tailored needs based on collection specifics, such as in our case the use of annotation to reflect on the sociological aspects of our collection; and (f) guidelines for the publication and reuse of the resulting data and evaluation of their FAIR-ness, bringing together existing research and recommendations to further guide the work based on needs and limitations.
As an example of the usage of the workflow, FAIR Memories will transform part of the Forgacs & Gundle collection and make it available both as a FAIR dataset via existing EOSC-integrated repositories, such as CLARIN, as well as via a website aimed at the public using Oral History Metadata Synchronizer[9], allowing user interaction with both audio and textual components. Considering the experiences of the aforementioned ROADS and STOA projects as well as associated research specific to the Italian context of making oral history collections publicly available, we are aware of the ethical and legal challenges that a public website such as the one we are proposing faces. These challenges will be confronted during our research and we hope that it will be possible to produce a public-facing interface for our example dataset that can provide an example for the Italian community to work through these complex issues.
In order to make visible the limits of applying standardized solutions to historically situated and ethically constrained sources, collaboration between digital humanists and domain experts is constitutive of the workflow we are proposing to help underline how the FAIR-ification process for analog oral history collections should be thought of as not just a purely technical operation. Decisions concerning digitization, metadata granularity, transcription practices, anonymization thresholds, semantic enrichment, and access conditions must be shaped jointly by technical constraints, community involvement when possible, and domain-specific interpretive knowledge in order to better confront the inherent fragmented nature of these collections (Calamai & Frontini, 2018). Semantic annotation is a particularly interesting aspect of the workflow in the context of our collection, conceived not merely as a technical enhancement but as a collaborative interpretive practice. Pre- and post-war Italy was a period marked by tumultuous, profound social and economic transformations. As such part of our annotation work will focus on how language use, including the interaction between widespread regional dialects and limited Italian-language proficiency reshaped by internal migration, and references to everyday practices of consumption with clear class connotations were expressions of belonging or exclusion from social groups (Alfonzetti 2009; Forgacs 2014; Wright 2000). In doing so we hope to show how this stage can function as a mediating layer between computational accessibility and cultural interpretation, enabling the re-emergence of historically marginalized perspectives within the constraints of responsible data reuse.

5. CONCLUSION AND FUTURE WORK

Over the next two years, our project is expected to produce the two primary deliverables highlighted in this poster presentation — a reusable workflow for digitization and FAIR-ification of analog oral history collections alongside a sample dataset from the Forgacs & Gundle collection — as well as a website allowing the public to read and listen to the interviews. The narrative version of the workflow will be made available via the SSHOMP while the interactive version will be made available via the University of Bologna's Digital Humanities Advanced Research Center's infrastructure, including a website and repository containing editable and reusable components. Whenever possible we intend to use the workflow as a way to also highlight and connect existing research, tools, and services related to oral histories in order to build on what has been done rather than simply reproduce existing work. We are also in early talks with other researchers and initiatives, such as the German Oral-History.Archive platform and the CLARIN Knowledge Center for Oral Archives in Italy to find ways to meaningfully collaborate with them as well as better understand how our workflow and dataset may be adapted by or useful to other research communities.
Today, with growing attention to research sustainability and Open Science as well as community restitution and marginalized experiences, reactivating analog oral history collections feels both timely and necessary. By taking a multidisciplinary approach informed by the digital humanities, FAIR Memories aims to support existing national and international efforts to enhance the discoverability and reuse of such collections while addressing the cultural and social complexity of this form of testimony and ensuring critically informed scholarly and public engagement.

[9] https://www.oralhistoryonline.org/

ACKNOWLEDGEMENTS

The authors acknowledge the OSCARS project, which has received funding from the European Commission's Horizon Europe Research and Innovation programme under grant agreement No. 101129751.

REFERENCES

Alfonzetti, G. (2009). *Italiano e dialetto tra generazioni, 2008, in G. Marcato. Dialetto. Usi, Funzioni e Forme, CLEUP*. https://www.academia.edu/10219387/Italiano_e_dialetto_tra_generazioni_2008_in_G_Marcato_Dialetto_Usi_Funzioni_e_Forme_CLEUP

Archiving Oral History: Manual of Best Practices - Oral History Association. (2019, November 7). https://oralhistory.org/archives-principles-and-best-practices-complete-manual/

Ardolino, F., Brasile, L., & Calamai, S. (2025). Uno stress test per il Vademecum per il trattamento delle fonti orali: L'esperienza con l'Archivio orale dell'Istituto Interregionale di Studi e Ricerche della Civiltà Appenninica. *L'Analisi Linguistica e Letteraria, 33*(3). https://www.analisilinguisticaeletteraria.eu/index.php/ojs/article/view/795

Barbot, L., Dolinar, M., Gray, E. J., Grisot, C., Illmayer, K., Kurzmeier, M., & McGillivray, B. (2024). Contextualizing Research Tools & Services Through Workflows in the SSH Open Marketplace. *Journal of Open Humanities Data, 10*, 22. https://doi.org/10.5334/johd.192

Bonomo, B., Casellato, A., & Garruccio, R. (2016). Maneggiare con cura: Un rapporto sulla redazione delle Buone pratiche per la storia orale. [*Mestiere di storico : rivista della Società italiana per lo studio della storia contemporanea : VIII, 2, 2016*]. https://digital.casalini.it/10.23744/1324

Calamai, S. (2025). Why This Journal, Why Now. *Oral Archives Journal, 1*, 5–19. https://doi.org/10.36253/oar-3334

Calamai, S., & Frontini, F. (2018). FAIR data principles and their application to speech and oral archives. *Journal of New Music Research, 47(4)*, 339–354. https://doi.org/10.1080/09298215.2018.1473449

Calamai, S., Scagliola, S., Ardolino, F., Draxler, C., Van Hessen, A., & Van Den Heuvel, H. (2022). *Ravensbrück Interviews: How to Curate Legacy Data to Make it CLARIN Compliant*. 1–9. https://doi.org/10.3384/ecp1891

Forgacs, D. (2014). *Italy's Margins: Social Exclusion and Nation Formation since 1861* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781107280441

Forgacs, D., & Gundle, S. (2007). *Mass culture and Italian society from fascism to the Cold War*. Indiana University Press.

Forgacs, D., & Gundle, S. (2010). *Oral History of Cultural Consumption in Italy, 1936-1954*. https://doi.org/10.5255/UKDA-SN-6479-1

Garellek, M., Simpson, A., Roettger, T. B., Recasens, D., Niebuhr, O., Mooshammer, C., Michaud, A., Lee, W.-S., Kirby, J., Gordon, M., & Yu, K. M. (2020). Letter to the editor: Toward open data policies in phonetics: what we can gain and how we can avoid pitfalls. *Journal of Speech Sciences, 9*, 03–16. https://doi.org/10.20396/joss.v9i00.14955

Pessanha, F. (2022). Non-verbal Signals in Oral History Archives. *Proceedings of the 2022 International Conference on Multimodal Interaction*, 668–672. https://doi.org/10.1145/3536221.3557036

Passerini, L., Trakilović, M., & Proglio, G. (Eds.). (2024). *The mobility of memory: Migrations and diasporas across european borders* (First paperback edition). Berghahn.

Rogers, I., & Mcallister, M. (2014). *Ghosts in the archives: Exploring the challenge of reusing memories*. https://acquire.cqu.edu.au/articles/conference_contribution/Ghosts_in_the_archives_exploring_the_challenge_of_reusing_memories/13433291/1

Viswanath, A. (2023). *Multi-modal Emotion Categorization in Oral History Interviews*. Fraunhofer-Gesellschaft. https://doi.org/10.24406/PUBLICA-1893

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR

Guiding Principles for scientific data management and stewardship. *Scientific Data, 3*(1), 160018. https://doi.org/10.1038/sdata.2016.18

Wright, S. (2000). *Community and Communication: The Role of Language in Nation State Building and European Integration*. Multilingual Matters. https://doi.org/10.2307/jj.33169474

Zaagsma, G. (2023). Digital History and the Politics of Digitization. *Digital Scholarship in the Humanities, 38*(2), 830–851. https://doi.org/10.1093/llc/fqac050

Zaagsma, G. (2024). Facing the History Machine: Toward Histories of Digital History. *History of Humanities, 9*(2), 451–491. https://doi.org/10.1086/731827

Ontologycore: A Formal Knowledge Graph of Internet Aesthetics

Anouk Flinkert[1], Ekaterina Krasnova[2], Shiho Nakamura[3]

[1]ALMA MATER STUDIORUM, UNIVERSITY OF BOLOGNA, ITALY - ANOUK.FLINKERT@STUDIO.UNIBO.IT
[2]ALMA MATER STUDIORUM, UNIVERSITY OF BOLOGNA, ITALY - EKATERINA.KRASNOVA2@STUDIO.UNIBO.IT
[3]ALMA MATER STUDIORUM, UNIVERSITY OF BOLOGNA, ITALY - SHIHO.NAKAMURA@STUDIO.UNIBO.IT

ABSTRACT (ENGLISH)

This paper presents Ontologycore, an ontology and knowledge graph designed to capture the structure and creative processes underlying internet aesthetics; cultural, visual, and lifestyle styles such as "Y2K" or "Dark Academia" that emerge and evolve primarily through social media platforms. Applying Rhodes' 4P theory of creativity and the CREON ontology for modeling creativity theories, the project formalizes six prominent internet aesthetics with the aim of answering the following question: to what extent can a formal ontology describe the inherently fluid and multidimensional character of internet aesthetics and the creative processes behind them? The methodology follows a bottom-up approach based on eXtreme Design principles, utilizing Protégé software to develop the ontology. The ontology organizes knowledge around four top-level classes corresponding to the 4P theory: Person (creators, influencers, consumers), Process (Convergence, Divergence, and Meta-convergence), Product (aesthetic elements and creative outputs), and Press (media and platforms). The concept of meta-convergence, the phase in which divergent aesthetic experimentation solidifies into a culturally recognizable style, is central to modeling how aesthetics stabilize over time. Accuracy was evaluated through competency questions and SPARQL queries using an expanded dataset of 124 unique aesthetic element instances across 26 aesthetics. Results indicate that aesthetics share common creative procedures but are distinguishable by their unique element configurations. While the qualitative experience remains an anthropological phenomenon, ontologies can formalize an aesthetics' "creative signature."
Keywords: Ontologies; Knowledge Graphs; Internet Aesthetics; Knowledge Representation; Creativity

ABSTRACT (ITALIANO)

Ontologycore: Un Knowledge Graph Formale degli Internet Aesthetics
Questo articolo presenta Ontologycore, un'ontologia e knowledge graph progettati per catturare la struttura e i processi creativi alla base degli internet aesthetics; cioè stili culturali, visuali e di stile di vita come "Y2K" o "Dark Academia" che emergono e si evolvono principalmente attraverso le piattaforma social media. Applicando la teoria dei 4P della creatività di Rhodes e l'ontologia CREON per la modellazione delle teorie della creatività, il progetto formalizza sei internet aesthetics prominenti con l'obiettivo di rispondere alla seguente domanda: in quale misura un'ontologia formale può descrivere il carattere intrinsecamente fluido e multidimensionale degli internet aesthetics e i processi creativi alla loro base? La metodologia segue un approccio bottom-up basato sui principi dell'eXtreme Design, utilizzando il software Protégé per lo sviluppo dell'ontologia. L'ontologia organizza la conoscenza attorno a quattro classi di primo livello corrispondenti alla teoria dei 4P: Person (creatori, influencer, consumatori), Process (Convergenza, Divergenza e Metaconvergenza), Product (elementi estetici e output creativi) e Press (media e piattaforma). Il concetto di meta-convergenza, ovvero la fase in cui la sperimentazione estetica divergente si consolida in uno stile culturalmente riconoscibile, è centrale nella modelazione di come gli aesthetics si stabilizzano nel tempo. L'accuratezza è stata valutata attraverso domande di competenza e query SPARQL utilizzando un dataset espanso di 124 istanze uniche di elementi estetici. I risultati indicano che gli aesthetics condividono procedure creative comuni ma sono distinguibili per le loro configurazioni uniche di elementi. Mentre l'esperienza qualitativa resta un fenomeno antropologico, le ontologie possono formalizzare la "creative signature" di un aesthetic.
Parole chiave: Ontologie; Grafi di Conoscenza; Internet Aesthetics; Rappresentazione della Conoscenza; Creatività

1. INTRODUCTION

The word "aesthetic" originated as the philosophical discussion about what beauty is, how we should approach it, and why it exists. Later, the field of art history defined aesthetics to refer to a set of principles motivating artists and certain periods of art history. In contemporary digital culture, Millennials and Generation Z started using that term as an adjective that describes what they personally consider beautiful, implicitly assuming styles, subcultures, and thematic trends. Current internet aesthetics (which might be better named "social web aesthetics") can be described as forms of identity in "digital hyperreality" (Ajmal, 2024), following Baudrillard. Baudrillard defines hyperreality as perceptions of a "real" without a reality behind it; a hallucinatory resemblance of the real with itself, different from imagination or fantasy (Baudrillard, 1994). Especially within the social media landscape represented by platforms such as Instagram and TikTok, where visibility and engagement are relentlessly incentivized, we can interpret this as what Sontag mentions; internet aesthetics have moved from a commodity to a performative act of artificial self-presentation (Sontag, 2008).

Another characteristic of internet aesthetics is their accelerated emergence and transformation. Ajmal aligns this phenomenon as a case of Zygmunt Bauman's "liquid modernity", noting that digital culture is inherently characterized by transience (Ajmal, 2024). While some aesthetics spike popularity virally as a meme-effect and fade just as rapidly, others evolve into subcultures, achieving a certain degree of cultural permanence. This research focuses on the latter, specifically examining six prominent aesthetics: Y2K, Grunge, Dark Academia, Cottagecore, Clean Girl, and Old Money, all of which are entries of the open collaboration website Aesthetics Wiki.[1]

To further examine the agents and dynamics involved in the formation of an aesthetic, we drew on Rhodes' 4P theory of creativity. Rhodes suggests that creativity is a phenomenon shaped by four interconnected components: *Person*, *Process*, *Product* and *Press* (Rhodes, 1961). This framework allowed us to analyze different actors and forces within aesthetics. In the case of Clean Girl for example, public figures such as Hailey Bieber function as *Person*, whose visual appearances and personas contribute to consolidation of the aesthetic. Tangible and intangible elements such as the smartphone case Rhode Lip Case produced by Bieber and the "glowy skin makeup look" constitute the *Product*, and social media platforms operate as *Press*, acting as environmental factors that influence the dissemination and perception of aesthetics. The final P, *Process*, leads us to map the creative process behind these phenomena, raising fundamental questions such as: *How do visual elements and cultural contexts converge into a recognized aesthetic?* and *Who decides what an aesthetic is and what is not?*.

From an art historical perspective, aesthetics has much in common with the montage art technique, as they synthesize existing audiovisual, textual or material elements into a new cohesive environment where they coexist naturally. Like montage, internet aesthetics rely on recombination and contextual re-framing. As Boon argues in *In Praise of Copying*, "heterogeneous energies" coalesce into a discrete one once they are unified under a shared name or form (Boon, 2013). This act of labelling is essential for the manifestation of an aesthetic, as it leads to consistently evoking a shared conceptual framework that remains recognizable even in the absence of clearly defined boundaries. The sociological perspective offers an insight by treating aesthetics as a cultural phenomenon of "meaning-making". Through their qualitative analysis, Giolo and Berghman draw on Barthes' concept of "cultural myth", where elements, or signifiers, perform the original social meanings they carry while receiving a novel layer of meaning (Giolo & Berghman, 2023; Barthes, 1957). Internet aesthetics thus function as semiotic systems in which familiar elements are re-contextualized to express new identities, affects, and values.

To formalize these creative semiotic systems, we utilized CREON, an ontology for modeling creativity theories (Lucifora et al., 2023). The above-mentioned Rhodes' 4P theory is formalized in CREON as a core design pattern for creativity, as validated by experimental research (Lucifora et al., 2025a). Another fundamental theory of creativity, Guilford's convergent/divergent thinking, is formalized as the backbone of creative production in a dynamic context, as evaluated by experimental research (Lucifora et al., 2025b). CREON characterizes the phenomenon of "meta-convergence" that helps us understand what is happening when aesthetics defunctionalize existing concepts (and their referents) in a new environment. Meta-convergence refers to the final phase of a creative process, after the disruptive expanding phase of divergence from an established cultural convergence, by solidifying and refining the resulting divergent product as a novel convergence. It is the elaboration phase that refines aesthetic expressions into

[1] https://aesthetics.fandom.com/wiki/Aesthetics_Wiki

culturally recognizable and contextually meaningful concepts. Incorporating this notion into our research allowed us to represent aesthetics as dynamic outcomes of creative processes.
Using the Aesthetics Wiki as a primary data source, we profiled six selected aesthetics according to the 4P model and formalized their creative processes using the CREON ontology.
The central research question of this study is: *To what extent can formal ontologies capture the multifaceted and transient nature of internet aesthetics and their underlying creative processes?* To address this, Ontologycore seeks to formalize what we term the "creative signature" of an aesthetic: the reproducible, structural patterns of elements and processes that define its recognizable identity. By mapping these digital-native phenomena onto a structured framework, we aim to demonstrate that while the qualitative experience of an aesthetic remains a nuanced anthropological phenomenon, its underlying logic can be systematically modeled to reveal how digital trends are constructed and sustained.

2. METHODOLOGY

Our methodology for developing the ontology of internet aesthetics followed a bottom-up, iterative approach. Rather than automated knowledge extraction, we began by constructing manually from a selected set of aesthetics from the Aesthetics Wiki, which was then refined, validated, and formalized into a knowledge graph with expansion of newly generated instances.

2.1 EXTREME DESIGN

In constructing the ontology, we referred to the principles from eXtreme Design (XD, Presutti et al., 2009.), attempting to apply them to the modeling of aesthetics as cultural and creative phenomena. This methodology focuses on iterative development driven by Competency Questions (CQs) and the reuse of Ontology Design Patterns (ODPs). In practice, we adapted XD principles to the aesthetic domain by beginning with a concrete set of six core aesthetics - Y2K, Cottagecore, Dark Academia, Grunge, Old Money, and Clean Girl. Their selection provided a diverse yet manageable basis for initial modeling. Development proceeded in short, collaborative cycles within the Web Protégé and Protégé environment, where each iteration involved designing an element of the ontology, creating relations, and refining the structure. This process transformed broad aesthetic concepts into a formal, computable schema through disciplined, team-based negotiation and gradual improvement.

2.2 USER STORIES AND COMPETENCY QUESTIONS

Guided by the XD methodology, we aimed to bridge the gap between formal ontology and its practical application. The core of the XD is that every part of the ontology must be built to answer a specific Competency Question (CQ). In order to obtain an external perspective on the approach to creating our project, we used Large Language Models (LLM), specifically the Gemini 1.5 Flash Model to assess the relevance of competency questions based on existing data and specific parameters of the "user" we imply, so LLMs could act as an observer of the ontology. By prompting an LLM to act as a specific stakeholder (e.g., a social anthropologist), we generated realistic user stories and contextual inquiries. For instance, an LLM acting as a researcher outlined a mixed-methods analytical approach, prompting us to consider questions of temporal context (*What are the temporal contexts associated with each aesthetic?*), element characterization (*Which aesthetic elements characterize particular aesthetics?*), and emotional profiling (*Which emotions are involved in different aesthetics?*).
Crucially, in the XD framework, CQs also function for test-drives (Presutti et al., 2009). For each CQ, we defined a corresponding SPARQL query with an expected answer. In parallel with the modeling of classes and properties described in the following section, the ontology development became an iterative cycle of implementing a module, running these query-tests, and refining the model until it passed. This process ensured that every class and property we created served a direct, verifiable purpose for a potential user, making the ontology a functional tool rather than an abstract construct.

2.3 CORE MODELING - CLASSES AND PROPERTIES

The structural design of Ontologycore was initiated by defining a comprehensive set of classes and properties that capture the nature of internet aesthetics. We initially identified a wide range of potential properties (over 100 in an initial draft phase) to connect aesthetics with their core elements, resources, context, media etc. Subsequently, we systematically consolidated and generalized these properties to create a more organic and informative structure. For instance, specific properties like `hasVisualElement` or `hasAudioElement` were replaced with the more versatile and semantically concise property

characterizedBy. To refine the ontology, we first utilized the reasoner HermiT, which highlighted inconsistencies and redundancies that helped us in the creation of clearer and more systematic connections. In addition to this, LLM assistance contributed in identifying semantically overlapping or redundant elements (cases where properties appeared structurally distinct but conveyed equivalent meaning in context) and which fall outside the scope of automated reasoning.
Furthermore, it was critically important not only to describe an aesthetic in terms of its external characteristics, but also to incorporate the previously mentioned 4P framework (Person, Process, Product, and Press) into the ontology's structure. Therefore, these four components were distinguished as separate top-level classes, each encompassing relevant subclasses. Thus, the *Person* class included both the creators of the aesthetics, influencers, and consumers. *Process* was divided into three main subclasses: Convergence, Divergence, and Metaconvergence. *Product* encompassed both the direct "products" of an aesthetic and its aesthetic elements. Finally, *Press* included media and technologies by which the aesthetic is created and popularized.
The core classes are illustrated through the knowledge graphs excerpts as presented in Figure 1. In this directed graph, the Aesthetic node functions as a central hub, maintaining the highest degree of incoming and outgoing relations. This dominance is established through its many object properties and its role in interconnecting all top-level classes. According to this conceptual model, instances such as DarkAcademia, Grunge, CleanGirl, Y2K, Cottagecore, and OldMoney are defined through their relationships with other top-level classes. Figure 2 demonstrates excerpts of aesthetic elements linked to the DarkAcademia aesthetic. As shown, the aesthetic is characterized by specific aesthetic elements such as Classical Literature, Candles, associated with emotions such as Melancholy, and disseminated through media platforms such as Tumblr, providing a general picture in understanding the essences in an aesthetic. Note that for visual clarity, the links related to *Process* are omitted in Figure 2; these are detailed in the subsequent process-focused visualization.

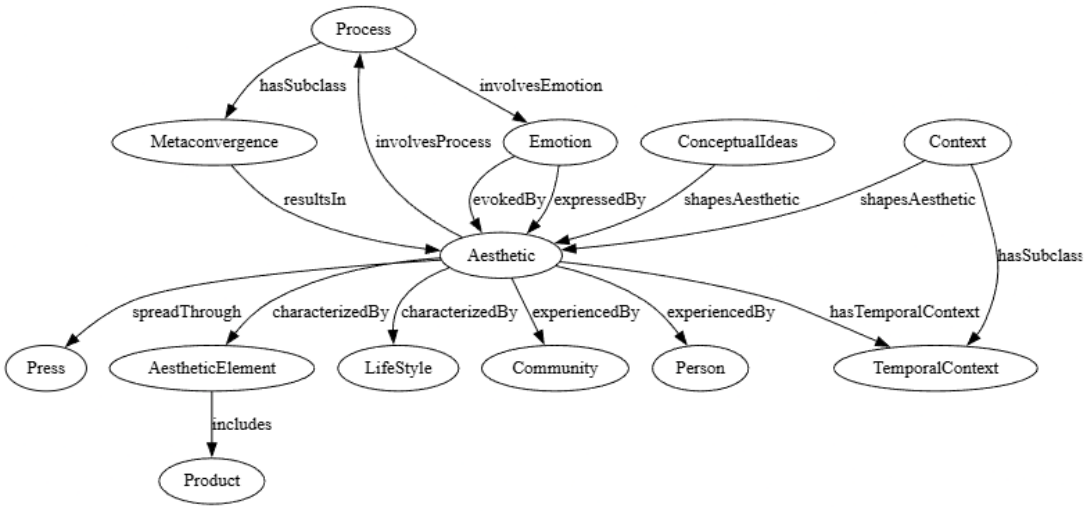


Fig. 1 The core classes of the ontology

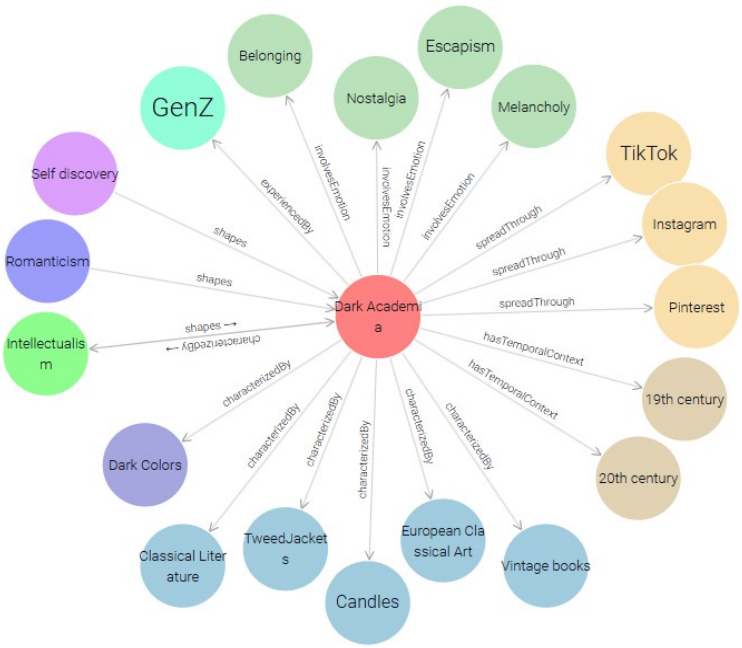


Fig. 2 Instantiation of the DarkAcademia aesthetic within the ontology

2.4 KNOWLEDGE GRAPH POPULATION AND EXPANSION

Building upon the refined ontology, we proceeded onto expanding the knowledge graph by extracting and organizing a larger dataset from the Aesthetics Wiki. This step allowed us to validate the core model against a wider corpus of community-defined aesthetics, test its scalability, and populate it with extensive instances, thereby transitioning from a prototypical framework to a populated knowledge base ready for analysis and application

Using the six manually curated core aesthetics and their creative processes as benchmarks, we utilized LLMs to generate instances and map from an additional 20 new representative aesthetics into the Ontologycore knowledge graph. The robustness of the ontology was evaluated through this process. Beyond accuracy testing via SPARQL queries, the LLM was tasked with providing comprehensive descriptions of the aesthetics by establishing relationships with existing class instances. This approach tested the ontology's external validity and its capacity to categorize new, real-world phenomena within the predefined conceptual framework.

3. RESULTS

The resulting Ontologycore consists of 67 classes and 41 properties. Through the expansion of the knowledge graph, we populated the model with 339 unique instances from 26 aesthetics, resulting in a total of 2588 triples[2].

Besides quantitative metrics, our structural analysis reveals how the ontology formalizes the evolution of digital trends. As illustrated in Figure 3, the relation between the Aesthetic and Process classes is mediated by the top-level object property involvesProcess Within the CreativeProcess hierarchy, specific axioms define the sequential order of Divergence, Convergence, and Metaconvergence. The final stage Metaconvergence (defined as: the critical phase where experimental aesthetic elements are optimized for coherence and cultural impact) closes the dicyclic directed graph, linking the output back to the Aesthetic class and enabling a self-reinforcing loop.

Figure 4 demonstrates the instantiation of this creative process model with the aesthetic Dark Academia as an example. It illustrates how the three phases Divergence, Convergence, and Metaconvergence have specific sub-process instances. Within the Divergence phase, activities of Inspiration, Ideation, and Bricolage are identified as foundational drivers. These transition into the Convergence phase,

[2]Full dataset of the ontology can be found in our GitHub repository: https://github.com/digitalctrlv/ontologycore

characterized by Selection and Synthesis, before reaching the Metaconvergence phase, where Recontextualization and Reproduction facilitate the aesthetic's cultural stabilization and cyclic reinforcement. Regarding the data source for these process instances, the creative processes for the six initial core aesthetics (including Dark Academia) were manually selected by the authors based on descriptions from the Aesthetics Wiki. In contrast, the process mappings for the newly expanded set of aesthetics were generated via the LLM-driven pipeline, maintaining structural consistency with the original curated instances.

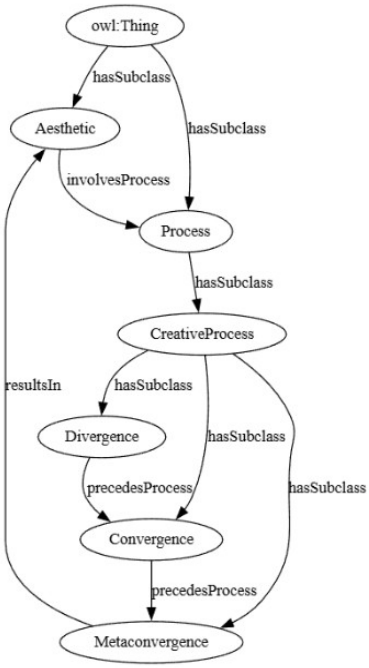


Fig.3 Model of Aesthetics' Creative Process

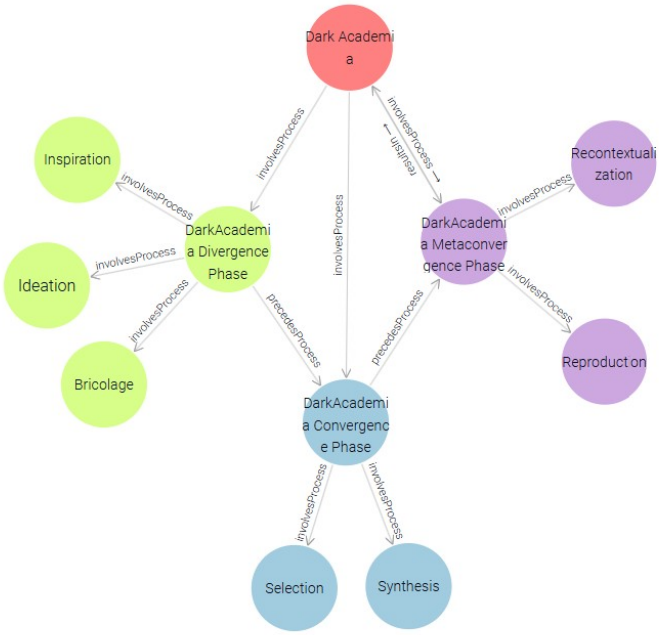


Fig.4 Creative Processes involved in DarkAcademia

As a further analysis, the results of our dataset (containing 124 unique AestheticElement instances) reveal a trend in how internet aesthetics are structured in Ontologycore. Our analysis indicates that aesthetics share abstract mechanisms more frequently than concrete elements. Specifically, the creative processes Inspiration (present in all 26 aesthetics) and Selection (present in 22) appear to be nearly universal core processes.

Therefore, our findings lead to the conclusion that aesthetics are currently more distinguishable by their unique AestheticElements than by their variant creative processes. While we identified specific clusters of specific combinations ("creative signature"), such as the cycle cluster shared by LightAcademia and Royalcore (Inspiration, Recontextualization, Refinement, Reproduction, Selection), the overall process signatures remain relatively uniform across the domain.

A possible explanation is that modern aesthetics emerge from a shared socio-technical background involving the same top-level classes: Person, Press, and TemporalContext. Nevertheless, this lack of process differentiation also suggests a need for ontological refinement. To create more meaningful and unique descriptions of the creative journey per aesthetic, the ontology requires a finer granularity.

4. DISCUSSION

Overall, this work provides a robust framework that functions as a classifier for aesthetics and can generate new, coherent aesthetic types. The model confirms that while aesthetics share abstract creative processes, they derive their unique identity from the specific, concrete elements they synthesize. By formalizing these relationships, we have demonstrated that the transient nature of internet folklore can be systematically modeled without stripping away its multifaceted cultural essence. Future work could focus on strengthening the logical mapping between creative processes and other key factors, such as Press (environmental/media factors) and Emotions, to enhance the ontology's predictive and explanatory

power regarding how aesthetics are built. Additionally, aligning Ontologycore with established upper ontologies and frameworks, such as CIDOC-CRM, would ensure greater interoperability within the broader cultural heritage and digital humanities domains. Ultimately, this study serves as a foundational step toward a formal computational understanding of the "creative signature" that defines our digital cultural landscape.

ACKNOWLEDGEMENTS

We would like to express our gratitude to our supervisor, Professor. Aldo Gangemi, for guidance and encouragement throughout this project. We acknowledge the use of generative AI, namely Large Language Models (Gemini 1.5 Flash) to generate use-case scenarios following the eXtreme Design principles, to generate instances of internet aesthetics and to validate the accuracy of the SPARQL query language and the ontology model.

REFERENCES

Ajmal, S. (2024). Charting youth subcultures, identity formation, and niche aesthetics in the age of social media. Journal of English Literature and Cultural Studies, 5(1), 32–41. https://doi.org/10.26655/JELCS.2024.01.4

Barthes, R. (1957). Mythologies. Éditions du Seuil.

Bauman, Z. (2000). Liquid Modernity. Polity Press.

Baudrillard, J. (1994). Simulacra and Simulation (S. Glaser, Trans.). University of Michigan Press.

Boon, M. (2013). In praise of copying. Harvard University Press.

Giolo, G., & Berghman, M. (2023). The aesthetics of the self: The meaning-making of Internet aesthetics. First Monday. https://doi.org/10.5210/fm.v28i3.12723

Lucifora, C., Scorolli, C., & Gangemi, A. (2023). CREON: A creative ontology based on psychological and neuroscientific studies. In Proceedings of CREAI@AI*IA.

Lucifora, C., Scorolli, C., & Gangemi, A. (2025a). Empirical exploration of the 4P theory of creativity using virtual reality. *Acta Psychologica*, *258*, 105243.

Lucifora, C., Gangemi, A., & Scorolli, C. (2025b). Creative thinking: Immersive, interactive, and multi-sensory experiences. *Creativity Studies*, 1-21.

Presutti, V., Daga, E., Gangemi, A., & Blomqvist, E. (2009). eXtreme design with content ontology design patterns. In Proceedings of the Workshop on Ontology Patterns (WOP).

Rhodes, M. (1961). An analysis of creativity. Phi Beta Kappa, 42, 305–310.

Sontag, Susan. (2008). On Photography. Penguin Classics.

Il ruolo del Test Plan per la validazione: il caso d'uso del Marketplace H2IOSC

VITTORIA FABIANI

CNR ISTITUTO PER IL LESSICO INTELLETTUALE EUROPEO E LA STORIA DELLE IDEE (CNR-ILIESI), ITALY-VITTORIA.FABIANI@CNR.IT

ABSTRACT (ITALIANO)

L'implementazione di un progetto di discipline umanistiche, come una piattaforma digitale di una infrastruttura di ricerca, comporta un processo in più fasi che nella maggioranza dei casi implicano il design delle interfacce e conseguentemente l'architettura e la configurazione di backend e di frontend. All'interno di questo processo di sviluppo, il Test Plan è il documento che assicura l'affidabilità della piattaforma e dei risultati forniti alla comunità scientifica. Esso si inserisce in un flusso documentale volto a sostenere l'intero ciclo di vita del progetto: dal design (modellazione dei dati) al monitoraggio dello sviluppo tecnologico, fino alla validazione e alla conservazione a lungo termine dei risultati.
Questo lavoro nasce dal caso d'uso sviluppato per la progettazione del Marketplace di H2IOSC[1] (Humanities and Cultural Heritage Italian Open Science Cloud), una piattaforma a servizio delle DH, dedicata a raccogliere e disseminare i prodotti della ricerca umanistica.
Parole chiave: metodologia; infrastrutture di ricerca; standardizzazione; validazione; affidabilità.

ABSTRACT (ENGLISH)

The role of the test plan in validation: the H2IOSC Marketplace use case
The implementation of a humanities project, such as a digital research infrastructure platform, involves a multi-stage process that in most cases includes interface design and, consequently, the architecture and configuration of both the backend and frontend.
Within this development process, the Test Plan is the document that ensures the reliability of the platform and the results provided to the scientific community. It is part of a documentation workflow designed to support the entire project lifecycle: from design (data modeling) to monitoring technological development, through to validation and the long-term preservation of results.
This work stems from the use case developed for the design of the H2IOSC[1] (Humanities and Cultural Heritage Italian Open Science Cloud) Marketplace, a platform serving the Digital Humanities (DH) field, dedicated to collecting and disseminating the products of humanities research.
Keywords: methodology; research infrastructure; standardization; validation; confidence.

1. INTRODUZIONE

1.1 Stato dell'arte

Il documento Test Plan rappresenta un elemento fondamentale nella gestione e nella validazione dei progetti scientifici, in particolare nei contesti in cui la riproducibilità e l'affidabilità dei risultati assumono un ruolo centrale per la riduzione sistematica degli errori sperimentali. Infatti, in un contesto scientifico, l'errore, oltre a rappresentare un "bug" informatico, è un fattore che può invalidare il lavoro di ricerca compromettendo l'intero studio.
Esso costituisce sia uno strumento di pianificazione che di validazione di un sistema, di un'applicazione o di un processo tecnologico. Inoltre, è un mezzo di coordinamento, che consente di allineare le attività dei diversi attori coinvolti (sviluppatori, analisti, tester, responsabili di qualità e ricercatori) e di garantire la coerenza con i requisiti definiti inizialmente come esigenza di ricerca. Infine, il Test Plan consente di misurare lo stato di avanzamento delle attività di test, rilevare eventuali criticità e rappresenta anche un documento di tracciabilità e conformità ai requisiti progettuali definiti dai ricercatori, come pure garantisce il rispetto degli standard normativi e di qualità (ad esempio ISO/IEC 29119, IEEE 829, ISO 9001) per supportare eventuali processi di audit o certificazione.
In molti ambiti scientifici ci sono modelli di test plan definiti da standard riconosciuti a livello internazionale come, ad esempio, il test plan descritto dallo 'Standard for Software Test Documentation'

(IEEE 829), i modelli ECSS definiti per lo spazio oppure il TP descritto in ISO 13485 (nell'ambito dei Dispositivi Medici).

1.2 Il flusso documentale

Il primo passo, nel progettare un ecosistema digitale (come per esempio il Marketplace di H2IOSC o altro progetto di ricerca), è definire i requisiti funzionali che derivano dalle esigenze espresse dalla comunità scientifica coinvolta nella definizione del progetto di ricerca. Essi devono descrivere non solo le operazioni tecniche del sistema, ma anche le modalità di interazione dell'utente con il contenuto culturale.
I requisiti forniscono una descrizione accurata di ciò che il sistema deve fare. A seconda del livello di dettaglio, essi danno luogo ad uno specifico documento. I principali documenti del flusso documentale del processo di specifica di un progetto sono descritti nella figura seguente (vd Fig.1).

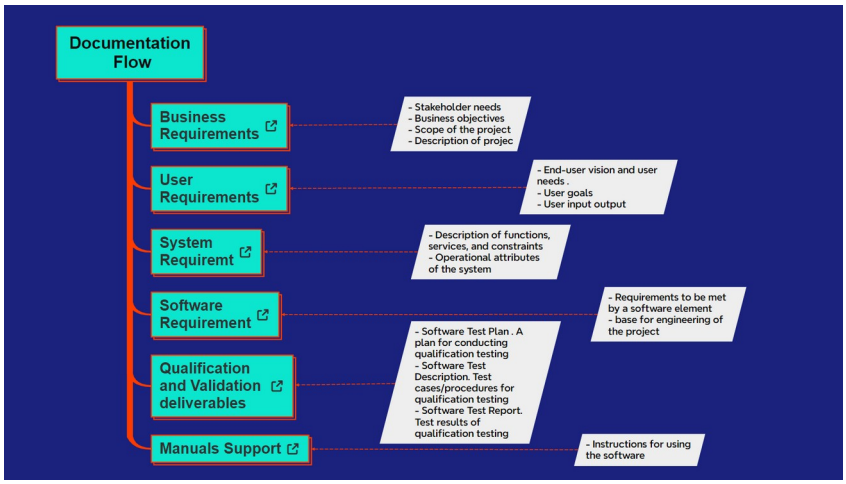


Figura 1. Flusso della documentazione di progetto

La redazione del documento di pianificazione dei test cioè il "Test Plan" si inserisce tra i documenti definiti di Qualificazione e Validazione (Qualification and Validation in Fig. 1).
Esso descrive in modo dettagliato l'approccio, le risorse, gli strumenti e i criteri di accettazione necessari per la conduzione dei test e delle verifiche di un progetto e trova ampia applicazione anche nell'ambito delle discipline umanistiche, dove la documentazione rigorosa è un elemento essenziale.

2. IMPORTANZA DEL TEST PLAN NEL CONTESTO SCIENTIFICO: UN MODELLO PER LE SSH

Il Test Plan riveste un ruolo importante nel contesto scientifico delle Digital Humanities poiché costituisce uno standard che specifica lo scopo, l'approccio, le risorse e il programma delle attività di test che validano un progetto di ricerca. Infatti, nelle SSH non si può sempre garantire replicabilità nel senso stretto (stessi risultati), ma si può garantire: trasparenza, grazie alla documentazione, controllabilità, grazie alla tracciabilità e affidabilità, grazie alla validazione. Quindi esso funge da strumento per il:

- Controllo della Qualità dei Dati: attraverso la definizione di criteri di accettazione e protocolli di verifica, il Test Plan assicura che i dati raccolti siano accurati, completi e coerenti. Ciò riduce la probabilità di errori sistematici e bias sperimentali.
- Supporto alla Comunicazione e alla Collaborazione: per i team di ricerca, il Test Plan funge da riferimento comune, migliorando la comunicazione tra ricercatori, tecnici e stakeholder, e assicurando un linguaggio condiviso sui criteri e sui metodi di test.
-

Nel contesto dello sviluppo di infrastrutture di dati, come nel caso del Marketplace di H2IOSC (sviluppato dal gruppo di lavoro del WP5[1]), il Test Plan, rappresenta un protocollo di validazione che garantisce l'integrità e la riproducibilità dei risultati. Poiché, a differenza di una piattaforma commerciale, una

[1] Workpackage 5 composto da Fabiani, V., Marras, C., Pasini, E., Sichera, P.

piattaforma per le Digital Humanities deve essere "scientificamente rigorosa", documentazione, tracciabilità e validazione assumono un ruolo cruciale nel rafforzare la solidità e la credibilità della ricerca.

Per chiarire meglio: la documentazione accurata consente di esplicitare il processo che conduce ai risultati, rendendo trasparenti le decisioni metodologiche adottate. In particolare, essa implica la formalizzazione dei criteri di selezione dei dati trasformando l'interpretazione da atto soggettivo a processo argomentabile e valutabile; la tracciabilità, consente di collegare dati e risultati, permette di ricostruire l'intero percorso di ricerca, dai dati grezzi fino ai risultati finali. Questo aspetto risulta particolarmente rilevante nelle SSH, dove i dati sono oggetto di rielaborazioni successive (quali le normalizzazioni) e le analisi dipendono dalle scelte interpretative. La tracciabilità garantisce quindi che ogni risultato sia verificabile; la validazione, in questo contesto, compensa la minore robustezza della replicabilità, spesso difficile da garantire nei progetti in ambito umanistico, configurandosi invece come solidità interpretativa e apertura al confronto critico. Essa non si basa tanto sulla riproduzione identica dei risultati, quanto su processi di verifica condivisa e trasparente. Ad esempio, si può realizzare attraverso il confronto tra annotatori, che permette di mettere in discussione, le interpretazioni, contribuendo così a forme di affidabilità alternative rispetto alla replicazione rigorosa
Nel complesso, documentazione, tracciabilità e validazione, pur non assimilando le SSH alle scienze naturali, dove la replicabilità forte è spesso un criterio centrale, rafforzano il rigore metodologico in coerenza con la specificità delle discipline umanistiche.

L'aderenza ai principi FAIR

L'adozione di un Test Plan e di un flusso strutturato dei requisiti contribuisce in modo significativo ad avvicinare i progetti SSH ai principi della scienza aperta, in particolare ai FAIR principles (Findable, Accessible, Interoperable, Reusable). Rendendo esplicito il passaggio dagli obiettivi di ricerca ai requisiti funzionali, fino ai microtask e ai test case, questo approccio permette di documentare la costruzione del dato e le scelte metodologiche sottostanti. Ne deriva una maggiore trasparenza e tracciabilità del processo, che non solo facilita la verifica e la valutazione critica dei risultati, ma ne aumenta anche la riusabilità e l'interoperabilità

In particolare, nell'ambito della *Findability*, un Test Plan ben strutturato consente di organizzare in modo sistematico requisiti, casi di test e risultati, attraverso l'assegnazione di identificatori univoci e la definizione esplicita delle relazioni tra gli elementi. Ciò non solo facilita il reperimento dei dati, ma rende individuabile e navigabile anche il processo che ha condotto alla loro produzione, ampliando il concetto stesso di "trovabilità".
In relazione alla *Accessibility*, la documentazione del workflow permette l'accesso non soltanto ai risultati finali, ma anche alle procedure che li hanno generati. Questo livello di trasparenza consente ad altri ricercatori di comprendere le modalità di utilizzo e le condizioni di interpretazione dei dati.
La dimensione dell'*Interoperability* è rafforzata dalla formalizzazione del flusso dei requisiti, che implica l'adozione di standard condivisi in termini di terminologia, formati e relazioni semantiche. Tale standardizzazione facilita l'integrazione tra progetti e strumenti differenti, risultando particolarmente rilevante nei contesti interdisciplinari, come le Digital Humanities, in cui la comunicazione tra sistemi eterogenei rappresenta una condizione essenziale per ampliare gli obiettivi della ricerca.
Infine, rispetto alla *Reusability*, il contributo di un Test Plan esplicito risulta determinante. La definizione chiara di condizioni, vincoli e criteri di validazione consente ad altri ricercatori di riutilizzare dati e metodologie con consapevolezza delle assunzioni sottostanti e dei limiti applicativi.
Nel complesso, l'integrazione di un Test Plan nel ciclo di ricerca contribuisce a rendere operativi i principi FAIR promuovendo trasparenza, riproducibilità metodologica e sostenibilità del riuso scientifico.

3. IL CASO DI STUDIO DEL MARKETPLACE DI H2IOSC

Per il Marketplace di H2IOSC, progetto che ha come obiettivo la costituzione di un cluster di infrastrutture federate, il documento di validazione è lo strumento operativo che consente di certificare, tramite procedure ben strutturare, le funzionalità della piattaforma del Marketplace.
Una funzionalità da pianificare in fase di test è, ad esempio, l'Harvesting. Il test deve verificare la visualizzazione dell'elenco dei lavori con opzioni di ricerca, filtro e ordinamento, la visualizzazione del

fornitore di riferimento, la data di pianificazione, lo stato (pianificato, completato, ecc.), nonché le operazioni di modifica, duplicazione ed eliminazione del lavoro. In questo caso di test, l'esito negativo può essere causato dalla incompatibilità con il modello dati della piattaforma.

Un altro esempio di funzionalità è la gestione dei Workflow. In questo caso, il test controlla che un Workflow applicativo, creato tramite WSO2 Micro Integrator, venga correttamente caricato sul Marketplace; controlla inoltre che, dopo l'esecuzione, il Workflow restituisca un risultato oppure un log con la descrizione degli errori riscontrati, come il seguente:

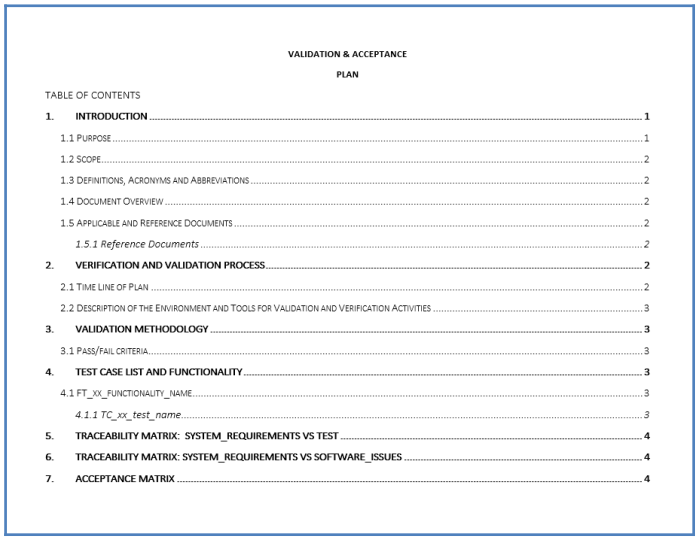
Errors: 1
Error validating item 001:
Validation failed
/properties/0/conceptIds: [en] Duplicate ids found | [it] Sono stati trovati ID duplicati.

Qualora si riscontri un esito negativo, indicativo di una non conformità rispetto ai requisiti iniziali, si procede con una segnalazione formale che attiva un intervento correttivo. Il successivo ciclo di re-testing permette di verificare la risoluzione dell'anomalia, aggiornando l'esito finale a 'SUPERATO' e validando, in tal modo la funzionalità testata. Questo processo di iterativo di raffinamento è ciò che garantisce il rispetto dei principi FAIR, poiché assicura che la piattaforma non sia solo disponibile, ma effettivamente funzionante e affidabile

Il modello per il Marketplace

Questo lavoro propone per il caso di studio del Marketplace di H2IOSC uno schema che riprende una struttura definita da standard ufficiali già consolidati in ambiti scientifici. Il template adottato (Vd. Fig.2) comprende il seguente insieme strutturato di sezioni:

- **Obiettivi dei test** – definizione chiara degli scopi dei test implementati.
- **Ambito e criteri di validazione** – scelta delle condizioni che definiscono le procedure di test.
- **Pianificazione temporale** – cronoprogramma delle attività, con milestone e deliverable.
- **Risorse e strumenti** – elenco di hardware, software, materiali e personale coinvolto.
- **Strategia di test** – descrizione metodologica (test funzionali, prestazionali, di non regressione, ecc.).
- **Criteri di accettazione** – definizione dei valori soglia o degli indicatori di successo dei test.
- **Descrizione dei test** – descrizione delle funzionalità e dei test case oggetto di validazione.
- **Tracciabilità** –tracciamento bidirezionale Requisito-test case e requisito-issue di sviluppo.



VALIDATION & ACCEPTANCE

PLAN

TABLE OF CONTENTS

1. INTRODUCTION 1
1.1 PURPOSE 1
1.2 SCOPE 2
1.3 DEFINITIONS, ACRONYMS AND ABBREVIATIONS 2
1.4 DOCUMENT OVERVIEW 2
1.5 APPLICABLE AND REFERENCE DOCUMENTS 2
1.5.1 Reference Documents 2
2. VERIFICATION AND VALIDATION PROCESS 2
2.1 TIME LINE OF PLAN 2
2.2 DESCRIPTION OF THE ENVIRONMENT AND TOOLS FOR VALIDATION AND VERIFICATION ACTIVITIES 3
3. VALIDATION METHODOLOGY 3
3.1 PASS/FAIL CRITERIA 3
4. TEST CASE LIST AND FUNCTIONALITY 3
4.1 FT_XX_FUNCTIONALITY_NAME 3
4.1.1 TC_xx_test_name 3
5. TRACEABILITY MATRIX: SYSTEM_REQUIREMENTS VS TEST 4
6. TRACEABILITY MATRIX: SYSTEM_REQUIREMENTS VS SOFTWARE_ISSUES 4
7. ACCEPTANCE MATRIX 4

Figura 2. Struttura del documento Test Plan

Dal requisito funzionale al dato: rendere “ispezionabile” l’intero processo del progetto

Il processo di sviluppo del MarkePlace di H2IOSC si configura come un percorso metodologico che parte dalle linee guida di alto livello di un capitolato tecnico per arrivare alla definizione di attività operative concrete e verificabili detti microtask (vd. Fig. 3).

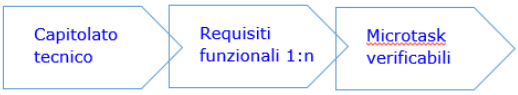


Figura 3. Fasi di progettazione

Tutto ha inizio con l'estrazione dei requisiti funzionali da soddisfare, fase cruciale poiché ogni singola affermazione del progetto è analizzata e mappata: l'obiettivo è garantire che ad ogni punto del capitolato corrisponda ad almeno uno (o anche più) requisiti specifici, assicurando così una copertura totale delle richieste.
Una volta definiti, questi requisiti vengono declinati in microtask dettagliati in modo da poter essere e testati con procedure operative (vd. Fig.4).

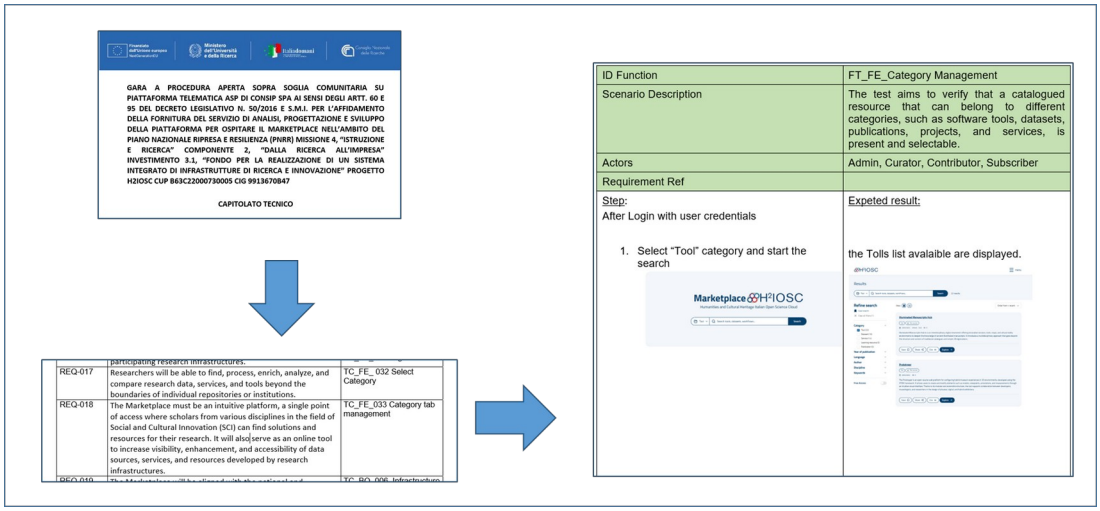


Figura 4. Esempio di scomposizione del capitolato

A partire da questa scomposizione in microtask, gli sviluppatori possono procedere con la fase di implementazione vera e propria, lavorando su unità di codice chiare, tracciabili, testabili e coerenti con le aspettative iniziali, come mostra la bidirezionale dei requisiti descritta nel modello all’interno dei paragrafi 5, 6 e 7, dedicati alle matrici di tracciabilità (vd. Fig.2)
Un flusso dei requisiti ben strutturato (capitolato → requisiti funzionali → microtask → test case) fa una cosa importante: traduce obiettivi astratti in unità operative tracciabili, in questo modo, ogni decisione è documentata, ogni è passaggio è motivato, ogni risultato è collegato a un requisito iniziale. Questo rende il processo “ispezionabile” e quindi analizzabile anche da altri.

Nel contesto dell’appalto

Nel contesto dell’appalto, L'Ente Appaltatore o committente ha la visione scientifica e definisce le necessità di alto livello, la Ditta vincitrice ha l’esecuzione tecnica e deve tradurre queste necessità "umanistiche" in punti esecutivi creando un flusso di documentazione scientifica e tecnica (Vd. Fig. 5)

Nel caso del nostro scenario, il Test Plan diventa lo strumento di garanzia: assicura che il Marketplace non sia solo un catalogo, ma un ecosistema dinamico dove ogni servizio rispetti gli standard di interoperabilità come richiesto dal committente in fase contrattuale.



Figura 5. Esempio di scomposizione del capitolato

In particolare, la documentazione dei test deve mantenere coerenza e tracciabilità rispetto a:

3. i documenti contrattuali, come il Capitolato Tecnico e l'Offerta Economica
4. i documenti di sistema, che descrivono l'architettura logica e fisica, e le interfacce dei vari moduli software;
5. i documenti di design, comprendenti le specifiche di dettaglio, i diagrammi UML, le definizioni di database e le scelte implementative;

L'integrazione tra tutti i documenti consente, inoltre, di sviluppare un "sistema documentale" standardizzato che può essere ulteriormente potenziato mediante l'adozione di repository e strumenti di orchestrazione automatica e versioning, per un allineamento continuo tra la documentazione dei test, le procedure operative e le specifiche funzionali.

Oltre il Marketplace

Si può pensare di adottare questo modello Test Plan ad altri progetti di Digital Humanities. Ad esempio, nell'ambito della linguistica computazionale, la definizione degli obiettivi, la costruzione del corpus, le annotazioni, la modellazione sono fasi fondamentali. Per ciascuna fase è necessario documentare: criteri di selezione dei dati (ad esempio fonti testuali o bilanciamento del corpus), operazioni di pulizia e normalizzazione, linee guida di annotazione e metriche adottate. In questo contesto il Test Plan svolge una funzione di formalizzazione e controllo poiché definisce:

- cosa deve essere verificato (es. consistenza delle etichette),
- come verificarlo (metriche, dataset di test, protocolli di validazione),
- quando e in quale fase del progetto eseguire i test.

Inoltre, è possibile includere test sulla qualità dei dati come ad esempio assenza di duplicati, correttezza della codifica, copertura linguistica. In questo modo il Test Plan chiarisce come deve essere verificata la qualità dei dati introducendo criteri formali di verifica e validazione.

4. CONCLUSIONI

Sappiamo bene che la robustezza di una ricerca dipende dalla solidità del processo che l'ha generata. In quest'ottica, il Test Plan non è un mero adempimento tecnico ma un pilastro della nostra metodologia di lavoro. Infatti, l'adozione sistematica di questo protocollo ci permette di rendere il processo di ricerca trasparente e resistente alle critiche accelerando i tempi di validazione scientifica. In sintesi, il Test Plan è molto più di una noiosa lista di controllo tecnica: è un "paracadute", un approccio metodologico per un progetto di ricerca.

BIBLIOGRAFIA

Cristofaro, S., Fabiani, V., Marras, C., E. Pasini, Sichera P., M. Yu. (2024). "Infrastrutture di ricerca come strumenti di "interculturalità. Digitale". In: Me.Te. Digitali. Mediterraneo in rete tra testi e contesti, Proceedings XIII Convegno Annuale AIUCD, Catania 28- 30 maggio 2024, Università di Catania, a cura di Di Silvestro A., Spampinato, D., Quaderni di Umanistica Digitale, pp. 572-576. DOI 10.6092/unibo/amsacta/7927

Degl'Innocenti, E., Pinna, F., Spadi, A., Spinelli, F. "Supporting Executable Scientific Workflows in a Clustered Infrastructure: DARIAH-IT and H2IOSC

Liburdi, A, Marras, C., Russo, A. «Infrastrutture, terminologie e policy per la ricerca umanistica: note per un confronto interdisciplinare». In Conferenza GARR 2018 - Data (R)evolution - Selected Papers. Cagliari, 2018. https://www.garr.it/it/chi-siamo/documenti/selected-papers/selected-papers-conferenza-2018/4711-selected-papers-conferenza2018-15-liburdi/file

IEEE Std 829-2008 — IEEE Standard for Software and System Test Documentation.

ISO/IEC/IEEE 29119-3:2021 Software and Systems-Engineering Software-Testing Test Documentation

Sichera, P., Marras, C., & Pasini, E. (2024, November). Orchestrazione api per workflow applicativi nell'ambito delle digital humanities. https://doi.org/10.5281/zenodo.14187534

Sichera, P., Marras, C., & Pasini, E. - Un endpoint OAI-PMH condiviso per l'orchestrazione dei servizi delle infrastrutture di ricerca nel progetto H2IOS. https://zenodo.org/records/19541925

Sichera, P., Cristofaro, S., Mazzagufo, L., SPAMPINATO, D., & DEL GROSSO, A. M. (2024, October). CHROMA model for H2IOSC [Presented at the Bazaar of the CLARIN Annual Conference 2024]. https://doi.org/10.5281/zenodo.13913607

Sichera, P., Fabiani, V., Marras, C., E. Pasini, Ongaro, P., Rossi, M. "Modelli e tecnologie integrate e innovative per una cittadinanza digitale equa e sostenibile", In: Diversità, Equità e Inclusione: Sfide e Opportunità per l'Informatica Umanistica nell'Era dell'Intelligenza Artificiale, Proceedings del XIV Convegno Annuale AIUCD2025. Verona, p. 663. ISBN 978-88-942535-9-7. DOI 10.6092/unibo/amsacta/8380.

Partecipazione, mediazione e conoscenza: Àndalas de Cultura quale ecosistema digitale del patrimonio culturale della Sardegna

CAROLINA FLORIS
UNIVERSITÀ DEGLI STUDI DI CAGLIARI, ITALIA – carolina.floris@unica.it

ABSTRACT (ITALIANO)

Il contributo prende in esame il progetto *Àndalas de Cultura* per riflettere sul ruolo che le infrastrutture digitali pubbliche possono assumere nei processi di mediazione del patrimonio e nella costruzione condivisa della conoscenza. Inserito nel più ampio contesto di rinnovamento delle politiche digitali a livello nazionale ed europeo, il caso viene letto come parte di un progressivo superamento di approcci emergenziali e tecnocentrici. Àndalas è assunto come un laboratorio osservativo attraverso cui indagare come pratiche di aggregazione, interoperabilità e riuso dei contenuti digitali, insieme a metodologie partecipative — dalla co-creation alla formazione degli operatori — incidano concretamente sulle modalità di accesso al patrimonio, sulla produzione di sapere e sulle relazioni tra istituzioni culturali e comunità in un contesto insulare policentrico. In questa prospettiva, l'infrastruttura si configura come uno spazio in cui il patrimonio prende forma attraverso pratiche d'uso, interpretazione e partecipazione.
Parole chiave: infrastrutture digitali; mediazione culturale; patrimonio culturale; SardegnaCultura; Àndalas de Cultura

ABSTRACT (ENGLISH)

Participation, mediation and knowledge: Àndalas de Cultura as a digital ecosystem for Sardinia's cultural heritage.
The contribution examines the Àndalas de Cultura project as a case study to reflect on the role that public digital infrastructures can play in heritage mediation processes and in the shared construction of knowledge. Situated within the broader national and European context of digital policy renewal, the case is interpreted as part of a gradual shift beyond emergency-driven and technocentric approaches.
Àndalas is approached as an observational laboratory through which to investigate how practices of aggregation, interoperability and reuse of digital content, together with participatory methodologies — from co-creation to capacity building — concretely shape access to heritage, knowledge production, and the relationships between cultural institutions and communities in a polycentric, insular context.
From this perspective, the infrastructure emerges not simply as a system for accessing content, but as a space in which heritage takes shape through practices of use, interpretation and participation.
Keywords: digital infrastructures; cultural mediation; cultural heritage; Sardegna Cultura; Àndalas de Cultura

1. PREMESSA: RIPENSARE IL DIGITALE NEL PATRIMONIO CULTURALE

Prima di addentrarsi nell'analisi del progetto Àndalas de Cultura, si ritiene necessario introdurre il contesto socio-tecnologico in cui si inserisce. Cominciata poco più di due decenni fa[1], la 'rivoluzione digitale' ha indubbiamente trasformato in maniera significativa ogni ambito della nostra quotidianità (Gruber & Glahn, 2015). In ambito culturale, essa esercita senza dubbio un impatto crescente sul modo in cui ci approcciamo alla conoscenza, alla comunicazione pubblica e alle strategie di gestione, valorizzazione e fruibilità del patrimonio culturale. In modo particolare, durante il periodo di crisi pandemica causata dal SARS-CoV-2, il sistema culturale ha sperimentato un'accelerazione significativa nell'adozione di strumenti digitali, rafforzando ulteriormente un processo di transizione tecnologica già in atto (Beccherle e Lazzaretti, 2023). Nonostante ciò, una gran parte dei contenuti prodotti e diffusi durante questa fase di emergenza ha evidenziato le limitazioni degli approcci adottati, spesso

[1] L'avvio convenzionale della cosiddetta 'rivoluzione digitale' viene generalmente collocato nel 1994, anno in cui il World Wide Web diventa accessibile al grande pubblico e inizia la sua diffusione su larga scala. (Bonacini, 2013).

caratterizzati da soluzioni tecnicamente datate e da una ridotta considerazione delle modalità contemporanee di narrazione e di interazione con il pubblico (Longo et al., 2020). È in questo contesto che il digitale applicato ai beni culturali è stato tradizionalmente inteso come risposta funzionale a bisogni pressanti, quali la continuità della fruizione, la visibilità online e la 'presenza' durante il periodo emergenziale (Mergel, Edelmann e Haug, 2019).
Questa tensione si riflette anche nella frammentazione che caratterizza il cosiddetto 'digitale culturale', articolato in una molteplicità di ambiti (Lughi, 2021). Questa condizione favorisce sovrapposizioni concettuali e interpretative, rendendo meno immediata la comprensione dei processi in atto e delle loro implicazioni complessive sul piano culturale. Il termine "digitalizzazione" è spesso utilizzato in modo ambiguo per riferirsi a processi diversi, con una sovrapposizione concettuale che privilegia strumenti e output rispetto ai processi e alle dinamiche sottostanti. In letteratura, infatti, digitization, digitalization e digital transformation vengono frequentemente usati in modo intercambiabile (Mergel, Edelmann e Haug, 2019; Valtysson, Kjellman e Audunson, 2022). Più precisamente, la digitization indica la conversione tecnica dei contenuti in formato digitale, la digitalization riguarda la loro integrazione nei contesti socio-istituzionali (Valtysson, Kjellman e Audunson, 2022; Uzelac e Lovrinić Higgins, 2025), mentre la digital transformation implica un cambiamento sistemico più profondo, risultato dell'interazione tra più innovazioni digitali che trasformano strutture organizzative, pratiche, ruoli e valori, ridefinendo le regole del gioco all'interno delle organizzazioni e degli ecosistemi (Agostino e Costantini, 2022). Nonostante la trasformazione digitale sia considerata fondamentale per rispondere alle esigenze del settore (Sandriester et al., 2025), la letteratura si è spesso concentrata sull'innovazione tecnologica e su cambiamenti incrementali, trascurando le pratiche e le dinamiche organizzative. Al contrario, essa richiede una prospettiva sistemica e olistica, che coinvolga sia le tecnologie, ma anche e soprattutto le persone e i processi (Koukopoulos e Koukopoulos, 2018).
Il contesto finora delineato si confronta con un modello istituzionale che, negli ultimi anni, ha progressivamente riconosciuto il ruolo strategico del digitale per il patrimonio culturale, promuovendo programmi e quadri di indirizzo su scala europea e nazionale (Beccherle e Lazzaretti, 2023; Capurro, Plets e Verheul, 2024). In ambito europeo, tale evoluzione si traduce nello sviluppo di infrastrutture e iniziative come Europeana, che segnano il passaggio verso un ecosistema basato su apertura, interoperabilità e riuso dei dati culturali (Longo et al., 2020; Capurro, Plets e Verheul, 2024; Uzelac e Lovrinić Higgins, 2025). In questo quadro, l'interoperabilità non è più intesa come un aspetto meramente tecnico, ma come una necessità strutturale per le istituzioni culturali, funzionale alla loro rilevanza e alla cooperazione transnazionale (Capurro, Plets e Verheul, 2024), mentre cresce parallelamente l'attenzione verso pratiche partecipative e modelli condivisi di produzione e accesso alla conoscenza (Beccherle e Lazzaretti, 2023). In ambito nazionale, il Piano Nazionale di Digitalizzazione (PND) e l'architettura PNRR hanno segnato un cambiamento significativo nel contesto italiano, introducendo una nuova enfasi sulla costruzione di un ecosistema digitale e sulla standardizzazione (Cerullo et al., 2024). In continuità con il modello europeo rappresentato da Europeana (infrastruttura di riferimento per l'aggregazione e il riuso dei contenuti culturali digitali) tali indirizzi trovano una concreta applicazione nello sviluppo di Ecomic, concepito come un ecosistema digitale integrato, interoperabile e federato, orientato alla connessione tra dati, attori e servizi e alla generazione di valore pubblico (ECOMIC, 2025).
Tuttavia, in ambito istituzionale persiste una tendenza a interpretare il digitale in chiave prevalentemente funzionale, come strumento di comunicazione, conservazione e fruizione, piuttosto che come processo trasformativo. In questa prospettiva, emerge una criticità legata al prevalere di approcci tecnocentrici, che privilegiano strumenti e soluzioni operative rispetto alla definizione di processi progettuali strutturati e allo sviluppo delle competenze necessarie per una trasformazione realmente integrata (Capodaglio, 2025; ICOM Italia, 2022[2]).

[2] In particolare, ICOM Italia nelle Osservazioni al Piano Nazionale di Digitalizzazione del Patrimonio Culturale (2022) ha rilevato come il Piano Nazionale di Digitalizzazione del Patrimonio Culturale attribuisca maggiore centralità alle soluzioni tecniche rispetto alla definizione dei processi progettuali, evidenziando la necessità di linee guida più chiare in relazione all'ideazione, alla sostenibilità, alle competenze e alla governance dei progetti digitali.

2. SARDEGNACULTURA: UN ECOSISTEMA PER L'ACCESSO DEMOCRATICO AL PATRIMONIO

All'interno di questa traiettoria e della necessità di osservare come le cornici programmatiche si concretizzino in pratiche operative, il contributo assume Àndalas de Cultura come caso di studio, avanzando la proposta di considerare il digitale non solo in qualità di strumento di promozione e gestione, ma come un'infrastruttura culturale e un dispositivo di mediazione (Smaniotto Costa, Artopoulos e Djukic, 2018). In tale contesto, il coinvolgimento attivo delle comunità diventa una componente intrinseca ai processi di innovazione digitale nelle Digital Humanities e nella museologia contemporanea, soprattutto quando la produzione e la circolazione di contenuti digitali implicano questioni di equità, cittadinanza culturale e costruzione condivisa di valore (Bonacini, 2018).
Il progetto Àndalas de Cultura, concepito nell'ambito dell'Agenda Digitale regionale nel marzo 2022 e realizzato attraverso fondi PO FESR 2014-2020 su iniziativa dell'Assessorato della Pubblica Istruzione e dei Beni Culturali della Regione Autonoma della Sardegna, nasce dalla necessità ormai matura di colmare il divario esistente nelle pratiche di comunicazione e di valorizzazione digitale del patrimonio culturale dell'isola, promuovendo una maggiore integrazione tra le risorse disponibili, le istituzioni e i diversi pubblici. In tale prospettiva, il progetto ha istituito, attraverso il portale SardegnaCultura, un punto di accesso unitario al patrimonio culturale regionale, sia materiale che immateriale, mediante l'impiego di tecnologie orientate a promuovere forme di fruizione, condivisione e partecipazione attiva da parte degli operatori del settore e delle comunità locali.
A seguito della pubblicazione del nuovo portale, avvenuta nell'ottobre 2023, si è cercato di integrare la molteplicità di risorse digitali generate nel tempo dall'amministrazione regionale, al fine di sviluppare un modello di fruizione che risultasse organico, sistemico, aperto e interoperabile. In tale cornice interpretativa, il progetto è stato concepito come uno strumento per rafforzare la conoscenza del patrimonio culturale e identitario: un sistema capace di generare effettive ricadute culturali, educative ed economiche sul territorio. Le tecnologie digitali sono concepite come infrastrutture al servizio della condivisione e della diffusione della conoscenza, in un'ottica di collaborazione tra le istituzioni coinvolte, gli operatori e le comunità (Velhinho e Almeida, 2023).
Il portale SardegnaCultura rappresenta il principale punto di accesso a questa infrastruttura di servizi digitali, con soluzioni che favoriscono il riuso delle informazioni e la creazione di nuovi percorsi di lettura e interpretazione del patrimonio. All'interno di questa visione, il progetto ha inteso attribuire un ruolo centrale alle istituzioni del territorio, promuovendo la costruzione di una rete in grado di ridurre l'isolamento e la frammentarietà informativa. La soluzione tecnologica adottata mette infatti a disposizione di tutti i soggetti culturali un ambiente comune per la raccolta, la conservazione digitale, la descrizione e la fruizione delle risorse.
Parallelamente, il progetto ha previsto azioni sistematiche di comunicazione, formazione e diffusione, finalizzate ad ampliare e diversificare i pubblici potenziali unitamente allo sviluppo di un'app dedicata alla fruizione di una selezione di luoghi della cultura, di servizi digitali rivolti a studenti ricercatori, famiglie e turisti, e di strumenti interattivi per la scoperta del territorio, rispondendo all'esigenza contemporanea di promuovere modalità di fruizione attive, partecipative e multilivello. Il caso sardo è assunto come laboratorio osservativo per indagare in che modo tali pratiche incidano sulla costruzione del sapere, sulle modalità di accesso al patrimonio e sulle relazioni tra istituzioni culturali e comunità in un contesto territoriale insulare e, ancora una volta, frammentato.

3. ÀNDALAS DE CULTURA: INFRASTRUTTURA E ARCHITETTURA DIGITALE

Dopo aver approfondito il contesto teorico e istituzionale, è ora possibile analizzare in modo più esaustivo le modalità operative attraverso cui il progetto ha tradotto tali principi in pratiche concrete.[3] Per comprendere a fondo l'approccio della Ragione Sardegna, è possibile fare riferimento alle teorie di Tim Berners-Lee, co-founder della World Wide Web Foundation, riguardanti il *Contract for the Web*[4]. Esso

[3] Tutti i dati quantitativi, descrittivi e progettuali relativi alla piattaforma Àndalas de Cultura presentati nel presente contributo derivano da documentazione interna non pubblicata, fornita dall'Amministrazione regionale nell'ambito della collaborazione scientifica con l'autrice.
[4] Il *Contract for the Web* è un'iniziativa globale promossa da Berners-Lee volta a promuovere una governance responsabile del web attraverso principi e azioni condivise tra istituzioni, imprese e cittadini, con particolare attenzione ai diritti, alla trasparenza e all'impatto sociale delle tecnologie digitali.

propone un modello di costruzione dei servizi digitali orientato a garantire un accesso ampio, inclusivo e partecipato alle informazioni. (Berners Lee, 2019).
In questa prospettiva, la dimensione infrastrutturale dimostra di superare l'approccio tecnocentrico, costituendo il dispositivo attraverso cui si rende possibile la mediazione tra istituzioni, territori e comunità (Velhinho e Almeida, 2023). Come evidenziato da De Pasquale nel Report di inquadramento concettuale di ECOMIC, gli ecosistemi digitali culturali non emergono spontaneamente, ma richiedono una costruzione intenzionale e una chiara visione della trasformazione digitale: in assenza di questa consapevolezza, il rischio è quello di ridurre il digitale a risposta a bisogni immediati, perdendone il potenziale come ambiente abilitante per nuove forme di conoscenza e partecipazione (ECOMIC, 2025). Nella stessa sede, Moro afferma invece che il passaggio cruciale è il cambio di prospettiva che considera il patrimonio da risorsa statica a piattaforma generativa, capace di attivare processi di co-creazione e re-mediazione culturale (ECOMIC, 2025).
L'infrastruttura di Àndalas si articola in tre livelli, rispondenti alle esigenze di conservazione, integrazione e accesso, i quali operano in modo interconnesso tra loro: rispettivamente, si concretizzano nel sistema di Digital Asset Management (DAM), nel metamotore di ricerca e, infine, nel portale SardegnaCultura (Figura 1). Mentre il sistema di gestione delle risorse digitali garantisce la stabilità e la continuità del patrimonio, il metamotore consente l'integrazione e la standardizzazione dei dati provenienti da archivi, biblioteche e musei. Tutto questo confluisce nel portale, che traduce tali dati in percorsi di consultazione e narrazioni digitali. Questi livelli risultano rilevanti sia per la loro specificità tecnica, ma sopratutto nella misura in cui abilitano interoperabilità, riuso e apertura verso ecosistemi più ampi, anche a scala nazionale ed europea, contribuendo alla costruzione di identità culturali condivise.

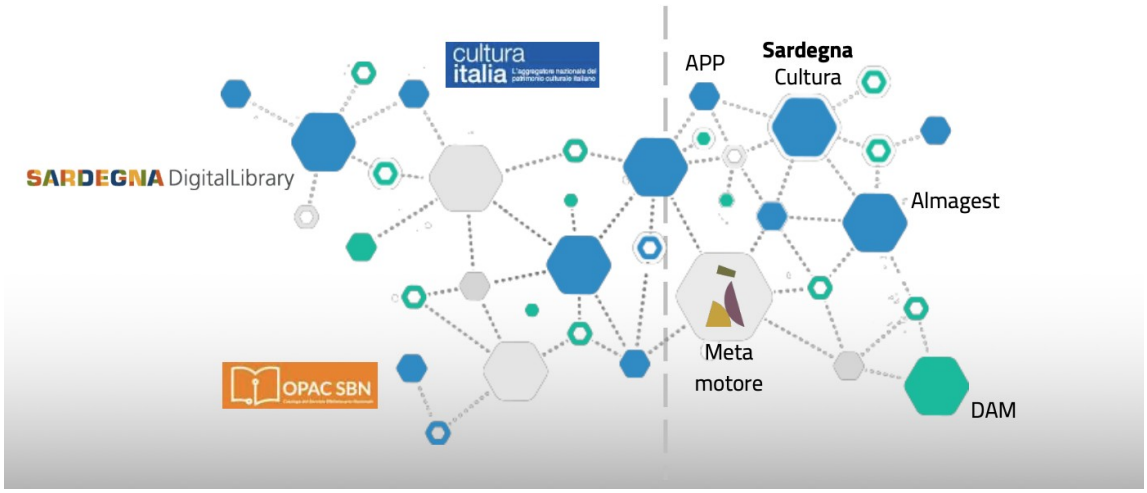


Figura 1: Ecosistema digitale Sardegna Cultura

Tali caratteristiche si collocano nel quadro delle politiche di open access e dei principi FAIR, orientati a garantire accessibilità, interoperabilità e la produzione di nuovo valore culturale (Longo et al., 2020). In linea con i principi della Convenzione di Faro, l'utente non è più solo destinatario, ma soggetto che contribuisce con diritto alla costruzione e circolazione del sapere culturale (Convenzione di Faro, 2005). Le tecnologie digitali applicate al patrimonio mirano infatti sempre più a coinvolgere il pubblico attraverso modalità esplorative, personalizzate e intertestuali, ponendo al centro i processi di co-creazione e mappatura culturale (Velhinho e Almeida, 2023).
Un ruolo centrale nella connessione tra l'infrastruttura e il territorio è svolto dall'applicazione mobile di Àndalas, sviluppata a partire da un lungo lavoro di raccolta e validazione dei dati sui luoghi della cultura. L'applicazione consente infatti l'accesso a oltre quattrocento siti culturali, considerando tra questi solo quelli effettivamente fruibili dal pubblico. Il sistema di ricerca consente di interagire con i dati tramite mappe interattive, filtri per località, tipologia e modalità di accesso, parole chiave e gallerie tematiche curate, consentendo all'utente di avere un'esperienza di fruizione personalizzata. In questo modo la logica della consultazione si sposta dalla semplice ricerca per categoria alla costruzione di una relazione immediata tra patrimonio e contesto, rendendo l'esperienza ancora più personale e cucita sulla persona. L'app, rispetto al portale SardegnaCultura, è pensata per un uso in movimento e, di conseguenza,

propone un sistema di esplorazione dei luoghi ancora più interattivo. La selezione dei luoghi e delle modalità di rappresentazione evidenzia tuttavia come le infrastrutture di questo tipo non siano neutre, ma incorporino scelte culturali e politiche: le infrastrutture digitali devono infatti essere comprese come elementi di policy, capaci di orientare pratiche, narrazioni e forme di accesso al patrimonio (Capurro, Plets e Verheul, 2024). Di conseguenza, anche istituzioni di piccole dimensioni e realtà periferiche trovano uno spazio stabile di visibilità, entrando in una rete condivisa che consente loro non solo di essere presenti, ma di essere riconosciute all'interno di una narrazione più ampia.
L'architettura dell' ecosistema digitale è quindi caratterizzata da una rete di valore, in cui le relazioni tra gli attori sono reciprocamente vantaggiose (Pesce e Franzè, 2025). È infatti soprattutto sul piano delle pratiche partecipative e delle relazioni costruite nel tempo che tali dinamiche diventano pienamente osservabili, come verrà discusso nel capitolo seguente.

4. ÀNDALAS DE CULTURA: PRATICHE PARTECIPATIVE E CO-COSTRUZIONE DEL SAPERE

A questa architettura si affiancano strumenti, procedure e pratiche operative, che trasformano l'infrastruttura in un ambiente orientato alla produzione, alla condivisione e alla rielaborazione dei dati. In particolare, il progetto ha articolato le proprie attività su due principali linee di intervento:

- quella formativa, rivolta all'accrescimento delle competenze da parte degli operatori culturali;
- quella partecipativa, rivolta alla co-costruzione del patrimonio culturale digitale da parte di musei, enti locali e istituzioni del territorio, e da parte dell'utenza pubblica attraverso la possibilità di creazione di percorsi narrativi (storie) creati a partire dai contenuti digitali resi disponibili dal portale SardegnaCultura.

Con riferimento al primo punto (attività formativa), il progetto ha infatti investito in percorsi di accompagnamento e formazione, finalizzati al rafforzamento delle competenze digitali e alla promozione di una partecipazione consapevole ai processi di produzione e gestione dei contenuti. Di particolare rilievo è l'iniziativa Àndalas Bootcamp, concepita come un ciclo di incontri su piattaforme online, con l'obbiettivo di approfondire le principali tematiche legate alla digitalizzazione, alla valorizzazione e alla comunicazione del patrimonio culturale. Il programma è stato definito attraverso una fase preliminare di consultazione rivolta a circa 1.438 invitati, consentendo di individuare priorità formative condivise sulla base delle effettive esigenze e del grado di competenza digitale del territorio. Il coinvolgimento ha previsto le seguenti fasi di attività:

- Nella prima fase, è stato definito il bacino di utenza mediante analisi, revisione ed aggiornamento dell'anagrafe delle istituzioni e luoghi della cultura in possesso dell'Amministrazione regionale, con particolare riferimento agli operatori di Biblioteche, Musei di ente locale ed Archivi comunali, quale target di riferimento del più ampio progetto Andalas de Cultura;
- Nella seconda fase, tramite strumenti di email marketing, è stata predisposta ed inviata a tutti i soggetti individuati una comunicazione rivolta ad illustrare e condividere le finalità del percorso formativo, chiedendo agli stessi di scegliere le tematiche di maggior interesse su circa 12 proposte nell'ambito dello sviluppo delle competenze digitali per i luoghi della cultura, mediante compilazione di un form on line. Ai destinati è stata inoltre riservata la possibilità di integrare la proposta con l'indicazione di ulteriori temi di interesse;
- Nella terza fase, i dati raccolti sono stati elaborati ed hanno consentito la definizione del percorso formativo con l'individuazione di esperti d'ambito per ciascuna tematica maggiormente votata dagli utenti.

Sulla base dei risultati emersi, nel 2023 sono state organizzate due edizioni del Bootcamp (aprile–giugno e luglio–agosto), i cui contenuti sono stati successivamente resi disponibili in modalità open. Il progetto ha promosso anche forme di restituzione narrativa del patrimonio, come ad esempio attraverso il format Àndalas a 10 voci. La ricerca ha previsto la realizzazione di una serie di brevi video-interviste condotte con rappresentanti della cultura sarda provenienti da ambiti professionali differenti (dalla musica alle tradizioni popolari, dagli archivi al settore museale), al fine di raccogliere in modo diretto le loro esperienze e pratiche professionali.

Con riferimento al secondo punto (attività partecipativa), il progetto ha fatto ricorso ai medesimi strumenti di email marketing e comunicazione per diffondere la conoscenza del progetto e le potenzialità degli strumenti resi disponibili al territorio. In questo modo l'Amministrazione ha inteso sensibilizzare le diverse realtà del territorio sulla necessità di una riduzione della frammentarietà del patrimonio culturale digitale esistente, fornendo un'ampia illustrazione delle tecnologie addottate ed delle possibili modalità di conferimento di risorse digitali esistenti. Al contempo, è stata assicurata la disponibilità a promuovere iniziative di co-progettazione di interventi di valorizzazione, che trovano in Àndalas de Cultura il contesto ideale per la conservazione e diffusione. A queste attività sono stati inoltre affiancati specifici interventi, che hanno incluso l'invito a partecipare a incontri e momenti di diffusione dei risultati del progetto, con l'obiettivo di stimolare l'interesse e la consapevolezza rispetto agli obiettivi condivisi, in una prospettiva di lungo periodo. Infine, ulteriori azioni di coinvolgimento del territorio hanno inteso rafforzare l'accesso agli strumenti e ai contenuti dell'iniziativa da parte dell'utenza pubblica dei luoghi della cultura mediante la dotazione, ad ogni biblioteca, archivio oltre che ad una selezione di musei del territorio regionale, di oltre 400 postazioni informatiche collocata presso ciascuna sede coinvolta.
In questa direzione, l'architettura della piattaforma integra dispositivi orientati alla partecipazione, come sistemi di commento e social tagging, che permettono agli utenti di intervenire direttamente sulle risorse: attribuire significati, proporre etichette, aggiungere livelli di interpretazione. Il patrimonio digitale si configura così come uno spazio aperto, in cui la produzione di senso non è più esclusivamente nelle mani delle istituzioni, ma si distribuisce tra una pluralità di soggetti. Per i soggetti accreditati quali, per esempio, le istituzioni scolastiche e culturali, sono disponibili funzionalità avanzate di annotazione basate su tecnologia IIIF, che consentono di associare contenuti testuali alle immagini e di condividerli in ambienti collaborativi. A queste pratiche si affianca l'utilizzo delle 'storie', attraverso cui gli utenti possono costruire percorsi narrativi personalizzati aggregando materiali selezionati dal metamotore.
È proprio in questo passaggio che emerge una delle tensioni centrali della digitalizzazione: oltre a rispondere a finalità conservative, essa deve necessariamente integrare dimensioni partecipative: la conservazione del patrimonio, infatti, risulta inefficace se non accompagnata dal coinvolgimento e dalla consapevolezza delle comunità (Beccherle e Lazzaretti, 2023). L'*engagement* si configura quindi come un processo strutturato di attivazione, fondato su pratiche di co-creation e, in parte, modelli riconducibili alla citizen science, che mirano a integrare gli attori del territorio nelle diverse fasi di progettazione, produzione e riuso dei contenuti digitali. Il coinvolgimento attivato dal progetto si configura come un processo articolato su più livelli, che supera una logica meramente orientata alla fruizione, spesso indirizzata esclusivamente a finalità turistiche, per estendersi a forme di partecipazione attiva nella produzione e interpretazione del patrimonio.

5. CONCLUSIONI: CRITICITÀ E PROSPETTIVE DI SVILUPPO

Il carattere territoriale e policentrico del progetto si traduce inevitabilmente in alcune criticità. È opportuno ricordare che Àndalas non si innesta su un contesto regionale privo di precedenti, ma si colloca all'interno di un percorso di lungo periodo che ha visto la Regione Sardegna investire in modo continuativo nella costruzione di un sistema informativo culturale, già previsto dalla legge regionale n. 14 del 2006. È infatti a partire da tale cornice normativa che, negli anni precedenti alla nascita del progetto, si sono progressivamente consolidate alcune infrastrutture fondamentali, come il portale SardegnaCultura e una prima Digital Library regionale. Queste hanno permesso di raccogliere, descrivere e rendere accessibile una parte rilevante delle risorse digitali prodotte dall'Amministrazione e dagli istituti culturali del territorio. Tali esperienze, leggibili come fasi preparatorie, hanno contribuito alla formazione di un patrimonio informativo diffuso, ma ancora caratterizzato da una forte frammentazione e disomogeneità. È in questo contesto che Àndalas de Cultura si configura come risposta a un'esigenza ormai matura: superare la dispersione delle risorse digitali e accompagnare la transizione da un insieme eterogeneo di banche dati a un ecosistema digitale unitario e coerente. Tuttavia, una parte significativa dei contenuti attualmente confluiti nella piattaforma proviene proprio da queste infrastrutture preesistenti e risente, di conseguenza, di differenze nei modelli descrittivi, negli standard adottati e nei livelli di aggiornamento. A quanto detto finora si affianca una forte variabilità delle competenze digitali presenti sul territorio. Il sistema regionale coinvolge infatti istituzioni molto diverse per dimensione, dotazione tecnologica, risorse umane e capacità progettuale, con effetti diretti sulla qualità, sulla continuità e sui processi di produzione e gestione dei contenuti. Nonostante le azioni formative e di accompagnamento promosse dal progetto, permangono differenze significative nei livelli di autonomia operativa e nella capacità di integrazione

stabile all'interno dell'ecosistema. Un ulteriore elemento critico è stato identificato nella forte connessione tra lo sviluppo dell'infrastruttura e i cicli di finanziamento pubblico. Come in molti progetti di innovazione digitale in ambito culturale, la crescita di Àndalas è stata scandita dall'accesso a programmi di finanziamento regionali, nazionali ed europei, che hanno consentito importanti avanzamenti, ma che, allo stesso tempo, rendono più complessa la continuità nel medio-lungo periodo. Tra le linee di sviluppo attualmente in fase di esplorazione rientra anche la sperimentazione di applicazioni di intelligenza artificiale e machine learning nell'ambito del progetto *Àndalas AI*. Tale orientamento apre nuove possibilità di analisi, classificazione e valorizzazione automatizzata delle risorse, ma richiede al contempo un attento presidio metodologico, etico e scientifico.
Nel complesso, l'esperienza di Àndalas de Cultura mostra come lo sviluppo di un ecosistema digitale pubblico possa attivare reti, competenze e pratiche collaborative, superando una visione esclusivamente tecnica del digitale. L'iniziativa si configura come piattaforma abilitante per la produzione condivisa di conoscenza, il rafforzamento delle istituzioni locali e il coinvolgimento delle comunità. Allo stesso tempo, emerge la necessità di approfondire i criteri di narrazione, le logiche di selezione dei contenuti e le modalità di rappresentazione digitale del patrimonio. Risulta inoltre centrale avviare una valutazione sistematica dell'impatto culturale, sociale ed educativo, attraverso strumenti analitici condivisi. In questa prospettiva, appare auspicabile la definizione di un toolkit di valutazione per lo studio comparativo e critico di progetti analoghi. Tale prospettiva apre ulteriori spazi di ricerca sul ruolo delle infrastrutture digitali pubbliche nei processi di mediazione culturale e di costruzione partecipata del valore.

RINGRAZIAMENTI

Si desidera esprimere un sentito ringraziamento per il sostegno finanziario ricevuto nell'ambito del Programma Regionale FSE+ 2021–2027, che ha reso possibile l'attivazione della borsa di studio presso l'Università degli Studi di Cagliari e lo svolgimento della presente ricerca nell'ambito della cultura digitale e della valorizzazione del patrimonio culturale.
Un ringraziamento particolare è rivolto alla prof.ssa Rita Pamela Ladogana e prof.ssa Simona Campus, per il costante supporto scientifico e metodologico fornito nel corso del percorso di ricerca.
Si ringraziano inoltre la dott.ssa Maria Stefania Podda, Direttrice del Servizio Supporti Direzionali, Sistemi Informativi e Biblioteca Regionale, e il dott. Cristiano Melis, Coordinatore del Settore Sistemi Informativi, per la disponibilità, la collaborazione e per aver messo a disposizione materiali, dati e competenze fondamentali per la realizzazione di questo studio, favorendo al contempo un coinvolgimento diretto nelle attività del progetto Àndalas de Cultura.

BIBLIOGRAFIA

Agostino, D., & Costantini, C. (2022). A measurement framework for assessing the digital transformation of cultural institutions: the Italian case. Meditari Accountancy Research, 30(4), 1141-1168.

Beccherle, P., & Lazzeretti, L. (2023). The role of digital technologies for culture-driven local development in Europe: A policy review. Il capitale culturale. Studies on the Value of Cultural Heritage, (28), 25-58.

Berners-Lee, T. (2019). I invented the World Wide Web. Here's how we can fix it. The New York Times, 24.

Bonacini, E. (2013). La valorizzazione digitale del patrimonio culturale in Europa e in Italia. Forme di fruizione e di valorizzazione museale attraverso le nuove tecnologie ei social media. Una proposta di turismo wireless per Catania. Tesi di Dottorato, Università degli Studi di Catania.

Bonacini, E. (2018). Partecipazione e co-creazione di valore culturale. #iziTRAVELSicilia e i principi della Convenzione di Faro. Il capitale culturale. Studies on the Value of Cultural Heritage, (17), 227-273.

Capodaglio, E. (2025). Nuovi paradigmi digitali per le esperienze culturali complesse. Modelli applicativi per piani strategici digitali per i musei: dalla conservazione, fruizione e comunicazione del patrimonio, allo sviluppo sostenibile. Tesi di Dottorato, Università degli Studi di Macerata.

Capurro, C., Plets, G., & Verheul, J. (2024). Digital heritage infrastructures as cultural policy instruments: Europeana and the enactment of European citizenship. International Journal of Cultural Policy, 30(3), 304–324.

Cerullo, L., Bartoli, M., Coppola, L. A., Landino, C., Madonna, A. D., Negri, A., ... & Rossetti, V. (2024). Verso la creazione di un Ecosistema digitale nazionale per la Cultura. DigItalia, 19(2), 11-48.

Consiglio d'Europa (2005). Convenzione quadro del Consiglio d'Europa sul valore dell'eredità culturale per la società. Consiglio d'Europa, Faro (Portogallo), 27.

ECOMIC, Ecosistema digitale per la cultura: Report di inquadramento concettuale (2025). Istituto centrale per la digitalizzazione del patrimonio culturale – Digital Library.

Gruber, M., & Glahn, C. (2015). E-learning for arts and cultural heritage education in archives and museums. Applying Digital Media to Culture; Hasebrook, J., Muhr, G., Schrader, A., Eds.

ICOM Italia. (2022). Report dell'indagine pilota su Digitale e digitalizzazione nei musei italiani (2020-2021): un'analisi e prospettive future, Commissione Tematica Tecnologie Digitali per il Patrimonio Culturale.

Koukopoulos, Z., & Koukopoulos, D. (2018). Evaluating the usability and the personal and social acceptance of a participatory digital platform for cultural heritage. Heritage, 2(1), 1-26.

Longo, D., Boeri, A., Turillazzi, B., & Orlandi, S. (2020). Cultural heritage and interoperable open platforms: strategies for knowledge, accessibility, enhancement and networking. WIT Transactions on Ecology and the Environment, 241, 371-382.

Lughi, G. (2021). Di cosa parliamo quando parliamo di cultura digitale. Agenda Culturale, 6.

Mergel, I., Edelmann, N., & Haug, N. (2019). Defining digital transformation: Results from expert interviews. Government information quarterly, 36(4), 101385.

Pesce, D., & Franzè, C. (2025). When digital platforms meet tradition: Phygital innovation in the cultural heritage. Journal of Engineering and Technology Management, 77, 101896.

Regione Autonoma della Sardegna. (2023). Documentazione interna del progetto Àndalas de Cultura (Materiale non pubblicato).

Sandriester, J., Harfst, J., Kern, C., & Zuanni, C. (2025). Digital transformation in the cultural heritage sector and its impacts on sustainable regional development in peripheral regions. Sustainability, 17(15), 6661.

Smaniotto Costa, C., Artopoulos, G., & Djukic, A. (2018). Reframing digital practices in mediated public open spaces associated with cultural heritage. Revista de Comunicação e Linguagens-Cidades do Futuro= Journal of Language and Communication-Cities of the Future, (48), 143-162.

Uzelac, A., & Lovrinić Higgins, B. (2025). From Data to Impact: Assessing the Value of Cultural Heritage in the Digital Age. Heritage, 8(4), 117.

Valtysson, B., Kjellman, U., & Audunson, R. (2022). The impact of digitalization on LAMs. In Libraries, archives, and museums in transition (pp. 117-129). Routledge.

Velhinho, A., & Almeida, P. (2023). The legacy of collective memory in digital culture: digitisation, cultural mapping and co-creation. Comunicação e sociedade, (43), e023003.

From Violin Lessons to Linked Open Data. Reconstructing Tartini's *Scuola delle Nazioni* through Digital Prosopography and Digital Editions

Selina Galka[1], Marcella Tambuscio[2], Rolf Wissmann[3], Cristina Scuderi[4], Georg Vogeler[5]

[1]University of Graz, - selina.galka@uni-graz.at,
[2]University of Graz, Austria - marcella.tambuscio@uni-graz.at,
[3]Schola Cantorum Basiliensis, Switzerland- rolf.wissmann@fhnw.ch,
[4]Università Statale di Milano, Italia- cristina.scuderi@uni-graz.at,
[5]University of Graz, Austria - georg.vogeler@uni-graz.at,

ABSTRACT (ENGLISH)

Giuseppe Tartini (1692–1770) was celebrated in the eighteenth century not only as a violinist, but also as maestro delle nazioni, attracting students from across Europe to his teaching activity in Padua. While his compositions and theoretical writings have been extensively studied, the pedagogical practices, didactic materials, and transnational networks through which his musical knowledge circulated remain largely unexplored. This research project addresses this gap by conceptualising Tartini's Scuola delle Nazioni not as a fixed institution, but as a dynamic, historically situated network of persons, texts, musical works, and mediating agents.

The project develops a digital prosopographical dataset on pupils, patrons, printers, and other actors, modelled as individual scholarly assertions explicitly linked to sources, temporal scope, and interpretative context, and expressed in RDF following CIDOC CRM–compliant patterns and FAIR principles. Musical sources and works are collected and systematically linked to international reference systems - most notably RISM - extending existing data models where necessary. In parallel, digital scholarly "assertive" editions of Tartini's pedagogical writings and related materials are created using TEI and MEI. The resulting interconnected data infrastructure, published in the long-term repository GAMS, enables advanced analytical and visual approaches, including mobility flows, temporal contact networks, multilayered social graphs, and spatial analyses of musical transmission.

Keywords: Digital Prosopography; Linked Open Data; Digital Scholarly editing; Semantic Web; Network Analysis

ABSTRACT (ITALIANO)

Dalle lezioni di violino ai Linked Open Data. Ricostruire la Scuola delle Nazioni di Tartini attraverso la prosopografia digitale e le edizioni digitali.

Giuseppe Tartini (1692–1770) fu celebrato nel XVIII secolo non solo come violinista, ma anche come *maestro delle nazioni*, attirando presso la sua scuola padovana studenti provenienti da tutta Europa. Sebbene le sue composizioni e i suoi scritti teorici siano stati ampiamente studiati, le pratiche pedagogiche, i materiali didattici e le reti transnazionali attraverso cui si diffusero rimangono in gran parte inesplorati. Questo progetto di ricerca affronta questa lacuna concettualizzando la Scuola delle Nazioni di Tartini non come un'istituzione fissa, ma come una rete storica dinamica di persone testi, opere musicali e agenti mediatori.

Il progetto prevede lo sviluppo di un dataset prosopografico digitale sugli allievi, i mecenati, gli stampatori e altri attori coinvolti, modellato come singole affermazioni esplicitamente collegate alle fonti, al periodo temporale e al contesto interpretativo, ed espresso tramite RDF seguendo pattern conformi a CIDOC-CRM e principi FAIR. Le fonti musicali e le opere sono raccolte e sistematicamente collegate a sistemi di riferimento internazionali - in particolare RISM - estendendo i modelli di dati ove necessario. In parallelo, verranno create edizioni digitali degli scritti di Tartini e di materiali correlati utilizzando TEI e MEI. Il risultato, una infrastruttura di dati interconnessi, sarà pubblicata nel repository a lungo termine GAMS, consentendo ulteriori analisi e visualizzazioni come flussi di mobilità, reti temporali, grafi sociali multilivello e analisi spaziali della trasmissione musicale.

Parole chiave: Prosopografia Digitale; Linked Open Data; Edizione Scientifica Digitale; Semantic Web; Analisi di Rete

1. INTRODUCTION

The eighteenth-century violinist, composer, and music theorist Giuseppe Tartini (1692–1770) was celebrated by his contemporaries as a *maestro delle nazioni* ("master of the nations", De Lalande 1769), attracting students from across Europe to his teaching activity in Padua from 1728 onwards. While Tartini's compositional output and theoretical writings have been widely studied, the pedagogical practices, didactic materials, and transnational networks through which his musical knowledge was transmitted have not yet been investigated. In fact, it seems only possible through a comprehensive digital framework capable of integrating sources, data, and interpretations. A research project addresses this by conceptualising Tartini's *Scuola delle Nazioni* as a complex digital research object, combining prosopographical data, semantic web technologies, and digital scholarly editions within a long-term repository. Rather than treating Tartini's teaching as an institutionalised "school" in a narrow sense, the project models it as a dynamic, historically situated network of persons, texts, musical works, and mediating agents such as patrons and printers (Scuderi 2024,2025).

2. THE CONCEPT OF SCHOOL

At the conceptual level, the project engages critically with the notion of "school" itself. In the eighteenth century, the term could refer to a physical place of instruction, a method of teaching, a body of shared practices, or a lineage of transmission extending across generations. In Tartini's case, these meanings overlap but do not coincide neatly. Teaching took place in Padua and Venice, but also continued indirectly through correspondence, copied notebooks, printed treatises, and the professional activities of former pupils. By modelling the *Scuola delle Nazioni* as a network rather than an institution, the project seeks to capture this fluidity and to situate Tartini's pedagogical activity within broader European patterns of mobility, patronage, and cultural exchange.

3. DIGITAL PROSOPOGRAPHY AND DATABASE OF MUSICAL WORKS

At the core of the project lies a digitally born prosopographical dataset recording pupils, patrons, printers, and other actors connected to Tartini. Core information about the persons will be names, lifespan (birth, death), geographical provenance, places of study and activity, timespan for their study with Tartini and timespan for patronage, relationships, relationships (as emerging from correspondences and sources) between all those actors, textual testimonies (on and by these actors) and their references, and categories relating to their musical production (roles, instruments played, musical production as deduced from RISM and other sources), as well as any further reference in secondary literature.
Rather than aggregating this information into flattened biographical narratives, the project adopts a factoid-based data model (Bradley/Short 2005; Bradley/Pasin 2013; Bradley 2020), in which each discrete statement — such as a teaching relationship, an activity at a specific place, or a patronage tie — is represented as an individual scholarly assertion explicitly linked to its source, temporal scope, and interpretative context. This approach allows the dataset to capture uncertainty, contradiction, and plurality of evidence, making visible the interpretative decisions underlying prosopographical reconstruction and enabling competing or evolving scholarly readings to coexist. The resulting dataset is expressed in RDF and aligned with international authority files and reference systems, including GND, VIAF, RISM, and GeoNames, following CIDOC CRM–compliant modelling patterns and FAIR data principles. E.g. each factoid could be modelled as a Document Interpretation Act (Pasin/Bradley 2013), instantiated as a crm:E13_Attribute_Assignment which binds three slots: the researcher responsible for the assertion via P14_carried_out_by, the primary source as a crm:E73_Information_Object via P16_used_specific_object, and the asserted state of affairs — a crm:E2_Temporal_Entity such as crm:E67_Birth, crm:E69_Death, or crm:E7_Activity — via P141_assigned. Contradictions are therefore encoded as coexisting source-bound assertions referring to the same entities or events: in this way, the project integrates critical historical reasoning directly into the data structure, ensuring both semantic interoperability and methodological transparency. To model the biographical dimension of the persons, the project considers Bio CRM (Tuominen/Hyvönen/Leskinen 2018), a domain-specific extension of CIDOC CRM designed for prosopographical research, which represents a person's life as a sequence of spatio-temporal events (e.g. concerts, professional activities, and social relationships) in which participants take explicitly typed roles. Bio CRM serves as a common harmonizing model for prosopographical data originating from different sources, formats, and historical periods.
In addition to personal data, musical sources play a crucial role. For this reason, the project aims to

clearly reference all relevant musical sources and works and link them to individuals (persons). The ambitious goal of the project is to link all of these musical sources to RISM data records, to expand them or to record them anew. On the one hand, this serves the project by improving the findability and interoperability of the data and, on the other hand, it serves RISM by expanding and improving it with high-quality research data. The RISM data model allows only two levels for recording the content of a source. If there is a composite manuscript containing several musical works in notated music, metadata can be recorded at both the collection level and the single item level. This flat data model reaches its limits in some of the present cases (e.g. D-B, Mus.ms.autogr.theor. Mariottini, A., D-Dl, MB.2.15). In such cases, a more flexible, multi-layered data model is used. In addition, metadata on musical works in the sense of 'works' according to the FRBR model are also recorded as separate entities if there is a relationship or reference at the work level.

4. DIGITAL EDITIONS

Another major part of reconstructing the school is the creation of digital scholarly editions of Tartini's pedagogical writings and related didactic materials. E.g., also the "Regole per arrivare a saper ben suonare il violino" will be part of the digital edition. Five manuscript copies have been preserved to date which will be collated and compared. Additionally, a contemporary French translation known under the title 'Traité des agréments de la musique' is important in this context, which has been printed in several editions. The first edition is also being published as part of the digital edition. Such materials provide direct insight into Tartini's teaching methods, musical concepts, and aesthetic priorities, but they also pose significant editorial challenges due to their variability, informal character, and complex transmission histories. The project addresses these challenges through a digital editorial approach that combines philological rigour with semantic enrichment. Textual sources are transcribed and encoded in TEI, while musical examples are represented using MEI, allowing for a precise and flexible representation of notation, variants, and material features. We are establishing a workflow for creating the digital editions, using Transkribus for the transcription of the text and MuseScore for the transcription of the music.
The editions are conceived as *assertive digital editions* (Vogeler, 2019): rather than merely reproducing sources, they embed scholarly claims through semantic annotations that connect pedagogical concepts, musical terminology, persons, places, and works directly to the prosopographical dataset. A central outcome of this work is the development of a structured glossary of Tartinian technical terms and concepts, modelled as a SKOS-based semantic resource.

5. ANALYSIS AND VISUALIZATIONS

The integration of semantic data and digital editions supports advanced forms of analysis and visualization. The database is explicitly designed to connect persons, teaching relations, events, institutions and musical sources and allow us to turn structured historical data into interpretable patterns of music spreading, contacts and power relationships through data and network science methods. In particular, the network perspective (Ahnert, 2019) is not intended merely to map relational proximity, but also to model different processes of musical knowledge transmission.
A first line of inquiry will focus on the professional diaspora for the students after attending Scuola delle Nazioni. We propose (Sankey) diagrams that trace students from their place (or institution) of study to subsequent career destinations: this view will make visible dominant corridors or secondary trajectories, allowing us to compare mobility by instrument played, role, nationality or patronage involvement. Rather than treating mobility as a simple origin-destination statistics, in the final platform the users will have the possibility to aggregate those paths by different features and develop further analysis and research. A second workflow will focus on interpersonal contact within the school environment: we are already realizing a temporal contact network by linking individuals who plausibly overlapped in the same time interval in the School (see Figure 1). By usual network analysis methods, we can highlight strategic people or clusters in this network, even on temporal dimensions, focusing on temporal slices: this dynamic perspective supports questions as whether contact opportunities concentrate around specific periods, teachers or institutional settings and how temporal proximity correlates with later shared careers or collaborations.

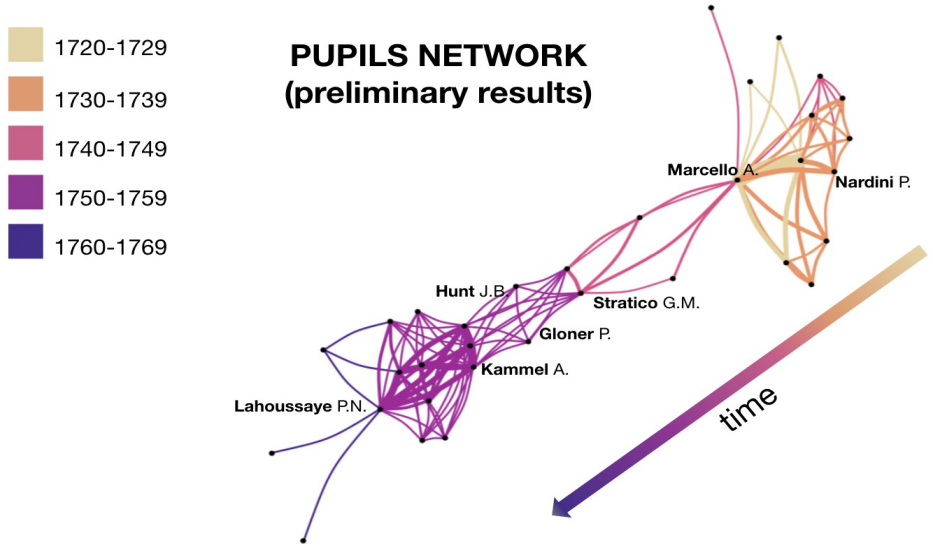


Figure 1. Social Network of Tartini's pupils (based on preliminary data). The figure represents a temporal contact network of selected pupils who attended Scuola delle Nazioni. Nodes correspond to individual actors, while edges indicate inferred contact relationships based on temporal overlap within the same pedagogical context. Colours encode different time periods, allowing the visualisation of diachronic patterns in the formation and transformation of the network. The visualisation illustrates how the structure of the network evolves over time, highlighting shifts in central actors and the emergence or fragmentation of groups across decades. As such, it should be interpreted not as a definitive reconstruction of social relations, but as an analytical model supporting the investigation of pedagogical transmission processes and the circulation of musical knowledge within Tartini's milieu.

To capture the complexity of social and professional ties, we will consider different layers representing different types of interaction: (1) teaching, (2) shared institutions and (3) co-presence in events. This layered structure allows us to test if some institutions act as bridges between otherwise disconnected teaching clusters or to compare structural roles in distinct layers (e.g. hub as teacher and broker among institutions). The multilayered representation also supports robustness: ties observed in multiple layers can be treated as stronger evidence of sustained interaction. At the micro-scale, we propose systematic ego-network reconstructions for selected individuals (students, teachers, patrons), enabling qualitative-quantitative comparison of personal trajectories. Ego-networks can be characterized through recurring patterns (dense clusters around specific teachers or patrons) and the comparison of their topologies (structures) provides a compact way to discuss similarities and differences in individual experiences and careers.

If the data collection will provide enough information about patronage we will have the opportunity to study these directed relations that encode access to resources, prestige and opportunities: do students with documented patronage links exhibit higher success or productivity? Centralities measures could reveal individuals or institutions acting as gatekeepers, as well in a diachronic perspective.

Finally, the database supports a detailed analysis on the transmission of music: we can map geography of testimonies by linking works and exemplars to holding institutions and places, reconstructing the spatial spreading traces of Tartini-related repertories. Statistics focused on prints and manuscripts will reveal shifts in production and preservation over time. Moreover, where metadata allows, we extend the network perspective to the production chain: a graph connecting authors, copyists, engravers and editors to identify clusters of production associated with specific regions or institutions.

In the final platform users will interact with networks and visualisations described in this section exploring relationships between teachers, pupils, patrons, and printers (Tambuscio/Vogeler, 2025); timelines to trace periods of study, mobility, and publication; and georeferenced maps to visualise the spatial reach of Tartini's *Scuola delle Nazioni* and the circulation of music through manuscript and print. In particular, the proposed visualisations are not considered as neutral displays of pre-existing data, but as interpretative interfaces focusing on different features of the database, grounded in specific modelling choices and assumptions, that will be explicitly claimed in the platform. By allowing users to move seamlessly between visualisations, source editions, and underlying data, the project fosters both scholarly exploration and critical engagement with the constructed nature of historical knowledge.

6. ARCHIVING AND INTERFACE

All data, digital editions, and visualisations are published within a long-term digital repository, which follows the OAIS reference model and adheres to FAIR principles of findability, accessibility, interoperability, and reusability (Stigler/Steiner 2018). All research outputs are stored under persistent identifiers and made available using open, standards-based formats, supported by rich descriptive and technical metadata (e.g. Dublin Core, METS, EDM), ensuring long-term preservation, citability, and interoperability with external research infrastructures. The data will additionally be made available in multiple export formats, e.g. including TEI, MEI, RDF, JSON, or CSV. The interface developed on top of the data - implemented using XSLT for the creation of a rich and dynamic HTML/JasvaScript frontend, and generated dynamically directly from the underlying data through predefined SPARQL queries or direct processing of TEI XML - brings together all project components within a single environment, integrating the digital prosopography, the digital scholarly editions, and related analytical views and visualisations. Challenges concern the coherent presentation of individual sources and their textual and musical variants within the digital editions, as well as the meaningful integration of text and music notation.

7. CONCLUSIONS

At the current stage, the projection is still in the phase of data collection, with ongoing efforts to expand the prosopographical dataset, integrate additional musical sources and refine the semantic model. The analytical workflows and visualisations presented here should therefore be considered as preliminary explorations: over the next two years, the project will focus on completing data acquisition, digital editions and further developing process-oriented analyses of musical transmission. These future developments will allow for a more comprehensive and robust assessment of Tartini's Scuola delle Nazioni as a complex system of knowledge production and circulation. In line with the themes of AIUCD 2026, the project reflects on digital infrastructures as sites of knowledge production and public engagement, discussing a methodological model for connecting prosopographical databases, digital editions and network analysis in the study of historical music transmission. By publishing data, editions, and visualizations in open-access form and by adhering to shared standards and ontologies, the project promotes reuse, sustainability, and transparency.

ACKNOWLEDGEMENT

This research was funded by the Austrian Science Fund (FWF) "Tartinians in Europe. The School of Nations and its Networks" DOI 10.55776/PIN5035023

REFERENCES

Ahnert, R., Ahnert, S. E., Coleman, C. N., & Weingart, S. B. (2020). The network turn: Changing perspectives in the humanities. Cambridge University Press. https://doi.org/10.1017/9781108866804

Bradley, J. (2020). Factoid prosopography. King's College London. https://www.kcl.ac.uk/factoid-prosopography

Short, H., & Bradley, J. (2005). Texts into Databases: The Evolving Field of New-style Prosopography. Literary and Linguistic Computing: The Journal of Digital Scholarship in the Humanities, 19(5), 3–24. https://doi.org/10.1093/llc/fqi022

De Lalande, Joseph Jérôme Le Français. Voyage d'un François en Italie, vol. 8, Paris: Desaint, 1769.

Pasin, M., & Bradley, J. (2015). Factoid-based prosopography and computer ontologies: towards an integrated approach. Digital Scholarship in the Humanities, 30(1), 86-97. https://doi.org/10.1093/llc/fqt037

Scuderi, C. (2024). Giuseppe Tartini (1692–1770): European Paths, Networks, Legacy. Eighteenth Century Music, 21, 273 - 276. https://doi:10.1017/S1478570624000101

Scuderi, C. (2025). Reframing Tartinian Networks Within 18th Century European Landscape: Issues and Perspectives from an International Project. Ex Chordis. https://doi.org/10.54103/3034-8781/30809

Stigler, J. H., & Steiner, E. (2018). GAMS – Eine Infrastruktur zur Langzeitarchivierung und Publikation geisteswissenschaftlicher Forschungsdaten. Mitteilungen Der Vereinigung Österreichischer Bibliothekarinnen Und Bibliothekare, 71(1), 207–216. https://doi.org/10.31263/voebm.v71i1.1992

Tambuscio, M., & Vogeler, G. (2025). Bridging Semantics and Structure: a Typed Prosopographical Network of Maximilian I's Court. Anthology of Computers and the Humanities, 3, 131-141.https://doi.org/10.63744/eWebTidsVFzV

Tuominen, J., Hyvönen, E., & Leskinen, P. (2018). Bio CRM: A data model for representing biographical data for prosopographical research. Proceedings of the Second Conference on Biographical Data in a Digital World 2017 (BD2017), CEUR Workshop Proceedings, 59–66.

Vogeler, G. (2019). The 'assertive edition': On the consequences of digital methods in scholarly editing for historians. International Journal of Digital Humanities, 1, 309–322. https://doi.org/10.1007/s42803-019-00025-5.

Addestramento di un modello di HTR in Transkribus per testi plurilingui di età moderna: il caso del *Ripulimento della lingua sarda* (Madau)

Michela Incollu[1], Giulia Murgia[2]

[1]Università degli Studi di Siena, Italy - m.incollu@student.unisi.it
[2]Università degli Studi di Cagliari, Italy - giulia.murgia@unica.it

ABSTRACT (ITALIANO)

Il contributo illustra le metodologie adottate per l'addestramento, tramite la piattaforma Transkribus, di un modello sperimentale di Handwritten Text Recognition (HTR) applicato a un manoscritto plurilingue di età moderna, il *Ripulimento della lingua sarda* di Matteo Madau (XVIII secolo). Viene riportato l'intero flusso di lavoro: preparazione dati e analisi del layout del documento; creazione della Ground Truth, con particolare attenzione alla metodologia impiegata per garantire la coerenza dei dati utili all'addestramento; criteri adottati per la selezione del modello base; avvio del processo di fine-tuning. I risultati del primo ciclo di addestramento sono presentati e valutati attraverso successivi test del modello su nuove pagine non viste durante l'addestramento. Le trascrizioni ottenute sono analizzate in relazione alla tipologia e alla frequenza degli errori di riconoscimento automatico. Tale analisi consente di misurare le effettive prestazioni del modello e di individuarne le principali criticità, in vista di futuri cicli di addestramento volti a migliorarne l'affidabilità.
Parole chiave: Transkribus; HTR; Ground Truth; fine-tuning; testo plurilingue.

ABSTRACT (ENGLISH)

Training HTR Text Model on Transkribus for Early Modern Multilingual Texts: the Case of Ripulimento della lingua sarda *(Madau)*
The present contribution illustrates the methodologies adopted for the training, on the Transkribus platform, of an experimental Handwritten Text Recognition (HTR) model, applied to a multilingual early modern handwritten text, *Ripulimento della lingua sarda* by Matteo Madau (18th century). The entire workflow is presented, including data preparation and document layout analysis, the creation of the Ground Truth, with particular attention to the methodology employed to ensure the coherence of the data used for training, the principles adopted for the selection of the base model and the initiation of the fine-tuning process. The results of the first training cycle are presented and evaluated through subsequent tests of the model on new pages not seen during training. The transcriptions obtained are analyzed in relation to the type and frequency of automatic recognition errors. This analysis allows for the evaluation of the actual performance of the model and to identify the major critical issues, useful for future training cycles aimed at improving its reliability.
Keywords: Transkribus; HTR; Ground Truth; fine-tuning; multilingual text.

1. INTRODUZIONE[1]

Il presente contributo si propone di esplorare le metodologie coinvolte nell'addestramento, condotto sulla piattaforma Transkribus, di un modello sperimentale di HTR (Handwritten Text Recognition) basato sulla tecnica del fine-tuning incrementale a partire da un modello base pre-addestrato. L'obiettivo principale è riflettere sull'applicazione di tali metodologie a un testo manoscritto plurilingue, il *Ripulimento della lingua sarda* di Matteo Madau (XVIII secolo), e analizzare le differenze di apprendimento del modello nelle due sezioni principali del manoscritto (saggio e dizionario), caratterizzate da diversi gradi di complessità, sia strutturale che linguistica.

[1]L'articolo è il frutto di una stretta collaborazione tra le autrici. Sono tuttavia da attribuire a Michela Incollu i parr.2,4,5,6 e a Giulia Murgia il par. 3. Il par. 1 è da attribuire ad entrambe le autrici.

Il progetto si inserisce in una prospettiva più ampia di valorizzazione del patrimonio linguistico e culturale sardo; trattandosi di un documento finora inedito, la sua digitalizzazione favorirebbe una fruizione più accessibile, anche da parte del pubblico non specializzato.

2. TRANSKRIBUS

Transkribus è una piattaforma di digitalizzazione e riconoscimento automatico del testo dotata di intelligenza artificiale, pensata per permettere anche a studiosi non esperti di coding e deep learning di accedere a strumenti avanzati di trascrizione automatica. Sviluppata nell'ambito del progetto READ dall'Università di Innbruck in occasione del programma di finanziamento e sviluppo europeo Horizon 2020 (*Recognition and Enrichment of Archival Documents* | *READ* | *Project* | *Fact Sheet* | *H2020*, s.d.), a partire da un iniziale progetto denominato tranScriptorium (Sánchez et al., 2013), è attualmente gestita dalla cooperativa READ-COOP. La piattaforma si serve del motore di riconoscimento PyLaia, un toolkit che integra reti neurali convoluzionali (CNN) per l'estrazione delle caratteristiche visive, ossia i tratti grafici, e reti neurali ricorrenti (RNN), per la modellazione sequenziale del contesto testuale (Puigcerver, 2017/2026). Quest'architettura consente un apprendimento più stabile e veloce rispetto a sistemi HTR precedenti, basati su reti neurali ricorrenti di tipo Multidimensional Long Short-Term Memory (MDLSTM), caratterizzati da una maggiore complessità computazionale (Puigcerver, 2017). Un ulteriore vantaggio è costituito dal fatto che la piattaforma supporta numerose lingue, antiche e moderne, e, permettendo la riutilizzabilità dei modelli pubblici in processi di fine-tuning incrementale, consente l'addestramento di nuovi modelli personalizzati che si adattino a documenti redatti in lingue minoritarie non ancora rappresentate.
Una criticità della piattaforma Transkribus, se valutata sul confronto con soluzioni e ambienti operativi analoghi, riguarda la dipendenza da un sistema di crediti a pagamento, necessari per l'analisi e il riconoscimento automatico di testo e layout; questo sistema rende la piattaforma meno adatta a gestire progetti su larga scala, in quanto la dipendenza economica tende ad aumentare progressivamente. Esistono alternative gratuite e open-source, come l'infrastruttura eScriptorium (Kiessling et al., 2019; Stokes et al., 2021), che permette, inoltre, l'esportazione e il riuso dei modelli HTR anche al di fuori dell'ambiente in cui sono stati prodotti. Transkribus, al contrario, permette il riuso di modelli pubblici solo all'interno del proprio ecosistema.
Tuttavia, l'integrazione "in-the-cloud" del motore PyLaia, offerta direttamente all'interno della piattaforma, rappresenta un notevole vantaggio di utilizzo; a differenza di eScriptorium, infatti, non è richiesta alcuna installazione locale, che richiederebbe all'utente competenze tecniche per la gestione dell'infrastruttura, né la disponibilità di risorse computazionali particolarmente avanzate.
In linea con le pratiche relative al public engagement, la piattaforma Transkribus riunisce una vasta comunità di studiosi e incoraggia la condivisione dei modelli HTR addestrati, in particolare per scritti e lingue sotto-rappresentate. In quest'ottica, l'addestramento di un modello HTR su un documento come il *Ripulimento della lingua sarda* e la sua successiva pubblicazione all'interno della sezione dedicata ai modelli condivisi costituirebbero un contributo importante, agevolando contestualmente future trascrizioni di manoscritti plurilingui.

3. DESCRIZIONE MANOSCRITTO

Il *Ripulimento della lingua sarda* di Matteo Madau è l'opera che, nel XVIII secolo, dà avvio agli studi di linguistica sarda (Murgia, 2023). L'autore diede alle stampe un primo *Saggio* dell'opera nel 1782 (Cagliari, Titard). Tuttavia, del *Ripulimento* esiste anche un manoscritto ad oggi inedito, scritto di pugno dallo stesso Madau, suddiviso in due tomi cartacei (rispettivamente di 196 e 293 carte), oggi appartenente al Fondo Baille della Biblioteca Universitaria di Cagliari, dove è conservato con segnatura S.P.6.1.39 e S.P.6.1.40. Rispetto al *Saggio* a stampa, il *Ripulimento* inedito manoscritto è l'edizione integrale che contiene in maniera compiuta la riflessione linguistica di Madau: i due tomi che lo compongono accolgono non solo il discorso introduttivo sulla lingua sarda, ma anche un ricco dizionario storico-etimologico bipartito in due sezioni, con ventimila voci sarde di presunta derivazione greca e ben centomila vocaboli derivati dal latino.
Le carte hanno una dimensione di 355 x 223 mm e sono fittamente scritte a piena pagina. Il numero di righe è mutevole e oscilla tra le 23 e le 51. Il testo appare redatto da un'unica mano.
Nel primo tomo, la sezione che precede il dizionario – contenente la dedica al re Vittorio Amedeo III, le indicazioni dell'autore a chi legge, l'introduzione, la presentazione dell'opera e le indicazioni ortografiche –

presenta un testo disteso su un'unica colonna, a piena pagina (cc. 1r-42r). Come si può notare dalla figura 1, lo spazio riservato ai margini destro e sinistro accoglie le note apposte da Madau, richiamate a testo mediante un esponente alfabetico. A rendere affollate le pagine contribuisce anche l'assenza di spazi tra un paragrafo e l'altro. Non mancano poi interventi dovuti a ripensamenti dell'autore, come cancellature e aggiunte in interlinea. Abbondante inoltre è l'uso della sottolineatura, impiegata per segnalare le citazioni in lingue diverse dall'italiano e le parole che, in una futura edizione a stampa mai realizzata, sarebbero state da rendere in corsivo.

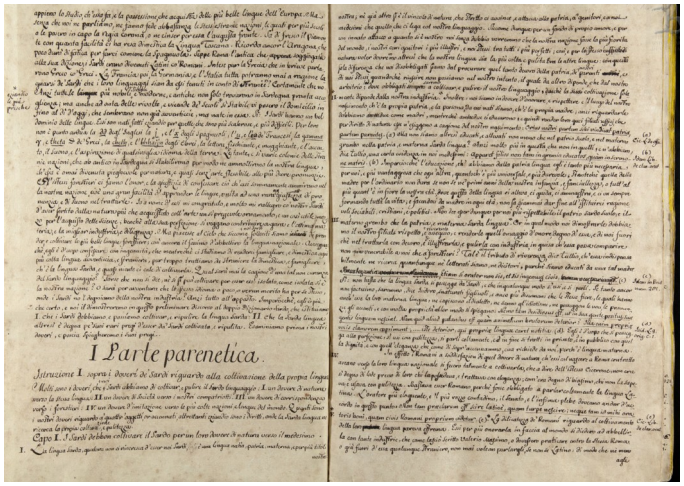


Figura 1. *Ripulimento della lingua sarda*.
Cagliari, Biblioteca Universitaria, Fondo Baille, S.P.6.1.39, cc. 10v-11r
Su concessione del Ministero della Cultura / Biblioteca Universitaria di Cagliari

A partire dalla c. 43r del primo tomo, inizia la sezione del *Dizionario di sardi voci, portate dal Greco*, nella quale i lemmi ordinati alfabeticamente si distendono su due colonne.



Figura 2. *Ripulimento della lingua sarda*.
Cagliari, Biblioteca Universitaria, Fondo Baille, S.P.6.1.39, cc. 45v-46r
Su concessione del Ministero della Cultura / Biblioteca Universitaria di Cagliari

Le figg. 1 e 2 rendono chiare le principali difficoltà che si pongono al riconoscimento automatico del nostro testo: la presenza di una lingua come il sardo, che non è contemplata tra i modelli proposti dalla piattaforma perché mai messa alla prova su Transkribus; la presenza di layout eterogenei (organizzazione dello specchio di scrittura su una o due colonne; presenza di glosse nei margini; uso della sottolineatura; ecc.); l'alternarsi di più alfabeti (latino e greco); il plurilinguismo dell'opera, scritta in italiano, ma con una cospicua presenza di sardo, latino e greco.

4. PREPARAZIONE DEI DATI[2]

La preparazione dei dati costituisce una fase fondamentale nel processo di addestramento di un nuovo modello di riconoscimento automatico del testo.
Per prima cosa vengono caricati, all'interno di una nuova collezione, i documenti da analizzare, in formato jpg e alla risoluzione di 300 dpi; la digitalizzazione del manoscritto di Madau è stata fornita dalla Biblioteca Universitaria di Cagliari, che, attraverso il servizio DIGIT, fornisce su richiesta le copie digitali per uso di studio o di ricerca.
Una volta completato il caricamento, occorre eseguire l'analisi e il riconoscimento del layout; questa fase preliminare, infatti, consente al software di individuare le aree testuali (sezioni e righe di testo) e di analizzarne i tratti grafici, così da apprendere le corrispondenze tra immagine e trascrizione.
Sulla piattaforma è possibile automatizzare l'analisi del layout, selezionando le pagine interessate e scegliendo, tra i modelli disponibili, quello che più si adatta alle caratteristiche del documento in esame. Nel nostro caso sono stati inizialmente presi in considerazione due modelli: *Mixed Line Orientation* e *Universal Lines*. La scelta è stata condizionata dalla struttura eterogenea del documento, caratterizzato dalla compresenza di una prima sezione con layout abbastanza regolare (saggio) e di una seconda, più cospicua, organizzata in testo a due colonne (dizionario), oltre alla presenza di numerose note a margine, esterne alla cornice di testo principale.
Sulla base di tali caratteristiche, il modello *Mixed Line* è apparso inizialmente l'opzione migliore, in quanto addestrato a gestire layout eterogenei; tuttavia, si è preferito optare per il modello *Universal Lines*, da applicare all'intero corpus in modo da ridurre il rischio di over-segmentation e ottenere una base più stabile su cui condurre le analisi successive.
Entrambi i modelli hanno mostrato da subito significative difficoltà a individuare correttamente il layout nella sezione con testo a due colonne, compromettendo così la segmentazione e alterando l'ordine naturale delle righe di testo. Per ovviare a queste difficoltà è stato avviato l'addestramento di un modello *Baseline* specifico per il layout del documento. Il nuovo modello (*Baseline Model Madau - experimental*), addestrato su un campione di pagine preparate tramite la selezione manuale delle aree testuali, ha prodotto risultati soddisfacenti solo se applicato a singole pagine, mentre si mostra ancora inefficace nella visione a doppia pagina che caratterizza le riproduzioni digitali del manoscritto. Necessita, pertanto, di ulteriori cicli di addestramento, in modo da poter raggiungere la soglia di *loss* inferiore al 10%; allo stato attuale, infatti, registra un valore pari a 14,84%.
Un'ulteriore fase centrale è rappresentata dalla preparazione della Ground Truth, ossia la raccolta di dati, verificati accuratamente, utilizzata per condurre l'addestramento, la convalida e infine il test dei modelli di riconoscimento automatico del testo. Secondo le linee guida Transkribus, reperibili all'interno della sezione *Training Text Models*, alla voce *Data Preparation*, per una Ground Truth consistente è necessario trascrivere manualmente un numero di pagine che varia a seconda delle caratteristiche del testo in esame. Nel caso di testi manoscritti, in particolare, si consiglia di raccogliere almeno 10.000 parole per ogni mano così da condurre un addestramento che possa raggiungere potenzialmente un *Character Error Rate* (CER) compreso tra il 2% e il 4%.
Un altro aspetto importante nella costruzione della Ground Truth riguarda l'accuratezza della trascrizione e la coerenza delle scelte editoriali adottate; si tratta di elementi indispensabili per un apprendimento più veloce ed efficiente del modello, in quanto, come evidenziato da studi recenti, la qualità della Ground Truth ha un impatto significativo sull'addestramento, e di conseguenza sulle prestazioni, dei modelli HTR. Significativamente, è stato dimostrato che una Ground Truth di dimensioni ridotte ma accurata e coerente contribuisca a minimizzare l'insorgenza di errori sistematici, che tendono invece a manifestarsi con maggiore frequenza laddove i modelli siano stati addestrati su una gran mole di dati approssimativi (Jungo et al., 2023).

[2] La metodologia impiegata per l'avvio e lo sviluppo del lavoro su Transkribus si fonda principalmente sulle linee guida ufficiali disponibili all'interno della piattaforma stessa, in particolare alla sezione contenente i materiali relativi all'addestramento dei modelli di riconoscimento automatico del testo (Training Text Recognition Models, s.d.), e alle pagine dedicate alla preparazione dei dati (Data Preparation, s.d.) e alla descrizione dei modelli (Model Setup and Training, s.d.).

Per ragioni di tempo si è scelto di concentrare il lavoro di trascrizione su un sottoinsieme rappresentativo di 20 pagine(22.576 parole, 2.444 righe),[3] tratte dal primo tomo del manoscritto di Madau. Le pagine sono state scelte in maniera casuale, in modo da raccogliere una quantità quanto più ampia di dati testuali, e ripartite in egual numero tra le due sezioni principali (saggio e dizionario). Le trascrizioni sono state quindi salvate e marcate come Ground Truth.

5. ADDESTRAMENTO

Completata la preparazione dei dati di Ground Truth, è stato possibile procedere con l'addestramento di un nuovo modello di riconoscimento automatico. Procedendo all'interno della collezione in esame, vengono selezionate le pagine precedentemente trascritte e marcate come Ground Truth, scegliendo dal menù a tendina la voce *New Train Model*. Le pagine vengono quindi suddivise automaticamente in due gruppi: training set, su cui avverrà l'addestramento vero e proprio, e validation set, corrispondente, secondo lo standard espresso dalle direttive di Transkribus, al 10% dei dati totali, utilizzato per monitorare le prestazioni del modello durante l'addestramento.
Successivamente vengono richiesti i metadati descrittivi relativi al nuovo modello: nome (Madau multilingual model [experimental]), breve descrizione (Experimental model for 18th century plurilingual handwritten text trained on Matteo Madau's work, "Ripulimento della lingua sarda"), periodo cronologico (XVIII secolo) e lingue presenti nel documento (Italian; Sardinian; Latin; Greek, ancient). All'interno della sezione riguardante i parametri avanzati, inoltre, ogni utente può verificare, ed eventualmente modificare, le impostazioni automatiche riguardo *Epochs* (numero di volte in cui l'algoritmo di apprendimento esaminerà i dati di addestramento e si valuterà sia sul training set che sul validation set), e *Early Stopping* (numero minimo di epochs per addestramento).
Una volta compilato il pannello relativo ai metadati, viene scelto il modello di base che verrà usato dal motore di riconoscimento di testo come punto di partenza per addestrare il nuovo modello, secondo il processo di *fine-tuning*. È stato selezionato un modello che potesse avvicinarsi maggiormente alle caratteristiche del documento, filtrando i diversi modelli disponibili per tipologia (handwriting) e per lingua e privilegiando modelli addestrati su testi multilingui. La scelta è ricaduta sul modello denominato *Italian 16th century antiquarian hand - Pirro Ligorio 0.3*, addestrato su un corpus plurilingue (latino, greco, italiano) di 57.003 parole tratte dall'*Enciclopedia dell'Antichità*, opera di carattere enciclopedico (Il modello HTR, s.d.; *Italian 16th century antiquarian hand - Pirro Ligorio 0.3*, s.d.). Queste caratteristiche, unite al tasso di errore CER del 2,70%, rendono il modello adatto a costituire una base valida per l'addestramento del nuovo modello. Inoltre, come evidenziato da recenti studi sperimentali condotti su corpora plurilingui, un singolo modello HTR multilingue costituisce un'alternativa efficace e conveniente a singoli modelli monolingui, soprattutto nel caso in cui le diverse sezioni linguistiche non siano facilmente individuabili e separabili in maniera automatica (Capurro et al., 2023). Questo permette di ovviare in maniera efficiente alla mancanza di un modello di riconoscimento specifico per la lingua sarda, che, in quanto lingua minoritaria, non è attualmente rappresentata nel panorama linguistico di Transkribus. A questo punto è stato possibile proseguire con l'avvio del training vero e proprio.
Durante il primo ciclo di addestramento, l'attività di apprendimento viene svolta sul training set e, progressivamente, le prestazioni del modello vengono valutate mediante il calcolo del Character Error Rate (CER), effettuato sui dati di convalida del validation set. Il processo di addestramento si conclude automaticamente quando il modello non mostra più ulteriori miglioramenti, ossia quando viene raggiunto il valore minimo di CER.
Nel nostro caso, al termine del primo ciclo di training, il modello ha raggiunto un CER del 3,84%. Si tratta di un risultato incoraggiante, soprattutto in questa fase iniziale, in quanto da una parte conferma la solidità del modello di base scelto per il fine-tuning, e dall'altra lascia ben sperare in un buon potenziale di miglioramento, da verificarsi nelle successive fasi di addestramento incrementale. La figura 3 riporta il grafico della curva di apprendimento (Learning Curve), generata alla fine del primo ciclo di addestramento, che mostra la progressiva diminuzione del CER all'aumentare del numero di epoch.

[3] Poiché ciascuna riproduzione fotografica mostra il volume aperto su due facciate, tale selezione corrisponde complessivamente a 40 pagine manoscritte.

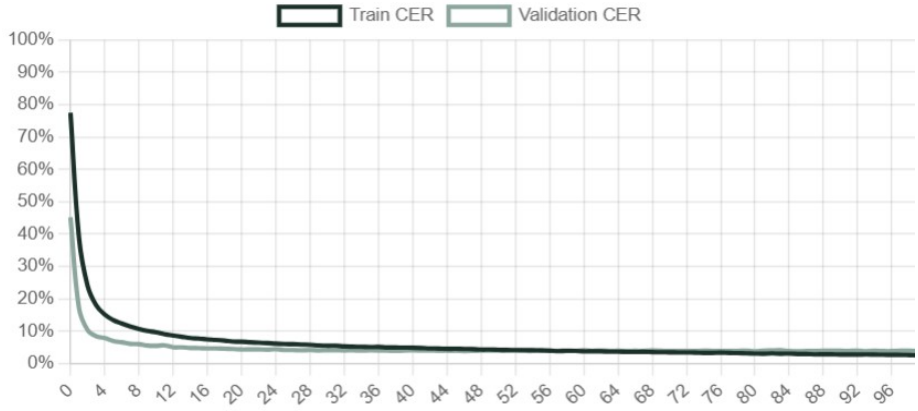


Figura 3. Grafico della *Learning curve* alla fine del primo ciclo di addestramento.

La fase successiva prevede di applicare il modello addestrato anche ad altre pagine mai viste. Questo permette sia di testare la reale capacità di generalizzazione del modello su dati non elaborati in precedenza, sia di ampliare progressivamente la Ground Truth, in modo da poter proseguire l'addestramento su un dataset sempre più ampio, con l'obiettivo di aumentare il tasso di attendibilità e ridurre ulteriormente il CER.

6. RISULTATI

Applicando il riconoscimento automatico a un insieme di pagine non incluse nell'iniziale Ground Truth, sono state generate nuove trascrizioni, successivamente analizzate mediante confronto diretto con le immagini del manoscritto. Da una prima analisi quantitativa degli errori è emersa una media di tre errori per riga, corrispondente a un tasso di errore per carattere prossimo al 3%, coerente con il CER ottenuto in fase di addestramento. Dal punto di vista qualitativo gli errori individuati si dividono principalmente in errori dovuti alla segmentazione non ottimale del layout o alla qualità dell'immagine, compromessa in casi sporadici da macchie e rumore generico, ed errori ortografici, riguardanti la resa delle maiuscole o il riconoscimento delle vocali accentate.

Le sezioni che presentano le maggiori criticità si confermano quelle caratterizzate da frequenti cambi di lingua, in cui si alternano, all'interno della stessa riga, parole in caratteri latini, riconducibili a sardo, latino o italiano, e termini in greco. Questa tendenza appare indicativa dei limiti intrinseci di un'architettura, come quella di PyLaia, che, pur dimostrandosi solida e veloce nell'apprendimento visivo-sequenziale, mostra una minore efficacia nella modellazione di contesti linguistici variabili, confrontata con i risultati dei più recenti Transformer, come TrOCR, secondo quanto emerge da recenti studi comparativi sulle moderne tecnologie HTR (Romein et al., 2025).

Nonostante ciò, il modello evidenzia un buon livello di maturità già dopo il primo ciclo di addestramento, pur richiedendo ancora una revisione sistematica delle trascrizioni.

Ulteriori cicli di addestramento incrementale potranno migliorare le prestazioni e garantire trascrizioni sempre più affidabili ed esportabili. La criticità principale resta legata all'analisi del layout, che richiede un'attenzione particolare e un addestramento mirato di un modello specifico per la sezione del dizionario, caratterizzato da una struttura più eterogenea e complessa.

Trattandosi di un'opera di notevole rilievo storico-linguistico e considerata la cospicua estensione del testo, l'avviamento di un progetto di digitalizzazione di questo tipo faciliterebbe la realizzazione di un'edizione critica integrale del manoscritto. L'esportazione delle trascrizioni in formato TEI XML permette infatti la condivisione dei dati nonché il loro uso in differenti contesti di ricerca e disseminazione, favorendo una fruizione più accessibile e partecipata anche al di fuori della comunità accademica, in una prospettiva di public engagement capace di rispondere alle odierne esigenze della ricerca nell'ambito dell'umanistica digitale (Pierazzo, 2016).

BIBLIOGRAFIA

Capurro, C., Provatorova, V., & Kanoulas, E. (2023). Experimenting with Training a Neural Network in Transkribus to Recognise Text in a Multilingual and Multi-Authored Manuscript Collection. Heritage, 6(12), 7482–7494. https://doi.org/10.3390/heritage6120392

Data Preparation. (s.d.). Recuperato 27 gennaio 2026, da https://help.transkribus.org/data-preparation

How to train a model in Transkribus. (2025, ottobre 22). https://blog.transkribus.org/en/how-to-train-a-text-recognition-model-in-transkribus

Il modello HTR. (s.d.). LIGORIO DIGITALE 1.0 WORKSPACE. Recuperato 2 febbraio 2026, da https://lidiws-limes.cfs.unipi.it/trascrizione/ilmodellohtr/

Italian 16th century antiquarian hand—Pirro Ligorio 0.3. (s.d.). Recuperato 4 febbraio 2026, da https://app.transkribus.org/models/public/text/42105

Jungo, M., Vögtlin, L., Fakhari, A., Wegmann, N., Ingold, R., Fischer, A., & Scius-Bertrand, A. (2023). Impact of Ground Truth Quality on Handwriting Recognition. Proceedings of the 12th International Symposium on Information and Communication Technology, 135–140. https://doi.org/10.1145/3628797.3628976

Kiessling, B., Tissot, R., Stokes, P., & Stökl Ben Ezra, D. (2019). eScriptorium: An Open Source Platform for Historical Document Analysis. 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW), 19–19. https://doi.org/10.1109/ICDARW.2019.10032

Model Setup and Training. (s.d.). Recuperato 27 gennaio 2026, da https://help.transkribus.org/model-setup-and-training

Murgia, G. (2023). «Tutto ciò è necessario alla Sarda nazione non men che a tutte le altre dell'Europa». La tradizione manoscritta e a stampa del Ripulimento della lingua sarda di Matteo Madau. Il «traffico delle lingue». Idiomi a contatto in Sardegna e nel Mediterraneo in età preunitaria, eds. R. Fresu, P. Maninchedda, G. Murgia, P. Serra, Cagliari, UNICApress, 2023, pp. 491-529. ISBN: 978-88-3312-112-3. https://doi.org/10.13125/UNICAPRESS.978-88-3312-108-6

Pierazzo, E. (2016). 3. Modelling Digital Scholarly Editing: From Plato to Heraclitus. In M. J. Driscoll & E. Pierazzo (A c. Di), *Digital Humanities Series* (1ª ed., Vol. 4, pp. 41–58). Open Book Publishers. https://doi.org/10.11647/OBP.0095.03

Puigcerver, J. (2017). Are Multidimensional Recurrent Layers Really Necessary for Handwritten Text Recognition? 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), 67–72. https://doi.org/10.1109/ICDAR.2017.20

Puigcerver, J. (2026). Jpuigcerver/PyLaia [Python]. https://github.com/jpuigcerver/PyLaia (Opera originale pubblicata 2017)

PyLaia: Teklia Documentation. (s.d.). Recuperato 27 gennaio 2026, da https://doc.teklia.com/pylaia/

Recognition and Enrichment of Archival Documents | READ | Project | Fact Sheet | H2020. (s.d.). CORDIS | European Commission. Recuperato 4 febbraio 2026, da https://cordis.europa.eu/project/id/674943

Romein, C. A., Rabus, A., Leifert, G., & Ströbel, P. B. (2025). Assessing advanced handwritten text recognition engines for digitizing historical documents. International Journal of Digital Humanities, 7(1), 115–134. https://doi.org/10.1007/s42803-025-00100-0

Sánchez, J. A., Mühlberger, G., Gatos, B., Schofield, P., Depuydt, K., Davis, R. M., Vidal, E., & de Does, J. (2013). tranScriptorium: A european project on handwritten text recognition. Proceedings of the 2013 ACM symposium on Document engineering, DocEng '13, 227–228. https://doi.org/10.1145/2494266.2494294

Stokes, P. A., Kiessling, B., Stökl Ben Ezra, D., Tissot, R., Gargem, E. H. (2021). The eScriptorium VRE for Manuscript Cultures. Classics@, 2021, 18 (1), https://classics-at.chs.harvard.edu/classics18-stokes-kiessling-stokl-ben-ezra-tissot-gargem/. ⟨hal-03991423⟩

Tarride, S., Schneider, Y., Generali-Lince, M., Boillet, M., Abadie, B., & Kermorvant, C. (2024). Improving Automatic Text Recognition with Language Models in the PyLaia Open-Source Library (arXiv:2404.18722; Versione 1). arXiv. https://doi.org/10.48550/arXiv.2404.18722

Training Text Recognition Models. (s.d.). Recuperato 27 gennaio 2026, da https://help.transkribus.org/training-text-recognition-models

Modelli digitali interattivi per gli archivi teatrali storici: mappare le relazioni semantiche al Teatro Metastasio di Prato (1827–1860)

MATILDE INNOCENTI

UNIVERSITÀ DEGLI STUDI DI FIRENZE, ITALY -
MATILDE.INNOCENTI@UNIFI.IT

ABSTRACT (ITALIANO)

A partire da un'analisi storico-archivistica sistematica, il contributo propone una riflessione teorica che si concretizza nello sviluppo di un modello digitale per la conservazione, l'accessibilità e la valorizzazione del fondo del Teatro Metastasio di Prato relativo al periodo 1827–1860. Il modello introduce una descrizione "aumentata", intesa come superamento della mediazione archivistica esclusivamente gerarchica, a favore di un approccio ergodico, capace di integrare entità eterogenee e di investigare in modo modulare e dinamico relazioni semantiche complesse. Interattivo e configurabile, il modello si fonda su un sistema basato su un modello relazionale, implementato in un database concepito come infrastruttura di interrogazione e interpretazione, in cui la definizione di categorie interpretative consente di incrociare i dati e di mettere in dialogo i contenuti all'interno di una struttura a matrice stratificata. Il modello si configura come una proposta di data modelling inteso come pratica interpretativa nell'ambito delle Digital Humanities. Tale impostazione permette di restituire e ricostruire, sia graficamente sia attraverso i testi, le dinamiche istituzionali, amministrative ed economiche che regolavano il funzionamento del Teatro. L'adozione di un repository pubblico mira, infine, a garantire la riusabilità e la sostenibilità del modello, prevedendo la possibilità di una sua espansione e integrazione nel tempo.
Parole chiave: archivi digitali; studi sul teatro storico; patrimonio culturale; modelli interattivi dei dati; analisi semantica delle reti

ABSTRACT (ENGLISH)

Interactive and Augmented Digital Models for Historical Theater Archives: Mapping Semantic Relations at the Teatro Metastasio (1827–1860).
Building on a systematic historical-archival analysis, this contribution proposes a theoretical reflection materialized in the development of a digital model for the preservation, accessibility, and valorization of the Teatro Metastasio archive in Prato (1827–1860). The model introduces an "augmented" description, moving beyond exclusively hierarchical archival mediation toward an ergodic, interactive approach capable of integrating heterogeneous entities and dynamically mapping complex semantic relationships. The model, configurable and interactive, is implemented as a system based on a relational model, designed as a digital infrastructure for inquiry and interpretation, where interpretive categories enable cross-referencing of data and dialogue between contents within a stratified matrix structure. The model can be understood as a data modelling approach conceived as an interpretative practice within Digital Humanities. This framework facilitates the reconstruction, both graphically and textually, of the institutional, administrative, and economic dynamics governing the theater's operations. The adoption of a public repository ensures the reusability and sustainability of the model, supporting its potential expansion, integration, and continuous adaptation over time.
Keywords: digital archives; historical theater studies; cultural heritage; interactive data models; network analysis

1. INTRODUZIONE

Negli ultimi anni la riflessione teorica e metodologica nel campo delle Digital Humanities ha evidenziato con crescente chiarezza i limiti dei modelli tradizionali di descrizione archivistica nel restituire la complessità dei patrimoni documentari storici. In particolare, archivi caratterizzati da una forte eterogeneità tipologica e da una fitta rete di relazioni funzionali tra documenti, soggetti e istituzioni pongono interrogativi rilevanti circa le modalità di mediazione, accesso e interrogazione delle fonti in ambiente digitale. In questo contesto, il concetto di "descrizione aumentata" si propone come una prospettiva teorica e operativa volta a suggerire nuovi modelli per ripensare la funzione dei sistemi

informativi archivistici. In questo senso, il contributo si colloca nel dibattito attuale delle Digital Humanities sul data modelling come pratica interpretativa, in cui la formalizzazione di entità e relazioni non si limita a supportare l'organizzazione dei dati, ma contribuisce attivamente a definire il quadro epistemologico attraverso cui il patrimonio culturale viene analizzato e compreso. Il modello proposto deve pertanto essere inteso non come una semplice soluzione tecnica, ma come una proposta metodologica per la rappresentazione di conoscenze archivistiche complesse attraverso relazioni semantiche strutturate. La ridefinizione della descrizione non riguarda esclusivamente la quantità o la granularità dei metadati, ma interessa la struttura stessa, orientandola verso modelli in grado di integrare entità diverse e di chiarire le relazioni che ne determinano il significato storico e i processi funzionali.
Il contributo si inserisce in questo quadro proponendo un caso di studio relativo al Fondo del Teatro Metastasio di Prato. L'obiettivo è mostrare come, a partire da un'analisi storico-archivistica approfondita, sia possibile progettare un sistema digitale che non si limiti alla rappresentazione dell'archivio, ma che funzioni come strumento di interrogazione critica e di restituzione visuale dei dati. Attraverso la definizione di categorie interpretative, la costruzione di un modello relazionale implementato in un database e l'adozione di strumenti di visualizzazione, il progetto mira a rendere leggibili le dinamiche decisionali, amministrative ed economiche che regolavano il funzionamento del Teatro.

2. IL FONDO DEL TEATRO METASTASIO COME CASO DI STUDIO

Il Fondo del Teatro Metastasio conservato presso l'Archivio di Stato di Prato documenta in modo esteso e continuativo la vita del principale teatro cittadino. Il presente studio prende in esame il periodo compreso tra il 1827 e il 1861, individuato come l'arco cronologico più coerente e densamente documentato per la ricostruzione della fase costitutiva e preunitaria dell'istituzione teatrale.
Il fondo si presenta come un insieme estremamente articolato, comprendente atti amministrativi e deliberazioni dell'Accademia del Teatro, regolamenti, protocolli delle adunanze, corrispondenza (interna ed esterna), contratti con impresari e maestranze, documentazione contabile, materiali promozionali, oltre a un nucleo grafico costituito da disegni tecnici, planimetrie, prospetti e memorie redatte da architetti e scenotecnici. Tale eterogeneità e complessità rende il fondo particolarmente significativo per l'analisi delle pratiche amministrative, tecniche e culturali legate alla gestione di un teatro ottocentesco.
L'analisi storico-archivistica condotta ha messo in evidenza come il funzionamento del Teatro Metastasio fosse il risultato di un sistema complesso di relazioni tra diversi organi decisionali, attori istituzionali e agenti della produzione spettacolare. Queste dinamiche, spesso distribuite su più tipologie documentarie e su serie archivistiche differenti, risultano difficilmente ricostruibili attraverso una consultazione lineare delle unità archivistiche. Questa configurazione ha orientato la scelta di adottare un approccio di modellazione digitale fondato sulla definizione di categorie interpretative e sulla formalizzazione delle relazioni tra entità. Il fondo, così configurato, si propone non solo come oggetto di descrizione, ma come spazio di sperimentazione per la costruzione di un sistema informativo capace di restituire la dimensione processuale e relazionale della documentazione teatrale.

3. DALLA DESCRIZIONE ARCHIVISTICA ALLA DESCRIZIONE AUMENTATA

La progettazione del modello digitale si fonda sull'assunzione che la descrizione archivistica non sia un'operazione neutra, ma una pratica interpretativa che riflette scelte teoriche e metodologiche. Nel contesto delle Digital Humanities, questa prospettiva si inserisce in approcci recenti che sottolineano il ruolo del modelling come forma di interpretazione scientifica, in cui le strutture dei dati incorporano specifiche domande di ricerca e prospettive analitiche, piuttosto che configurarsi come rappresentazioni neutrali della realtà.
La definizione delle categorie non risponde unicamente a esigenze di ordinamento, ma costituisce uno strumento per rendere esplicite le strutture di funzionamento dell'istituzione teatrale. Dal punto di vista della Storia del Teatro, la comprensione di tali strutture consente di ricostruire i processi decisionali, di delineare l'articolazione dei ruoli all'interno degli organi direttivi e di analizzarne le modalità di coordinamento operativo: un approccio questo ancora marginale nella storiografia dello spettacolo dal vivo, che ha tradizionalmente privilegiato l'analisi dei testi e delle pratiche rappresentative rispetto alla dimensione organizzativa e istituzionale. Questa lacuna è stata progressivamente affrontata nell'ambito dei digital theatre studies, dove la necessità di ricostruire archivi performativi dispersi ed eterogenei ha favorito lo sviluppo di approcci relazionali e basati su reti per la storiografia teatrale.
In questo quadro, emerge l'esigenza di strumenti e approcci ipertestuali capaci di connettere testi e

materiali eterogenei, offrendo una visione più integrata e dinamica. La Storia dello Spettacolo, infatti, si fonda proprio sulla ricomposizione e sulla messa in relazione di documenti eterogenei, come quelli conservati nel Fondo Metastasio. Il modello digitale adottato consente di valutare in che misura l'ordinamento funzionale proposto sia in grado di restituire una rappresentazione coerente del funzionamento teatrale, rendendo tale ricostruzione un punto di osservazione privilegiato per l'analisi storica. L'approccio adottato valorizza inoltre l'elevata replicabilità del modello, che potrebbe auspicabilmente essere applicato ad altri fondi teatrali.
Il modello elaborato individua un insieme di entità principali, tra le quali tipologie di documenti, persone, luoghi, eventi, organi, ciascuna dotata di attributi specifici e collegata alle altre attraverso relazioni semantiche formalizzate. Tale struttura consente di superare una rappresentazione puramente gerarchica del fondo, avvicinandosi a una logica reticolare ispirata ai principi di Records in Contexts (RiC) e agli orientamenti degli standard descrittivi di seconda generazione, che privilegiano la rappresentazione delle relazioni rispetto a più rigide strutture gerarchiche.
La descrizione aumentata non si limita ad integrare informazioni supplementari, ma rende interrogabili le relazioni funzionali che legano, ad esempio, un atto deliberativo a un organo decisionale, a un evento specifico o a una serie di corrispondenze. Il modello permette così di ricostruire processi decisionali, catene di responsabilità e dinamiche amministrative che difficilmente potrebbero emergere dalla descrizione tradizionale delle unità archivistiche se studiate singolarmente.

4. IL DATABASE COME INFRASTRUTTURA DI INTERROGAZIONE

A partire dal modello concettuale, è stato progettato un sistema basato su un modello relazionale, implementato in un database, concepito non come semplice backend descrittivo, ma come infrastruttura di ricerca. Il database consente di effettuare interrogazioni incrociate tra le diverse entità, rendendo possibile l'esplorazione dei materiali secondo criteri molteplici: temporali, istituzionali, funzionali o relazionali. Questo approccio riflette un più ampio spostamento nelle Digital Humanities dalla semplice conservazione dei dati al data modelling, in cui l'attenzione si concentra sull'interpretabilità e sulla relazionalità dei dati, piuttosto che sulla loro mera accumulazione.
Una caratteristica centrale del progetto è la possibilità di interrogare il sistema in modo altamente specifico, combinando più parametri e visualizzando i risultati attraverso differenti modalità di restituzione. Le timeline consentono di seguire l'evoluzione delle decisioni e degli eventi nel tempo; i grafici a bolle consentono di analizzare la distribuzione e il peso di determinate voci economiche; le reti di corrispondenza rendono visibili i flussi comunicativi e i rapporti tra i diversi attori coinvolti nella gestione del Teatro.
Il database sarà accessibile attraverso un repository pubblico, scelta che risponde a esigenze di trasparenza metodologica, riusabilità e sostenibilità. La pubblicazione del modello e delle strutture dati consente infatti di valutare criticamente le scelte adottate e di favorire il riuso dell'infrastruttura in contesti di ricerca analoghi. Le visualizzazioni non sono concepite come strumenti accessori o meramente divulgativi, ma come vere e proprie forme di restituzione critica dei dati archivistici. Attraverso la rappresentazione grafica delle relazioni e dei processi, il sistema consente di far emergere dinamiche che resterebbero altrimenti implicite nella documentazione.
In particolare, la possibilità di intrecciare dati amministrativi, corrispondenza ed eventi permette di osservare il funzionamento degli organi teatrali come sistema complesso, evidenziando ruoli, competenze ed interazioni. Le visualizzazioni diventano così strumenti di conoscenza, capaci di affiancare ed integrare l'analisi storica tradizionale.

5. CONCLUSIONI E PROSPETTIVE

In questa prospettiva, il contributo non introduce nuove tecnologie in senso stretto, ma propone un avanzamento metodologico nel modelling e nell'interpretazione dei dati archivistici all'interno del quadro delle Digital Humanities.
Il caso di studio del Fondo del Teatro Metastasio mostra come l'adozione di modelli di descrizione aumentata e di infrastrutture digitali orientate all'interrogazione possa contribuire in modo significativo alla conservazione, all'accessibilità e alla valorizzazione di archivi storici complessi. La combinazione tra analisi archivistica, modellazione concettuale, database interrogabile e visualizzazione dei dati consente di restituire la natura dinamica dei processi documentati, sostenendo nuove prospettive di ricerca. Il progetto si colloca all'interno del dibattito contemporaneo sugli archivi digitali e i sistemi informativi interculturali, offrendo un esempio di integrazione tra rigore storico e proposta digitale. In prospettiva,

l'approccio adottato potrebbe essere sperimentato anche in altri contesti archivistici teatrali, contribuendo, in dialogo con esperienze analoghe, allo sviluppo di infrastrutture sostenibili e condivise per lo studio del patrimonio culturale performativo.

BIBLIOGRAFIA

Alonge, R. (1988). Teatro e spettacolo nel secondo Ottocento. Torino: Einaudi.

Angelini, F. (1988). Il teatro italiano dell'Ottocento. Roma-Bari: Laterza.

Ciotti, F. (a cura di). (2023). Digital humanities. Metodi, strumenti, saperi. Roma: Carocci.

EGAD – Experts Group on Archival Description & International Council on Archives (ICA). (2023). Records in Contexts – Conceptual Model (RiC-CM). Paris: ICA.

Franceschinelli, R. (a cura di.). (2021). Spazi del possibile. I nuovi luoghi della cultura e le opportunità della rigenerazione. Milano: Franco Angeli Editore.

Flanders, J., & Jannidis, F. (2015). Data Modeling. In S. Schreibman, R. Siemens, & J. Unsworth (a cura di), A New Companion to Digital Humanities. Chichester: John Wiley & Sons, Ltd.

Garbero Zorzi, E. (Ed.). (1988). Storia del teatro in Italia. L'Ottocento. Torino: Einaudi.

Garbero Zorzi, E., & Zangheri, L. (1998). I teatri storici della Toscana. Venezia: Marsilio.

Guercio, M. (2019). Archivistica informatica. I documenti in ambiente digitale. Roma: Carocci.

Lucchesini, P., & Casale, C. (1995). Storia del Teatro Metastasio (Vol. II). Firenze: Pontecorboli Editore.

Mandarano, N. (2019). Musei e media digitali. Roma: Carocci Editore.

Meldolesi, C., & Taviani, F. (1980). Teatro e spettacolo nel primo Ottocento. Roma: Laterza. Nembi, O. (1995). Storia del Teatro Metastasio (Vol. I). Firenze: Pontecorboli Editore.

Pacini, G. P. (2001). Il teatro in Toscana tra Sette e Ottocento. Pisa: ETS.

Previtali, G. (2023). Che cosa sono le digital humanities. Roma: Carocci Editore.

Romiti, A. (2020). Archivistica generale. Primi elementi (6ª ed.). Lucca: Civita.

Schnapp, J. (2015). Digital humanities. Mattei, M. G. (Ed.). Milano: Egea.

Sendra, P., & Sennett, R. (2022). Progettare il disordine. Roma: Treccani.

Silvestro, N. (2022). Manuale di archivistica (5ª ed.). Napoli: Edizioni Simone Giuridico.

Sorba, C. (2001). Teatri. L'Italia del melodramma nell'età del Risorgimento. Bologna: Il Mulino.

Svensson, P. (2016). Big Digital Humanities: Imagining a Meeting Place for the Humanities and the Digital. Ann Arbor (MI): University of Michigan Press.

Valacchi, F. (2021). Gli archivi tra storia, uso e futuro. Dentro la società. Roma: Editrice Bibliografica. Valacchi, F. (2022). Diventare archivisti. Roma: Editrice Bibliografica.

Vitali-Rosati, M. (2018). On Editorialization: Structuring Space and Authority in the Digital Age. Amsterdam: Institute of Network Cultures.

Tra fotografia e segno: la trascrizione informatizzata dei registri storici Alinari

PAMELA KRZEMIEN
UNIVERSITÀ DI PISA, ITALY - PAMELAKRZEMIEN@LIVE.IT

ABSTRACT (ITALIANO)

Il progetto di ricerca è incentrato sulla valorizzazione dei registri storici Alinari pertinenti il fondo Patronato: si tratta di due registri, ordinati rispettivamente in ordine cronologico e alfabetico, prodotti dalla storica società fotografica Fratelli Alinari per documentare la produzione di negativi e positivi su commissione (1893-2019).
Al fine di poter ricollegare i negativi fotografici alle informazioni manoscritte, si è resa necessaria una trascrizione informatizzata dei registri attraverso un procedimento HTR; per tale procedura è stata utilizzata la piattaforma Transkribus.
Data l'eterogeneità della tipologia di grafie e l'ampiezza dell'arco cronologico, sono stati sperimentati alcuni modelli per il *recognition text* a partire da due periodi temporali: quello relativo ad una scrittura più calligrafica, tipica della fine del XIX sec./inizio XX, e quello relativo ad una grafia più confusa e imprecisa, *post* Seconda Guerra Mondiale. Per l'organizzazione del *layout* è stato necessario realizzare un *table model* dato che risultava totalmente assente nella lista dei modelli presenti all'interno del programma.
Il risultato è stato parzialmente migrato su Excel e alla trascrizione delle singole commissioni sono state affiancate manualmente le stringhe identificative delle immagini delle lastre.
L'importanza del progetto risulta fondamentale considerati: la fragilità dei manoscritti cartacei; la temporanea non accessibilità del materiale Alinari; la praticità nel reperimento delle informazioni necessarie; la possibilità di una pubblicazione dell'intero inventario sul sito della Fondazione; ed infine per la fattiva possibilità di estendere la trascrizione informatizzata agli altri registri prodotti dalla stessa società.
Parole chiave: Handwritten Text Recognition; Transkribus; Intelligenza Artificiale; registri; fotografia.

ABSTRACT (ENGLISH)

Between photography and sign: the computerized transcription of Alinari historical registers.
The research project focuses on enhancing the historical Alinari registers pertaining to the Patronato collection: these are two registers, arranged chronologically and alphabetically respectively, produced by the historic photographic company Fratelli Alinari to document the production of commissioned negatives and positives (1893-2019).
In order to connect the photographic negatives to the manuscript information, a computerized transcription of the registers was necessary using an HTR process; the Transkribus platform was used to document this process.
Given the heterogeneity of the handwriting typologies and the breadth of the chronological span, several models for text recognition were tested starting from two time periods: one relating to a more calligraphic handwriting, typical of the late 19th/early 20th century, and one relating to a more confused and imprecise handwriting, after Second World War. To organize the layout, it was necessary to create a table model, as it was completely absent from the list of templates available within the program.
The result was partially migrated to Excel, and the transcription of the individual commissions was accompanied by the image identification strings of the final ones.
The importance of the project was fundamentally considered: the fragility of the paper manuscripts; the temporary inaccessibility of the Alinari material; the ease of finding the necessary information; the possibility of publishing the entire inventory on the Foundation's website; and finally, the practical possibility of extending the computerized transcription to the other registers produced by the same company.
Keywords: Handwritten Text Recognition; Transkribus; Artificial Intelligence; registers; photography.

1. INTRODUZIONE

Nel 2019, con la legge del 13 novembre capo IV, art. 55, la Regione Toscana ha acquistato il patrimonio fotografico della società Fratelli Alinari IDEA, costituito da 5.000.000 di pezzi tra negativi, positivi, attrezzature fotografiche, album, libri, unicum.
Da allora, in virtù di attendere una definitiva e adeguata sistemazione, tutto il materiale è stato trasferito, temporaneamente, entro un caveu, in una condizione assolutamente difficoltosa in quanto a ricerca e valorizzazione.
Nonostante la situazione complicata dell'archivio, le attività ordinarie e straordinarie della neonata Fondazione Alinari per la Fotografia non sono mai mancate; tra queste è da sottolineare la grande ed impegnativa campagna di digitalizzazione dei positivi e negativi della cosiddetta lastroteca storica.
Il seguente progetto è parte di un programma più ampio di valorizzazione di uno dei fondi storici Alinari ossia il cosiddetto "Patronato", relativo alle commissioni fotografiche esterne.
Come sottolineato anche da Tomassini (2019) tale fondo è rimasto poco affrontato in letteratura, nonostante l'importante mostra del 1977 al Forte Belvedere: *Gli Alinari fotografi a Firenze 1852-1920* (Settimelli & Zevi, 1977), dove si valorizzavano anche le fotografie escluse dai cataloghi storici, ossia quelle realizzate su commissione, per realizzazione di ritratti, documentazioni di beni o altre evenienze.
In quanto ad un'analisi specifica degli ambienti e del modo di interpretare oggetti, soggetti e mestieri, un'analisi piuttosto interessante ne dette dunque Arturo Carlo Quintavalle nella sua monografia *Gli Alinari* (2003), mentre a citarne per la prima volta i documenti di corredo fu Emanuela Sesti (2022).
Il fondo "Patronato" è composto da 28.120 negativi tra i quali 11.177 sono lastre su vetro (Scartoni, in corso di stampa); per orientarsi cronologicamente e fattivamente nella lastroteca sono stati dunque prodotti negli anni una discreta quantità di registri. Il seguente progetto mira alla valorizzazione dei due in particolare citati da Sesti: la *Rubrica numerica e nominativa delle Commissioni Private dal 1 gennaio 1893 al 1.4.2019* e il *Registro numerico delle commissioni private dal gennaio 1893 al aprile 2019*.
I due documenti, prodotti rispettivamente in ordine alfabetico e cronologico, risultano essere due mezzi di corredo fondamentali per l'identificazione e la datazione di una parte dei negativi del Patronato e, pertanto, ne è risultata necessaria una trascrizione informatizzata, attraverso l'utilizzo di tecnologie HTR.
Considerando che l'utilizzo di tecnologie OCR non sarebbe stato efficace e avendo a disposizione materiale manoscritto di circa 950 pagine, è stato deciso di utilizzare la piattaforma Transkribus.

2. METODOLOGIA

Il progetto di trascrizione dei registri è da racchiudere entro sei fasi principali:

- Analisi dei documenti (grafie, pagine da escludere, qualità della digitalizzazione ecc.)
- Test di *recognition text* e *layout analyses* attraverso l'utilizzo di *supermodels*
- Creazione del modello tabellare per una corretta organizzazione del *layout*
- Applicazione del modello tabellare e correzione
- Creazione e sperimentazione di differenti modelli di *recognition text*
- Applicazione, correzione e validazione dei risultati

L'analisi dei due documenti similari, ma con organizzazione e quantità di dati differenti, ha posto fin dal principio delle perplessità circa l'ampiezza dell'arco cronologico da affrontare in ragione anche della discreta quantità di mani diverse, susseguitesi all'interno degli stessi registri. La qualità delle pagine digitalizzate è risultata piuttosto buona anche se talvolta riusltavano non propriamente orientate correttamente in modo perpendicolare; altre pagine invece sono state valutate da escludere per la trascrizione, in quanto vuote o digitalizzate due volte a causa di post-it nei documenti che richiedevano una doppia documentazione.
Dopo un'iniziale sperimentazione di *supermodels* pubblici, quali "Text Titan I ter", risultati ampiamente fallaci nella trascrizione, sono emerse due considerazioni: la fondamentale preparazione del *layout* entro strutture tabellari e la necessaria realizzazione di nuovi modelli realizzati *ad hoc*.
Difatti ogni pagina del registro cronologico è caratterizzata da una conformazione estremamente complessa, date 9 colonne e una media di 55 righe per pagina. Di conseguenza, l'applicazione manuale di ciascuna tabella su ogni pagina, restituiva un lavoro impegnativo e lungo, elemento che ha spinto alla realizzazione di un modello apposito per la struttura tabellare. Questa attività è risultata necessaria considerando la totale assenza di un modello *table* pubblico nella piattaforma, se si fa eccezione per

"Amsterdam Tables" il quale non ha dato nessun tipo di risultato e non pare riportare alcuna informazione circa l'addestramento.
Riguardo la realizzazione del modello applicabile sul documento, si è resa necessaria la sperimentazione di più prototipi da applicare al contenuto, anche se è risultata impossibile crearne uno per ogni grafia: la presenza di ciascuna mano sulle pagine risultava di volta in volta insufficiente per la realizzazione del *Ground Truth*.

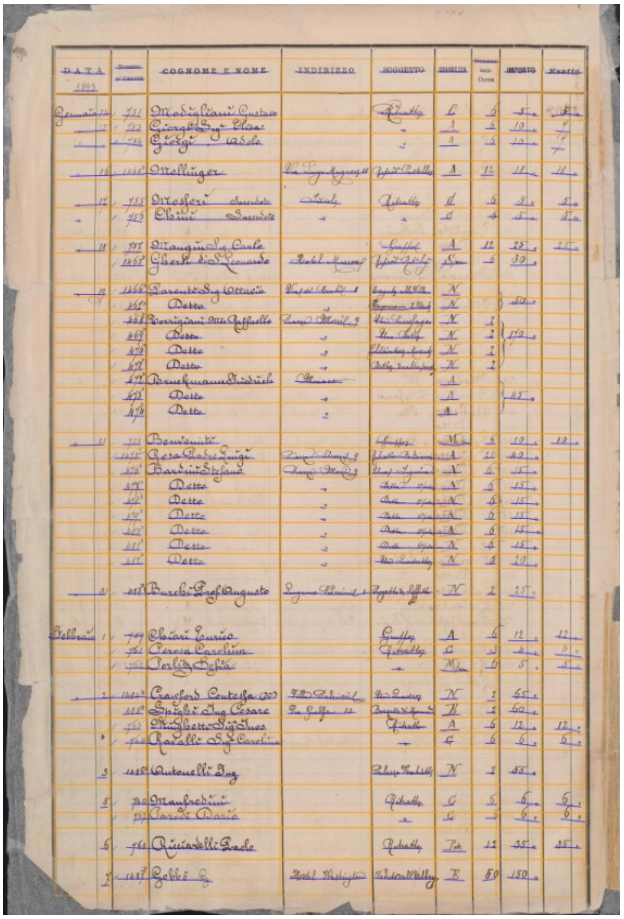


Figura 1 - Registro numerico delle commissioni private dal gennaio 1983 al aprile 2019

Di

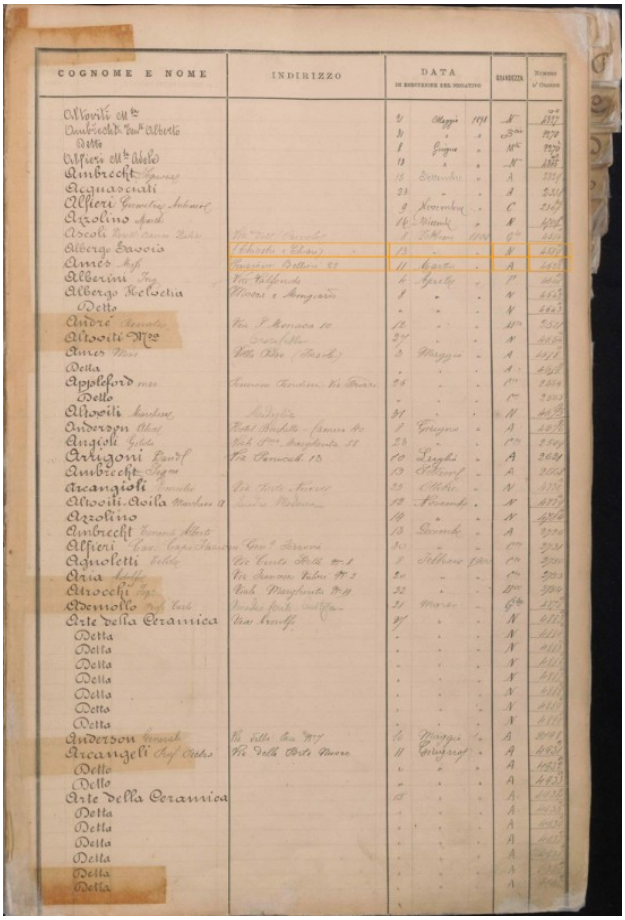


Figura 2 - Rubrica numerica e nominativa delle Commissioni Private dal 1 gennaio 1893 al 1.4.2019

conseguenza, più che la realizzazione di un modello per ogni scrittura, è stato deciso orientativamente di realizzarne due: uno per le grafie più precise, omogenee ed ordinate (fine del XIX-inizio XX sec.); l'altro per quelle maggiormente disordinate e differenti (seconda metà del XX sec[1]).

[1] Un elemento da non sottovalutare e dover tenere in considerazione, indicativo anche della forte differenza non solo aziendale della Fratelli Alinari, ma anche in termini di produzione su commissione, è la differente quantità di pagine prodotte entro uno stesso anno, se messe a confronto tra la fine XIX/inizio XX secolo, e la seconda metà dello stesso. Per comprendere esattamente il significato di ciò, basterà quindi fornire due numeri: 50 pagine consecutive del registro cronologico, a partire dal 1 gennaio 1893, corrispondono a poco più di un anno di attività degli Alinari, mentre lo stesso quantitativo di pagine circa, a partire dal 1942, attraversa l'attività di 38 anni (1942-1960). Da tale differenza in termini quantitativi di commissioni, non solo ne deriva una evidente gestione diversa della società, i cui momenti cardine di cambiamenti societari sono da individuarsi negli anni 1920 (conduzione Ricasoli), 1958 (conduzione Vittorio Cini), 1973 (conduzione Renato Zevi), 1982 (conduzione Claudio de Polo Saibanti), ma se ne comprende anche il motivo per il quale i cambiamenti delle mani che scrivono siano così frequenti. Inoltre, come si può osservare anche solo sfogliando i registri, dopo la Seconda Guerra Mondiale questi documenti paiono formalmente più una raccolta di appunti che non un vero e proprio schedario preciso come era stato concepito nell'ultimo decennio del XIX sec, sotto la direzione del brillante Vittorio Alinari. Nel *Report Finale: il patrimonio fotografico Alinari*, sono ben descritti i cambiamenti societari sopra accennati e difatti, a proposito del fondo Patronato si specifica che "Si trattava di un tipo di lavorazione già secondario nell'economia della ditta, e addirittura quasi abbandonato dalla conduzione Ricasoli in poi" (Maffioli, M., Tomassini, L., 2020).

Per la selezione del materiale da somministrare al vaglio dell'Intelligenza Artificiale è stato escluso il Registro Alfabetico in quanto si caratterizzava come più scarno nella quantità e varietà dei vocaboli. Viceversa, il registro cronologico risultava caratterizzato da un maggior numero di parole organizzate in colonne: numero d'ordine/ cognome e nome/ indirizzo/ soggetto/ dimensioni/ numero delle copie/ importo/ Esatto (vd. Figg. 1 e 2).
Per la realizzazione dei modelli è stato necessario procedere con una modalità di trascrizione mista, quindi è stato sperimentato prima l'utilizzo del supermodello "Text Titan I ter" e successivamente "XX Century Italian Handwritting", per poter valutare quale dei due potesse essere più performante in relazione al materiale da trascrivere. In entrambi i casi la trascrizione è risultata essere estremamente inappropriata quanto lacunosa, se si fa eccezione per alcuni gruppi di numerazioni riconosciuti generalmente in modo corretto. Quindi, con una precisa correzione manuale dei documenti, sono stati allenati diversi modelli, le cui informazioni sono state sintetizzate in tabella (vd. Tab. 1).
I due modelli di *recognition text* di partenza sono stati: "Registro manoscritto italiano fine XIX sec." e "Registro manoscritto italiano 1942-2002", i quali, per comodità, verranno definiti rispettivamente "XIX" e "1942-2002".
Lo scarto necessario nella quantità di pagine sottoposte all'analisi dell'Intelligenza Artificiale è dovuto, come già anticipato, alla differenza di mani ma anche di quantità di grafie diverse entro uno stesso quantitativo di pagine. In particolare, per il modello "1942-2002" si è cercato di implementare una maggior varietà di mani proprio perché risultava impossibile la creazione di tanti modelli per tante scritture differenti, ognuna contenuta entro pochissime pagine, se non addirittura entro qualche riga di una medesima.

Nome modello	Tipologia modello	Cronologia	*Base model*	Pagine di *training*	Pagine di validazione	Numero di parole	CER
Registro italiano XX sec.	*Table*	1942-2002	-	76	3	39801	28,63%
Registro manoscritto italiano 1942-2002	*Text*	1942-2002	XX century italian handwriting 1.0	72	7	37600	9,83%
Registro manoscritto italiano fine XIX sec.	*Text*	1893-1895	Italian - 20th century Minutes of Mediobanca's Board of Directors & Exec. Committee	45	5	15691	5,13%
Registro manoscritto italiano fine XIX sec. 2	*Text*	1893-1896	Registro manoscritto italiano fine XIX sec.	63	7	22407	7,51%
Registro manoscritto italiano primo quarto XX sec.	*Text*	1895-1913	Registro manoscritto italiano fine XIX sec. 2	72	8	47433	6,78%
Registro manoscritto italiano XX sec.	*Text*	1908-1921	Registro manoscritto italiano 1942-2002	63	7	40740	3,58%

Tabella 1 – Sintesi della sperimentazione dei modelli

Conseguentemente il CER relativo a ciascuno dei due modelli di partenza è differente di circa 4 punti percentuali: "XIX" riporta una percentuale di errore del 5,13, su 15.691 parole, mentre "1942-2002" una percentuale del 9,83 su 37.600 parole. Per la base di integrazione di quest'ultimo modello, la scelta di partenza è stata "XX Century Italian Handwriting 1.0". Risultando buona, ma non eccellente, la trascrizione restituita sulle pagine di prova, per la base di integrazione del modello "XIX" è stato selezionato un altro modello italiano dalla calligrafia piuttosto simile ossia "Italian - 20th century Minutes of Mediobanca's Board of Directors & Exec. Committee".
Pur risultando piuttosto soddisfacente, anche se piuttosto ricco di imperfezioni soprattutto nella restituzione dei nomi propri, è stato fatto un ulteriore tentativo di creazione modelli per una trascrizione più efficace: denonimati "Registro manoscritto italiano fine XIX sec. 2" e "Registro manoscritto italiano primo quarto XX sec." hanno restituito risultati differenti circa una corretta trascrizione. In entrambi i casi l'idea all'origine è stata quella di aumentare il corpus di pagine in modo da ottenere un vocabolario possibilmente più ampio, oltre a testare l'utilizzo di un *base model* precedente, per valutarne la performance: nel primo modello il CER è passato da 5.13% a 7,51% su 22407 parole anziché 15.961, nel secondo il CER è passato da 5,13% a 6,78% su 47433 parole anziché 15.961[2].
Un ulteriore test di valutazione per un modello migliore, sulla grafia di fine XIX-inizio XX sec., è stato fatto invece selezionando come *base model* proprio "1942-2002", dunque un modello sviluppato in precedenza sullo stesso documento ma con pagine differenti. Il risultato è stato un CER del 3,58% con 40740 parole.

3. CRITICITÀ

La sperimentazione di Transkribus per la trascrizione dei registri ha restituito differenti criticità, soprattutto per il riconoscimento del *layout*. L'ideazione di un *table model*, ha dato certamente dei risultati positivi sulle pagine dello stesso registro sulle quali il software è stato allenato, tuttavia non ha sempre restituito dei risultati soddisfacenti circa l'applicazione all'altro, con sviluppo grafico differente all'interno delle pagine (vd. Fig. 2). Come si può ben notare nella maggioranza delle immagini, pare che l'assenza di informazioni scritte nella parte centrale del registro, ha portato in moltissimi casi al non riconoscimento dello stesso, come continuo in maniera orizzontale. Questa problematica sulla regolarità del testo risulta in linea con quanto evidenziato anche nel progetto dell'archivio storico Mediobanca, nel quale viene evidenziata una difficoltà di lettura e riconoscimento di una tabella, come elemento a sé stante, rispetto al contenuto scritto dentro o fuori dalla stessa (Carboni & Molino Lova, 2025). Di conseguenza si è resa necessaria la correzione manuale delle tabelle su tutto il registro alfabetico.
Oltre a ciò una problematica piuttosto rilevante è stata quella di un mancato riconoscimento da parte dei modelli di *analyses layout* delle linee: le pagine difatti non sono caratterizzate da un flusso di testo continuo ma da molti elementi disseminati, quali lettere, numeri e virgolette. Queste ultime, in particolar modo, sono risultate critiche per una corretta individuazione da parte dell'IA: autonome rispetto alle altre parole, piccole e spesso anche sbiadite nella qualità degli inchiostri.
Questo problema, in minima parte risolvibile attraverso le impostazioni avanzate di riconoscimento del *layout*, ha causato in maniera notevole un allungamento dei tempi soprattutto per la parte di correzione manuale in seguito all'applicazione del modello "Universal Lines".
Un ulteriore punto di criticità è relativo alla numerazione presente nei documenti: tale numerazione si divide in: semplice e alfanumerica; quest'ultima è composta da un numero a quattro/cinque cifre e la lettera B. Tale lettera infatti risulta spesso scritta in posizione apicale, talvolta distante dal numero di riferimento e anche piuttosto piccola: il risultato che ne deriva è quello di una casuale distribuzione della lettera B sopra ogni tipo di numerazione, segno che l'intelligenza artificiale non pare in grado di decifrarla, discostandosi in maniera così prorompente rispetto all'andamento orizzontale delle cifre.
A questo proposito occorre segnalare un altro importante problema nella trascrizione del documento dato dalle abbreviazioni delle parole: queste, nella maggior parte dei casi, prevedevano che il termine

[2] I test dei modelli, basati su *base model* creati in precedenza con un nucleo di pagine identiche, sono stati eseguiti per valutare effettivamente il grado di performance di nuovi modelli, partendo da modelli precedenti. Nonostante il sito di Transkribus riporti la scarsa efficacia dei modelli che partano da altri con nucleo di pagine identico, l'esperienza di Mediobanca (Carboni & Molino Lova, 2025) con questo tipo di operazione è invece risultata buona. Si veda: https://www.transkribus.org/blog/using-base-model-when-training-a-custom-model (cons. 09/04/2026).

dell'abbreviazione fosse posto anch'esso in apice, dunque senza seguire la linea orizzontale della parola di riferimento. A causa di un'abbreviazione non uniforme, per cui il termine "riproduzione" risulta essere abbreviato come "rip.ne"/ "ripr.ne"/ "riprod.ne"/ "riproduz.ne", si è optato per la decisione di non sciogliere i vocaboli, data anche dalla cronologia piuttosto recente dei documenti.
Infine, parole a caratteri molto grandi o disposte non orizzontalmente, hanno causato anch'esse criticità a livello di riconoscimento da parte dell'IA.

4. RISULTATI E OBIETTIVI

Attualmente il progetto risulta essere ancora in corso e si procede alla trascrizione con l'applicazione dei modelli, seguendo sia un criterio che individua CER con punteggi migliori, sia un accoppiamento in termini cronologici: "Registro manoscritto italiano primo quarto XX sec.", ad esempio" pur avendo 1,5 punti percentuali in più rispetto a "XIX", risulta riportare risultati migliori sui primi anni del XIX sec., essendo stati ovviamente inclusi nel *training*. Rispetto alla seconda metà del XX sec. non si tratta di grandi differenze nelle grafie, tuttavia bisogna ammettere che la differenza in termini qualitativi di restituzione risulta evidente. Dall'analisi dei risultati dei modelli si può certamente confermare quanto presente sul sito di Transkribus: la diversificazione del *base model* porta sicuramente a risultati migliori nell'abbassamento del CER.
Per la seconda metà del XX secolo si potrebbero fare ulteriori indagini, tuttavia la diversità delle scritture entro poche pagine non permette di avere ampio margine di successo e pertanto sarà necessaria un'importante correzione manuale.
Ad ora risulta completa ed esportata in formato Excel le pagine del registro cronologico relative all'arco temporale 1893-1920: su queste si è proceduto con l'inserimento di una colonna aggiuntiva per l'integrazione manuale della striscia identificativa della lastra corrispondente. Come già anticipato, le lastre negative conservano ancora oggi il proprio numero d'inventario originale, tuttavia alcune ne sono sprovviste e altre ancora invece hanno una numerazione uguale e duplicata. Inoltre, a causa delle difficoltà di ricerca e abbinamento del materiale originale con i files creati dalle aziende di digitalizzazione, è stato necessario, come priorità della fondazione, predisporre un nuovo documento che riporti anche la denominazione attuale delle lastre.
Entro quest'ottica, la digitalizzazione degli archivi Alinari non si giustifica unicamente con uno scopo conservativo dei materiali, ma fa di questa operazione un nuovo punto di partenza per ulteriori linee di sviluppo: dall'elaborazione dei dati con software come Microsoft Power BI a collegamenti ipertestuali attraverso l'inserimento di link rimandanti alle immagini pubblicate sul catalogo Alinari, visionabile direttamente online. Tale operazione, una volta applicata, riassumerebbe entro le proprie finalità due dei quattro obiettivi individuati entro le *Linee guida per la digitalizzazione del patrimonio culturale* (Ministero della Cultura, 2022).
Da questo progetto risulta evidente anche una moltiplicazione delle possibilità di ricerca: solo pr fare qualche esempio, si potrebbero indagare le variazioni dei prezzi delle fotografie negli anni, la percentuale di richiesta dei differenti formati o la variazione dei soggetti richiesti. Tutto ciò propone una modalità differente di studio dei documenti, come illustrato in studi che hanno concentrato la propria attenzione su Transkribus (Muehlberger, G., et al., 2019; Gallegati, 2025).
Circa le modalità di ricerca successive all'elaborazione finale della trascrizione, sarà quindi possibile mettere in risalto le informazioni all'interno dei file, i quali nella semplice digitalizzazione risultano comunque di difficile gestione[3].
Infine, una volta terminata la fase di trascrizione dei registri, il contenuto verrà reso fruibile e pubblico agli studiosi, seguendo modalità ancora da definire: non si esclude la pubblicazione sul sito della piattaforma, la quale permetterebbe anche una metadatazione inclusiva delle stringhe identificative attuali rispetto ai numeri d'inventario storici.

RINGRAZIAMENTI

L'iniziativa si inserisce all'interno del progetto "Dicolab. Cultura al digitale" - sub-investimento 1.1.6 "Formazione e Miglioramento delle competenze digitali" del PNRR M1C3 ed è parte di una borsa di ricerca dell'Università di Pisa con la Fondazione Alinari per la Fotografia. Per l'avvio del progetto è stata istituita

[3] Quello dell'opacità dell'informazione data dalla pura digitalizzazione dei manoscritti era un tema già sottolineato da Edwards (2007).

una collaborazione tra il Dipartimento di Civiltà e Forme del Sapere con il Dipartimento di Informatica Umanistica, dal quale ne è nato un tirocinio formativo e una tesi della dott.ssa Alice Mannaioni.
Si ringraziano i proff. Cosimo Chiarelli, Enrica Salvatori e Vittore Casarosa; la Fondazione Alinari per la Fotografia, in particolar modo Claudia Baroncini, Francesca Bongioanni e Rita Scartoni; gli archivisti della biblioteca dell'Università Cattolica del Sacro Cuore di Milano Matteo Ruggeri e Lorenzo Babini per i preziosi suggerimenti.

BIBLIOGRAFIA

Carboni, S., Molino Lova, T., (2025), L'utilizzo dell'Handwritten Text Recognition nei sistemi di consultazione degli archivi: il caso dell'archivio storico Mediobanca "Vincenzo Maranghi" e Transkribus. *Il Mondo degli Archivi*. 1 https://www.ilmondodegliarchivi.org/wp-content/uploads/2025/05/Lutilizzo-dellHandwritten_v2.pdf

Dunley, R., (2018). *Machines Reading the Archive: Handwritten Text Recognition Software.*

Edwards, J., (2007). *Easily adaptable handwriting recognition in historical manuscripts*. PhD dissertation, University of California, Berkeley, CA. http://www.eecs.berkeley.edu/Pubs/TechRpts/2007/EECS-2007-76.html

Gallegati, S., (2025). Transkribus e la trascrizione automatica di manoscritti. Un caso studio sulle carte di Vittorio Alfieri. *Digitalia*. 2-2025. DOI: 10.36181/digitalia-00147

Maffioli, M., Tomassini, L., (2020). *Report Finale: il patrimonio fotografico Alinari*. Società Italiana per lo Studio della Fotografia, SISF. https://www.sisf.eu/wp-content/uploads/2021/04/Report-scientifico-finale-Accordo-Regione-Toscana-SISF.pdf

Ministero della Cultura (2022). *Linee guida per la digitalizzazione del patrimonio culturale*. https://docs.italia.it/italia/icdp/icdp-pnd-digitalizzazione-docs/it/v1.0-giugno-2022/index.html

Muehlberger, G., Seaward, L., Terras, M., et al., (2019). Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study. *Journal of Documentation*, 75(5), 954–976.

Quintavalle, A., C., (2003). *Gli Alinari*, Alinari.

Scartoni, R. (in corso di stampa). Commercio e produzioni nelle fotografie Alinari. In Giuntini A., Tomassini L. (a cura di) *L'associazionismo fiorentino e toscano a 150 anni dalla nascita dell'Associazione Macellai di Firenze e provincia.*

Sesti, E., (2018). La fotografia d'arte dei Fratelli Alinari dal 1852 a oggi e gli archivi fotografici entrati nelle collezioni Alinari al tempo di Vittorio Cini. In Bassanello, M., Guerriero, S. & Turetta, I. (a cura di). *Per un archivio fotografico dell'arte italiana. Vittorio Cini e la Fratelli Alinari* [atti della giornata di studi, 2018, 5 giugno], Venezia, Fondazione Cini, Marsilio. 38-55.

Settimelli, W., Zevi, F., (1977). *Gli Alinari fotografi a Firenze*, Alinari

Tomassini, L., (2019). Il patrimonio fotografico Alinari: excursus storico e questioni attuali. *Rivista di studi della fotografia* n. 10, 100-120. https://doi.org/10.36253/RSF-12248

Biblioteche filosofiche private e memoria culturale:
http://picus.unica.it/

Andrea Lamberti[1], Giovanna Granata[2]
UNIVERSITÀ DEGLI STUDI DI CAGLIARI, ITALY -
[1] andrea.lamberti@unica.it; [2] giovanna.granata@unica.it

ABSTRACT (ITALIANO)

La creazione di spazi virtuali che accolgano in maniera strutturata dati e materiali è nell'attuale era digitale un passaggio essenziale per la conservazione e la ricostruzione della nostra memoria culturale. Gli straordinari mezzi impiegati negli ultimi anni per preservare la documentazione, attraverso gli strumenti e i processi di digitalizzazione, non sempre hanno coinciso con una maggiore accessibilità. La sovrabbondanza di dati, riversati *online* in modo disorganizzato, ha di fatto costituito un limite alla loro fruizione, soprattutto laddove non sono stati previsti criteri e parametri con i quali ordinare e ricercare le informazioni.
Nei primi anni Duemila, il progetto "Biblioteche filosofiche private di età moderna e contemporanea", che ha portato alla realizzazione del sito web http://picus.unica.it/, ha rappresentato una vera e propria novità nel campo dell'umanistica digitale, con lo scopo di conservare e diffondere parti significative della nostra storia culturale, attraverso una ricerca al confine tra biblioteconomia e filosofia. La proposta progettuale ha restituito visibilità a un patrimonio altrimenti destinato ad essere trascurato: i cataloghi e gli inventari delle raccolte librarie private con un impatto significativo nella storia della filosofia e della cultura europea.
Nel 2026, il sito web è stato oggetto di un lavoro profondo di rinnovamento che ha inteso sfruttare al massimo le possibilità offerte dalla conversione digitale e dalla sua pervasività, limitandone i rischi di dispersione delle informazioni e ampliando le possibilità di condivisione della conoscenza.
L'aggiornamento e il potenziamento del sito hanno portato al recupero di una notevole quantità di dati aperti, in accordo con la politica dei principi FAIR. In tal modo la banca dati "Picus" è stata resa uno strumento con un grande potenziale di uso nel campo della ricerca e della didattica, nella prospettiva di un recupero attivo della memoria culturale da parte del vasto pubblico.
Parole chiave: memoria culturale; biblioteche digitali; storia; filosofia.

ABSTRACT (ENGLISH)

Private philosophical libraries and cultural memory: http://picus.unica.it/.
In the digital age, it is essential to create virtual spaces that store data and materials in a structured way in order to preserve and reconstruct our cultural memory. Although the extraordinary digitisation efforts employed in recent years have been effective in preserving documentation, they have not necessarily resulted in greater accessibility. In fact, the vast amount of disorganised data uploaded online limits its use, especially when there are no criteria or parameters for ordering and searching information.
In the early 2000s, the 'Private Philosophical Libraries of the Modern and Contemporary Age' project led to the creation of the "Picus" website (http://picus.unica.it/) and represented a real innovation in the field of digital humanities. The project aimed to preserve and disseminate important aspects of our cultural heritage by combining library science and philosophy. The project has brought to light a heritage that would otherwise have been forgotten: the catalogues and inventories of private book collections that have significantly influenced the history of European philosophy and culture.
In 2026, the website was extensively renovated with the aim of maximising the potential offered by digital conversion. This upgrade expanded the scope for knowledge sharing, while limiting the risks of information dispersion. Updating and enhancing "Picus" database and its functionalities led to the extraction of a considerable amount of open data in accordance with the FAIR Principles Policy. As a result, the website has become a valuable research and teaching tool for the general public, actively recovering cultural memory.
Keywords: cultural memory; digital libraries; history; philosophy.

1. IL PROGETTO DI RICERCA TRA BIBLIOTECONOMIA E FILOSOFIA

Il progetto "Biblioteche filosofiche private di età moderna e contemporanea" è stato sviluppato a partire dal 2007, grazie a una collaborazione tra l'Università degli Studi di Cagliari e la Scuola Normale di Pisa, che ha portato all'allestimento della banca dati online "Picus" (http://picus.unica.it/), coinvolgendo numerosi specialisti ed esperti internazionali, e riconosciuti istituti culturali, tra i quali la Fondazione Biblioteca
Benedetto Croce, l'Istituto italiano per gli studi storici, e l'Istituto per il Lessico Intellettuale Europeo e Storia delle idee (ILIESI).
Il progetto ha inteso portare l'attenzione degli studiosi sulle biblioteche private quali 'laboratori creativi', in grado di fare luce sulle dinamiche dei processi di ideazione e di scrittura, fondamentali per poter studiare nel dettaglio le relazioni culturali che caratterizzano una determinata epoca storica (Ragghianti, Savorelli 2012). Il progetto è stato in grado di restituire e di rimettere in circolazione in maniera unitaria e strutturata una enorme quantità di dati finora slegati, perché dispersi in biblioteche, in archivi fisici o in altre piattaforme online.
Studi consolidati hanno attestato che l'indagine sulle collezioni librarie è essenziale per un'analisi dei testi come reti complesse e stratificate, perché consente di ricostruire l'insieme delle tracce e degli elementi che spiegano la genesi dei testi, fornendo informazioni cruciali sui contesti storico-culturali nei quali hanno preso forma e sui quali hanno influito (Canone 1993). I risultati prodotti dalla ricerca sui fondi librari privati, una volta riversati online in http://picus.unica.it/, hanno reso possibile la creazione di una base orientativa, formata dagli studi generali sul tema, e di una banca dati, nella quale figurano i cataloghi, gli inventari e i documenti digitalizzati che descrivono le biblioteche private.
Tale struttura di riferimento oltre a rendere universalmente accessibile una grossa quantità di informazioni sugli autori e sui profili, noti e meno noti, che hanno segnato la nostra storia intellettuale, costituisce un raffinato dispositivo di analisi delle raccolte librarie capace di assolvere a diversi scopi, con ricadute in diversi campi di ricerca. Sul piano biblioteconomico si tratta di uno strumento che mette nelle condizioni di avere contezza dei passaggi e delle trasformazioni subite dalle biblioteche private in rapporto alla genesi e allo sviluppo dei sistemi bibliotecari istituzionali. Sul piano della storia della filosofia e della cultura la disponibilità di cataloghi di biblioteche private fornisce elementi per ricostruire le letture e gli interessi propri della formazione di un autore, che rientrano tra le fonti esplicite e implicite delle sue opere; permette inoltre di seguire la circolazione e la fortuna di determinati testi, mettendo nelle condizioni di ricostruire il modo nel quale vengono letti, recepiti e interpretati (Granata 2023). Più in generale, le biblioteche private rappresentano punti di osservazione privilegiati sui rapporti reciproci tra le differenti discipline e sulle forme di organizzazione e di trasmissione del sapere che mutano nel tempo. Offrono la possibilità di rileggere in maniera originale consolidate categorie storiografiche e canoni culturali, ponendo in nuova luce i caratteri delle tradizioni che hanno segnato certi periodi storici o particolari spazi geografici (Ragghianti, Savorelli 2014).
Punto di partenza ideale del progetto è stato Pico della Mirandola, dal quale deriva il nome "Picus" presente nell'URL del sito web (http://picus.unica.it/). La biblioteca di Pico della Mirandola, nel Quattrocento, alle soglie dell'età moderna, rappresenta uno dei primissimi esempi, se non il primo, di una collezione libraria che, prima di essere frutto di passione erudita, è intesa da Pico come strumento di ricerca e di indagine, nel lavoro materiale di costruzione delle opere e del suo pensiero.
La biblioteca di Pico della Mirandola è stata tra le prime a essere pubblicata alla fine dell'Ottocento: prima nel 1897, sulla base di un catalogo di vendita (Calori Cesis 1897, pp. 31 -76), poi nel 1936 da Pearl Kibre, sulla base dell'inventario manoscritto redatto da Pico (Kibre 1936). Negli stessi anni uscivano il catalogo della biblioteca di Michel de Montaigne (Villey 1908) e di quella di Galileo Galilei (Favaro 1886 e 1887).
Attraverso la comparsa di queste biblioteche filosofiche private prendeva così corpo un filone di studi inaugurato da Wilhelm Dilthey che, nel 1889, aveva posto grande attenzione alla rilevanza delle carte, delle lettere e dei libri posseduti per la comprensione storiografica dei filosofi e delle epoche nelle quali hanno operato, a partire dall'idea di una 'storia della filosofia' che si fa 'storia della cultura' (Dilthey 1889; Orsucci 2010).
L'uscita del catalogo della biblioteca di Pico alla metà degli anni Trenta assume inoltre un significato del tutto particolare perché contrassegna una importante ripresa delle ricerche sul Rinascimento. L'anno dopo lo studio di Kibre, nel 1937, Eugenio Garin pubblica un volume consacrato alla vita e alla dottrina di Pico della Mirandola, nel quale, tra le altre cose, corregge e discute il lavoro di Ferdinando Calori Cesis

sull'inventario dei libri posseduti dall'autore (Garin 1937, pp. 106-116). Gli studi di Kibre e di Garin furono immediatamente posti in risalto in una recensione di Paul Oscar Kristeller sul «Giornale critico di filosofia italiana» (Kristeller 1938).
Kristeller esamina accuratamente lo studio di Garin, ne critica l'interpretazione della dottrina di Pico, troppo schiacciata ai suoi occhi su quella di uno «scolastico eretico» (Kristeller 1938, pp. 375-376). L'accostamento con lo studio di Kirbe induce Kristeller a rilevare la centralità degli antichi inventari delle biblioteche in seno al lavoro storiografico, strumenti essenziali per rendere conto delle categorie filosofiche e delle continuità e discontinuità con le differenti tradizioni culturali. Siamo all'inizio della riflessione sul concetto di Rinascimento e sul suo rapporto con l'eredità scolastica e aristotelica che segna la storiografia italiana nel corso del Novecento, e che, è utile qui ricordare, nasce insieme all'interesse per le biblioteche filosofiche private come strumento per riportare gli autori a un determinato orizzonte di idee e di pensiero. Per tutte queste ragioni, la biblioteca di Pico vale perciò come caso paradigmatico della centralità dello studio delle raccolte librarie per la storia della cultura moderna e contemporanea. È il concreto avvio di un insieme di pratiche che vedono intrecciarsi circolazione dei libri e circolazione delle idee, trasmissioni dei patrimoni librari e costruzioni dei canoni filosofici e culturali. Gli inventari, i cataloghi e i documenti progressivamente archiviati e digitalizzati in "Picus" non fanno che tenere testimonianza e traccia di queste pratiche in un periodo di lungo corso, dal Rinascimento alla contemporaneità.
I documenti resi accessibili nel sito sono di molteplice genere. Si ritrovano disponibili: a) gli elenchi di libri compilati dagli stessi possessori per uso personale, presenti negli archivi privati, con l'insieme delle loro carte personali; b) gli atti notarili con gli inventari dei beni tra i quali i libri del possessore; c) i cataloghi di vendita all'asta di collezioni librarie private; d) i cataloghi di fondi privati confluiti nelle biblioteche istituzionali, per via di donazione o di altra forma di acquisizione (Usalla 2019).
Questi materiali sono stati digitalizzati per lo più in formato pdf che è possibile scorrere per pagine o utilizzando la mediazione di una struttura ipertestuale. Le differenti fonti catalografiche sono associate a una scheda descrittiva con notizie sul possessore e sulla sua raccolta libraria.

2. IL NUOVO SITO WEB E LE SUE FUNZIONALITÀ

A partire dal 2026, la banca dati online di "Picus" è stata al centro di un processo di aggiornamento e di revisione, reso possibile grazie a un finanziamento della Fondazione di Sardegna. Il sito web è stato adeguato ad avanzati standard di sicurezza per evitare il processo di obsolescenza. Inoltre, mettendo a frutto le nuove possibilità offerte dall'avanzamento della tecnologia nel campo delle *Digital Humanities*, è stato rivisto e potenziato per raggiungere migliori livelli di sfruttamento delle informazioni, prefigurando ulteriori sviluppi delle modalità di ricerca, nel solco delle prospettive aperte dall'uso dell'IA.
L'obiettivo generale è stato di creare un ambiente virtuale in grado di promuovere l'uso del sito web su vasta scala. Questo risultato è stato perseguito attraverso l'attuazione di due azioni particolari:

- la riorganizzazione grafica dei contenuti, pensati in funzione dell'esperienza dell'utente, in modo da renderli facilmente accessibili e fruibili;
- il potenziamento delle funzionalità di ricerca per passare ad una interrogazione trasversale e strutturata dei dati nelle schede e nel materiale digitalizzato.

Dal punto di vista grafico, la scelta è stata quella di dare maggiore spazio possibile alla ricerca e alla esplorazione delle informazioni e dei contenuti. Questo risulta particolarmente evidente sia nei modi nei quali è stata ridisegnata la pagina di accesso alle schede delle biblioteche private ('Indice delle biblioteche'), sia nei modi nei quali è stata ripensata l'organizzazione dei contenuti nelle schede descrittive. Nella soluzione precedente, la pagina 'Indice' presentava la lista dei nomi degli autori o possessori in ordine alfabetico, con la possibilità di accedere alle schede per tramite del nome. Nel nuovo sito i singoli autori o possessori sono presentati per mezzo di card, giustapposte in ordine alfabetico, che contengono il nome, la data e il luogo di nascita e di morte, e alcuni tag informativi relativi al profilo e alle aree di interesse del possessore (per es. Rinascimento, Giansenismo), e tag sulla sua biblioteca (per es. catalogo d'asta; inventario manoscritto). Questi tag si ritrovano poi nelle schede descrittive alle quali si accede tramite le differenti card. Nelle schede di presentazione i contenuti sono stati organizzati in due sezioni: la prima di 'Informazioni' generali sul profilo del possessore (biografia, opere, bibliografia essenziale); l'altra riguarda invece propriamente la 'Biblioteca', con le notizie specifiche sulla sua raccolta

libraria (consistenza, eventuali dispersioni, attuale localizzazione, studi in merito), e con la possibilità di consultare il catalogo e gli altri documenti disponibili in formato digitale. Ogni scheda, a fondo pagina, rimanda ad altre tre schede correlate.
Già solo questa nuova disposizione grafica rappresenta un notevole incremento in termini di possibilità di esplorare le collezioni mettendo in collegamento informazioni di natura diversa, suggerendo nessi e percorsi che prima non erano visibili. Soprattutto stimola l'utente a costruire un dialogo tra autori, testi e raccolte librarie, mentre naviga nel sito.
Per quanto riguarda il potenziamento delle funzionalità operative di ricerca, la maggiore novità è stata l'introduzione di una navigazione "a faccette" attraverso l'uso di filtri dinamici così da restringere i risultati per data, categoria, autore o tag. Per realizzare questo tipo di navigazione, si è provveduto ad avviare una indicizzazione semantica dei titoli, del corpo del testo delle pagine e del materiale digitalizzato, per consentire un accesso più diretto ai contenuti attraverso l'interrogazione per parole. Questo lavoro, però, non è stato affatto conclusivo, soprattutto per quanto concerne il materiale digitalizzato.
Gran parte dei documenti presenti nel sito è stato acquisito come scansione di immagine e non come conversione di un testo. Questo genere di risorse digitali, a differenza di quelle testuali, pone notevoli problematiche sul piano del recupero dei metadati, cioè di informazioni strutturate di natura 'indicale' sulla base delle quali è possibile costruire i filtri dinamici di ricerca. La soluzione perseguita è stata quella di estrarre testo dalle immagini digitali attraverso l'uso di tecnologie OCR (Optical Character Recognition). Queste tecnologie hanno sono applicabili esclusivamente alle immagini dei testi a stampa e non a quelle dei manoscritti. Inoltre, il riconoscimento dei caratteri diventa sempre meno efficace e preciso nella misura laddove i testi sono più antichi e i caratteri utilizzati meno "mappati". Nonostante ciò, i risultati raggiunti sono stati però rilevanti, avendo portato a rendere ricercabile una buona parte del materiale digitalizzato che prima era possibile solo scorrere per pagine.
La riorganizzazione dei contenuti, la loro strutturazione con l'introduzione di tag, il rafforzamento delle funzionalità di ricerca con l'introduzione di filtri dinamici e l'impiego di tecnologie OCR per la lettura dei file di immagini di testo hanno radicalmente trasformato http://picus.unica.it/, mettendo a disposizione della comunità scientifica internazionale e del largo pubblico un servizio di facile consultazione e dall'alto potenziale informativo.

3. LE PROSPETTIVE DI SVILUPPO

In un tale contesto è possibile intravedere opportunità ulteriori di gestione dei dati nel sito web, anche per mezzo dell'uso dell'IA, combinato con gli sviluppi nel campo della HCR (Hi-Definition Recognition) e di nuovi applicativi come Traskribus, specializzati nella lettura di manoscritti. Si tratta di mettere a punto una seconda fase con l'obiettivo di sfruttare una serie di soluzioni tecnologiche innovative per una 'risemantizzazione' più approfondita e massiva delle informazioni presenti nei documenti digitalizzati, siano essi immagini di testo a stampa oppure immagini di carte manoscritte. Tale obiettivo si intende inoltre perseguire per mezzo dell'uso di algoritmi di Machine learning, di modo che, nel passaggio dall'immagine al testo, le stringhe di testo siano 'interpretate', lette e rese ricercabili per quello che sono, cioè come citazioni bibliografiche.
Lo sviluppo di "Picus" nella direzione di questa puntuale "risemantizzazione" delle fonti digitalizzate metterebbe nelle condizioni di recuperare una quantità ingente di dati strutturati e aperti. Questo è un presupposto essenziale per portare il sito web verso la costruzione di una piattaforma, adatta a ospitare, gestire e sviluppare altre iniziative analoghe nel campo della storia culturale, così da farne un vero e proprio laboratorio per lo sviluppo di soluzioni progettuali da impiegare nel campo dell'umanistica digitale, con l'obiettivo di un recupero attivo della memoria.

BIBLIOGRAFIA

Calori Cesis F. (1897). Memorie storiche della Mirandola. XI. Mirandola

Canone E. (ed.). (1993). Bibliothecae selectae. Da Cusano a Lopardi. Firenze: Olschki Dilthey W. (1889), Archive für Literatur. *Deutsche Rundschau*, 58, 360–375

Favaro A. (1886). La libreria di Galileo Galilei descritta e illustrata. *Bullettino di bibliografia e di storia delle scienze matematiche e fisiche*, 19, 219–293

Favaro A. (1887). Appendice prima alla libreria di Galileo Galilei descritta e illustrata. *Bullettino di bibliografia e di storia delle scienze matematiche e fisiche*, 20, 372–376

Garin E. (1937). Giovanni Pico della Mirandola: vita e dottrina. Firenze: Le Monnier

Granata, G. (2023) Philosophers' Private Libraries (1600–1800). In Rindert Jagersma, Helwi Blom, Evelien Chayes, Ann-Marie Hansen (Eds.), Private Libraries and Their Documentation, 1665–1830. Studying and Interpreting Sources (pp. 389-402). Leiden: Brill

Kibre P. (1966). The library of Pico della Mirandola. New York: AMS Kristeller P.O. (1938). *Giornale critico della filosofia italiana*, XIX, 378-381

Ragghianti R., Savorelli A. (Eds.). (2014). Biblioteche filosofiche private. Strumenti e prospettive di ricerca. Pisa: Edizioni della Normale

Villey P. (1908). Les sources & l'évolution des Essais de Montaigne. Paris: Hachette

Ragghianti R., Savorelli A. (2012). Biblioteche filosofiche private: strumenti di lavoro, documenti e contesti. In Rusconi R. (Ed.), Il libro antico tra catalogo storico e catalogazione elettronica, (pp. 109-132). Roma: Accademia Nazionale dei Lincei

Orsucci A. (2010). Il problema delle biblioteche filosofiche nella storiografia novecentesca: qualche breve divagazione. In Crasta F. M. (Ed.), (2010). Biblioteche filosofiche private in età moderna e contemporanea. Atti del convegno Cagliari, 21-23 aprile 2009 (pp. 265-275). Firenze: Le Lettere

Usalla L. (2019). Biblioteche dei filosofi. Biblioteche filosofiche private in età moderna e contemporanea. In Granata, G. (Ed.), Biblioteche e saperi. Circolazione di libri e di idee tra età moderna e contemporanea (pp. 3-15). Roma: Edizioni di Storia e Letteratura

Beyond Archives: A Three-Stage Workflow for Relational Digital Libraries of Written and Oral Memory Sources

GIULIA LEMBO

[1]University of Naples "L'Orientale", Italy – g.lembo2@unior.it

ABSTRACT (ENGLISH)

The project *Beyond Archives* aims to develop a relational digital library for memory sources in conflicted historical settings, using the post-war Julian–Dalmatian exodus to Naples as a laboratory. It connects institutional records and oral testimonies in one environment while keeping differences in provenance, access conditions, and narrative contexts readable. The approach adopts a three-stage workflow applied in parallel to written and oral streams: digitization produces stable surrogates and minimal metadata; formalization creates aligned, inspectable structural and textual layers; publication reassembles these layers into traceable access paths across facsimiles, transcriptions, and TEI encoding. We present early results primarily from the written stream, where semantic layout segmentation and ATR are trained in eScriptorium/Kraken and exported as PAGE/ALTO to preserve coordinates for correction and TEI-oriented modelling. Successive segmentation iterations show stable behavior on the main text stream, while stamps, marginalia, and signatures remain the main sources of variance. Across the ATR trajectory, early cycles yield sizeable gains over the public baseline, and the strongest operational checkpoint is currently ATRModel-5. In parallel, the project has reached a prototype stage in the TEI pipeline and has activated the oral stream through a first exploratory full-pass ASR. Finally, we discuss implications for targeted ground-truth expansion, out-of-sample testing on new lots, and planned releases of checkpoints and paradata to support reuse.
Keywords: automatic text recognition (ATR); layout segmentation; eScriptorium; digital libraries; Italian typescripts

ABSTRACT (ITALIANO)

Beyond Archives: Un workflow in tre fasi per digital library relazionali di fonti memoriali scritte e orali .
Il progetto *Beyond Archives* mira a sviluppare una digital library relazionale per fonti della memoria in contesti storici conflittuali, usando l'esodo giuliano-dalmata verso Napoli nel secondo dopoguerra come laboratorio. Il progetto connette in un unico ambiente documenti istituzionali e testimonianze orali, mantenendo leggibili differenze di provenienza, condizioni di accesso e contesti narrativi. L'approccio adotta un workflow in tre fasi, applicato in parallelo ai flussi scritto e orale: la digitalizzazione produce surrogati stabili e metadati minimi; la formalizzazione crea strati strutturali e testuali allineati e ispezionabili; la pubblicazione ricompone questi strati in percorsi di accesso tracciabili tra facsimili, trascrizioni e codifica TEI. Si presentano i risultati preliminari soprattutto sul flusso scritto, in cui la segmentazione semantica del layout e l'ATR sono addestrate in eScriptorium/Kraken ed esportate in PAGE/ALTO per preservare le coordinate utili alla correzione e alla modellazione orientata alla codifica TEI. Le successive iterazioni della segmentazione mostrano una buona tenuta sul testo principale, mentre timbri, marginalia e firme restano le principali fonti di variabilità. Nella traiettoria ATR, i primi cicli producono miglioramenti rilevanti rispetto al baseline pubblico, e il checkpoint operativo più solido è attualmente ATRModel-5. In parallelo, il progetto ha raggiunto una fase prototipale nella pipeline TEI e ha attivato il ramo orale attraverso un primo full-pass ASR esplorativo. Infine, si discutono le implicazioni per un'espansione mirata del ground truth, test out-of-sample su nuovi lotti e rilasci pianificati di checkpoint e paradata per garantirne il riuso.
Parole chiave: riconoscimento automatico del testo (ATR); segmentazione layout; eScriptorium; digital libraries; dattiloscritti italiani

1. INTRODUCTION

Archives and memory sources are never neutral (Hedstrom, 2002). As historically situated institutions, they reflect uneven capacities to record, preserve, and circulate experience, crystallizing power relations

that become especially visible in contexts of displacement and political conflict (Bastian, 2003; Giuva, Vitali & Zanni Rosiello, 2007; Pavone, 2009). In such settings, institutional records and personal narratives coexist within broader memorial networks that also include community practices, informal collections, and oral transmission (Hedstrom, 2002; Gilliland, 2017). *Beyond Archives* is situated in this landscape and aims to design a relational digital library on the post-war Julian-Dalmatian exodus to Naples. It brings into relation institutional records and oral testimonies as complementary components of an interconnected memorial field (McKemmish & Piggott, 2013), foregrounding differences in provenance and production contexts while keeping tensions readable for critical inquiry. The project is grounded in a documented phase of data selection that defines the corpus perimeter while acknowledging the interpretive and non-neutral character of that construction. It investigates how to manage, represent, and preserve heterogeneous sources linked to contested diasporic events through a modular workflow, with documentation and paradata oriented to transparency and reuse in line with the FAIR principles (Findable, Accessible, Interoperable, Reusable) (Wilkinson et al. 2016). Its practical goal is to keep surrogates, derived texts, and structural coordinates aligned and inspectable. This contribution outlines the workflow for written and oral materials in parallel and reports early results primarily from the written stream, focusing on semantic layout segmentation and Automatic Text Recognition (ATR) in eScriptorium with the Kraken engine (Kiessling, 2019; Kiessling et al., 2019).

2. ARCHIVES, MEMORY, AND DIGITAL MEDIATION

Scholarship in archival science and memory studies has examined the power relations embedded in archival infrastructures and their effects on how the past is transmitted and represented[1]. In diasporic and conflictual settings, travelling and multidirectional memory further show that institutional frameworks and community narratives may diverge in both content and modes of transmission (Rothberg, 2009; Erll, 2011; Gilliland, 2017). Taken together, this literature suggests that preservation and access practices should accommodate plural traditions and communities of records, while keeping visible the differences in provenance, transmission, and authority that shape what becomes comparable and reusable in the first place (Bastian, 2003; Gilliland, 2017). In digital environments, mediation becomes more complex when records are digitized and re-presented through new representational layers (Hedstrom, 2010; Di Marcantonio, 2025). Scanning and repository deposit provide necessary surrogates, but they require incremental textual and structural layers to support query, inspection, and controlled comparison (Del Grosso et al., 2018; Valacchi, 2024). Recent work on the circulation of archive terminology in digital humanities highlights uneven alignments with archival paradigms and the proliferation of thematic and hybrid digital collections, strengthening the case for methodological clarity about transformations, linking, and description (Long et al., 2025). These dynamics further expose the non-neutral nature of digital documentary aggregation, which incorporates selection, modelling, and re-description (Hoekstra & Koolen, 2019). Digital reuse adds another layer of complexity: records and descriptions circulate across platforms, are reaggregated, and acquire meanings shaped by emergent contexts of use (Borgman, 1999; Tammaro, 2008; Damiani, 2021; Michetti, 2023). Accordingly, documenting how records are transformed, re-described, and linked becomes part of archival responsibility, particularly for vulnerable communities and contested events (Benoit & Eveleigh, 2019), including to mitigate risks of thematic and arbitrary aggregations (Valacchi, 2024). In this perspective, documentation and paradata make the mediations embedded in digital workflows more traceable and keep interpretive decisions open to scrutiny over time.

3. CASE STUDY AND CORPORA

The case study is the post-war Julian–Dalmatian exodus to Naples, shaped by the contested histories of Italy's "eastern border" (Ballinger, 2003). Naples functioned both as a transit hub and as a site of prolonged accommodation through camps and collective facilities, some active until the 1980s (Lazzarich, 2021). This configuration and the polarized narratives that followed provide a lens for memory-making in contested settings, where sources with different forms and provenances require plural approaches to preservation (Gilliland, 2017; Radstone & Wilson, 2021). Against this background, the project brings into

[1] This issue has been addressed by several scholars from different perspectives; for a non-exhaustive selection, see Giuva, Vitali, & Zanni Rosiello (2007); Pavone (2009); Lian (2017); Caswell (2014 and 2016); Bastian (2003); Michetti (2015); Valacchi (2024).

relation institutional records preserved at the State Archives of Naples (Archivio di Stato di Napoli, ASNa) and a corpus of oral testimonies. The institutional corpus derives from six boxes in the Prefecture of Naples fonds[2]. The series concerns refugee management (1946–1958) and includes correspondence, nominal lists, circulars, and reports. The corpus consists mainly of mid-twentieth-century Italian typescripts affected by stamps, marginalia, and handwritten corrections, totaling roughly 5,000 records. These recurrent layouts make the collection suitable for semantic layout segmentation and ATR. The first experimental cycle focused on box 269, used as an initial testing ground to consolidate the approach and develop the first recognition models before extending and refining them on the wider corpus. Oral materials are being developed through semi-structured interviews with direct witnesses, conducted under ethical protocols and described according to shared standards[3]. While this contribution focuses on the written stream, the workflow is conceived to process written and oral materials as parallel streams, modeled in a shared TEI[4] encoding framework and interlinked within the same digital library.

4. METHODOLOGY AND WORKFLOW

The *Beyond Archives* workflow is organized in three stages: digitization (§4.1), formalization (§4.2), and publication (§4.3). These stages are preceded by a preliminary phase of source selection, which defines the documentary perimeter and documents corpus-building choices. The workflow is designed to process written and oral sources as parallel streams within a single infrastructure. In the written stream, page images are imported into eScriptorium[5], where region and line segmentation are iteratively refined and used to drive ATR on fine-tuned Kraken models. Exports in PAGE and ALTO XML[6] preserve alignment between images, layout structure, and derived texts, and support later TEI-oriented modeling (Liégeois et al., 2015) and publication tasks. This paper focuses on two implemented modules that ground the first results in §5: semantic layout segmentation (§4.2.1) and ATR fine-tuning (§4.2.2). Across stages, the approach remains modular and incremental, aiming to stay transferable while keeping transformations inspectable through documentation and paradata, consistent with transparency-oriented practice, integrated workflow perspectives and FAIR-aligned principles (Mazzagufo et al., 2025)[7].

4.1 DIGITIZATION

Digitization corresponds to *Capturing* in TaDiRAH[8] and produces stable surrogates together with minimal technical consolidation for later transformations. For the written stream, this entails high-quality image capture, consistent file organization and naming, and essential technical metadata, in line with Italian national guidance[9]. Master images and key derivatives will be deposited in an institutional repository of a DSpace-GLAM[10] type with stable identifiers and IIIF[11] delivery, separating preservation storage and documentary description from the evolving publication layer. For the oral stream, digitization similarly stabilizes audio surrogates and minimal metadata on rights, consent, and access conditions, so that transcription can later be handled as a documented transformation under clear constraints.

---

[2] See: Archivio di Stato di Napoli, Prefettura di Napoli – Gabinetto, III Versamento, serie V, "Varie"; Buste: 269, 270, 868, 885, 2218, 2275. Inventory edited by Giovanna Buonaurio.

[3] See: Vademecum per il trattamento delle fonti orali (2023); Tavolo Permanente per il Trattamento delle Fonti Orali https://sites.google.com/view/tavolopermanenteperlefontioral/home; CLARIN infrastructure: https://www.clarin.eu [Unless otherwise noted, all web resources were accessed on 15 January 2026.]

[4] Text Encoding Initiative (TEI): https://tei-c.org.

[5] eScriptorium instance: https://obtic-gpu1.mesu.sorbonne-universite.fr:8501.

[6] Page Analysis and Ground-Truth Elements (PAGE XML): Pletschacher & Antonacopoulos (2010); Analysed Layout and Text Object (ALTO XML): https://www.loc.gov/standards/alto/.

[7] For further examples of project environments combining digital workflows, textual processing, and publication infrastructures, see also H2IOSC (Bucciero et al. 2024; Melaccio et al. 2025); DiDcholEd – Digital Scholarly Editions (https://discholed.huma-num.fr/exist/apps/discholed), and Artificial Recognition and Cataloging of Ancient Narrative Object - ARCANO (https://app.transkribus.org/en/sites/arcano).

[8] Taxonomy of Digital Research Activities in the Humanities (TaDiRAH): http://www.tadirah.info; "Capturing": https://vocabs.dariah.eu/tadirah/capturing.

[9] "Linee guida per la digitalizzazione del patrimonio culturale": https://docs.italia.it/italia/icdp/icdp-pnd-digitalizzazione-docs/it/v1.0-giugno-2022/index.html.

[10] DSpace: http://www.dspace.org; DSpace-GLAM, Bollini et al. (2017).

[11] International Image Interoperability Framework (IIIF) https://iiif.io. See also: Salarelli (2017).

4.2 FORMALIZATION

Formalization adds representational layers that make sources computable and interpretable for structured modeling. In TaDiRAH terms, it moves from *Data Recognition* toward *Enriching*[12] and related activities. For the written stream, segmentation and ATR are performed in eScriptorium so that transcriptions remain aligned with images and layout structure. PAGE/ALTO exports preserve coordinates for modeling, inspection, and controlled correction. Oral materials follow an analogous layered logic: Automatic Speech Recognition (ASR) (i.e. Whisper; Radford et al., 2022; Jurafsky & Martin, 2023) is paired with revision for research use and ethical compliance, producing derived texts traceable to recording context. Convergence is achieved through TEI, which provides a shared computational language for written and oral materials and keeps representational decisions inspectable through dedicated guideline modules (Burnard & Larrousse, 2013). In the written stream, the TEIHeader preserves provenance, archival genealogy, and reference to the originating fonds. In the oral stream, it will provide core metadata and explicit links to the corresponding audio/video recordings. To keep results reproducible across iterations, paradata record model versions, dataset partitions, and documented limits through split-defining artifacts and brief notes on recurrent error patterns (Wilkinson et al. 2016; Chiffoleau, 2024).

4.2.1 Semantic layout segmentation

Layout segmentation is trained in eScriptorium through ground-truth annotations to distinguish zones relevant to documentary interpretation. The project adapts SegmOnto as a controlled vocabulary, extending it with region classes so recurring components such as headings, body text, signatures, stamps, and marginal notes remain consistently addressable (Gabay et al., 2021). This supports downstream PAGE-to-TEI mapping by preserving provenance- and context-relevant elements as explicit structures. Figure 1 illustrates this alignment, linking page images, semantic regions, and derived text within the same workspace. Training proceeds iteratively through successive segmentation checkpoints, from an initial in-domain baseline through later rounds of ground-truth expansion and controlled comparison on internal and external subsets. Representative pages are manually annotated, splits are recorded to keep versions comparable, and each iteration enriches the training pool with underrepresented or complex layouts, including stamps and marginalia. Outputs are exported in PAGE XML and derivatives to preserve alignment for transcription and publication. As discussed in §5, line-level behavior becomes more stable, while block-level fragmentation remains more pronounced in archival and paratextual zones. Evaluation is reported in §5 through qualitative review and quantitative comparisons on controlled subsets and can be complemented by additional metrics.



Figure 1. eScriptorium workspace illustrating alignment between the page image, semantic layout regions, and derived transcription, enabling inspectable transformations from surrogate to structured text.

4.2.2 Automatic text recognition fine-tuning

ATR is implemented by fine-tuning Kraken in eScriptorium (Kiessling, 2019; Kiessling et al., 2019), starting from a generic, multilingual, and diachronic baseline for printed and typewritten material, CATMuS-Print [Large] (Gabay & Clérice, 2024), and adapting it to the typography and degradation patterns of the Prefecture corpus. Ground truth is produced by manually correcting lines while preserving layout information in PAGE XML, so transcription remains aligned with images and layout structure. Ground-truth pages are grouped into corpus subsets and compiled into stable train/validation/test

[12] TaDiRAH activities: "Data Recognition": https://vocabs.dariah.eu/tadirah/dataRecognition; "Enriching": https://vocabs.dariah.eu/tadirah/enriching.

partitions through reusable dataset artifacts, including .arrow files, enabling like-for-like comparisons across baseline and fine-tuned evaluations. Fine-tuning proceeds incrementally across successive checkpoints to obtain a domain-specific model that can generalize across the collection. As discussed in §5, the training trajectory includes an initial phase of strong cumulative gains, a later consolidation checkpoint, and subsequent diagnostic runs with more explicit split control and ablation logic. This later phase clarifies the limits of further fine-tuning and supports a more empirical selection of the operational checkpoint. Parameters, split design, and test results are recorded in the methodological record (Bubula et al., 2025). Evaluation in §5 relies on Character Error Rate (CER) and Word Error Rate (WER) and is interpreted in fit-for-purpose terms for TEI encoding and downstream reuse, while alignment-preserving exports support controlled correction and traceable reuse (Echavarria & Le Meur, 2025).

4.3 PUBLICATION AND INTEROPERABILITY

The publication layer is designed to integrate surrogates, derived texts, and TEI-encoded representations in an environment that preserves traceability across layers. The planned stack uses TEI Publisher[13] with an eXist-db backend and connects TEI views to IIIF image services, enabling navigation between facsimiles, transcriptions, and encoded structures (Chiffoleau, 2020). For the written stream, this environment will support enriched consultation of documentary units whose preservation and inventory-level description remain anchored in the institutional repository layer. For the oral stream, the same architecture is intended to support access to transcripts, the corresponding audio/video materials, and their descriptive metadata. The publication layer retains references to segmentation and ATR checkpoints and concise notes on ground-truth creation, correction policies, and post-processing, so derived texts remain auditable and comparable over time. Interoperability relies on IIIF manifests, minimal descriptive metadata, stable identifiers, and TEI-based modeling that preserves provenance and documentary context through the TEIHeader. The publication environment thus extends repository-based description through enriched and relational access. Sustainability is supported through open standards and planned releases on Zenodo and HTR-United[14] of trained checkpoints, model cards, split-defining artifacts, and a curated subset of ground truth for reuse and transparent reporting (Wilkinson et al., 2016).

5. EARLY RESULTS

The first implementation cycle has produced different kinds of results across the workflow. The most advanced and comparable findings concern the written stream, where semantic layout segmentation and ATR have been iteratively refined and evaluated. In parallel, the project has reached a prototype stage in the TEI pipeline and has activated the oral stream through a first exploratory full-pass ASR on the primary recording of the first interview. These latter results are discussed more briefly, but they confirm that the workflow is operative across parallel streams with different degrees of maturity.

5.1 SEGMENTATION RESULTS

Segmentation results confirm a consistent pattern across successive iterations of the written-stream workflow. Performance is strongest on the core textual layout, while zones with higher archival semantics remain less stable and require targeted supervision. The earlier cycles established a usable in-domain baseline for head/body layout and line hypotheses. Later comparisons show a gradual reduction in block inflation together with more mature control of line segmentation, although the trajectory is not perfectly linear across all classes.

On the stable held-out benchmark, predicted block counts decrease from 124 in SegModel-B to 100 in SegModel-C and 92 in SegModel-D, against 21 ground-truth blocks. At line level, the same benchmark shows 134 predicted lines for B, 128 for C, and 137 for D, against 119 ground-truth lines. The strongest improvement therefore concerns the containment of regional over-segmentation, while line behavior remains comparatively more stable. On the validation set of SegModel-D, MainZone:P lines reach recall 0.927 and precision 0.838, while MainZone:Head lines reach recall 0.896 and precision 0.687. The regional profile remains more fragile, with good recall but much lower precision. The model is now sufficiently stable to support transcription-oriented workflows on the main text stream, especially for

[13] TEI Publisher: https://teipublisher.com/index.html.
[14] HTR-United: https://htr-united.github.io.

body text and, to a lesser extent, headings. Signatures show partial improvement, while stamps and marginal areas remain the main source of instability.

5.2 AUTOMATIC TEXT RECOGNITION RESULTS

Across successive ATR cycles, fine-tuning on in-domain ground truth yields substantial gains over the public baseline and reaches a consolidation point that remains the strongest operational reference in the current workflow. Starting from CATMuS-Print [Large] (Gabay & Clérice, 2024), the first three cycles already show strong cumulative improvement, with ATRModel-3 reaching 95.20% character accuracy and 82.28% word accuracy on the combined in-domain test split, about +6 points in character accuracy and +19 points in word accuracy over the baseline. The most robust checkpoint for the written stream is currently ATRModel-5, which reaches 96.81% character accuracy and 87.83% word accuracy on lot-5. Later runs remain methodologically informative, while checkpoint selection is grounded in held-out test performance and documented comparison across runs.

A coarse analysis of recurrent errors for ATRModel-5 shows a profile dominated by substitutions (52.8% of all errors), followed by insertions (39.4%) and deletions (7.8%). The most recurrent error families are omissions of frequent graphemes (15.54%), vowel confusions (11.52%), spacing and segmentation errors (6.72%), and punctuation or weak graphic signs (5.48%), with smaller contributions from confusions between formally similar graphemes (3.48%) and accented or diacritic variants (1.00%). These distributions confirm that residual errors remain concentrated in a limited set of recurrent phenomena and support fit-for-purpose evaluation for TEI-oriented reuse and controlled correction workflows.

6. DISCUSSION AND FUTURE WORK

The first implementation cycle of Beyond Archives shows differentiated but coherent progress across the workflow. On the written stream, segmentation and ATR have produced clear advances under controlled conditions. In parallel, the TEI pipeline has reached a prototypical stage, and the oral stream is structurally activated and documented, although its current outputs remain more exploratory and still under consolidation. Overall, these developments confirm the value of a modular pipeline in which transformations remain inspectable through documentation, paradata, and explicit links between surrogates, derived texts, and structured representations.

For segmentation, the clearest result is the sharper distribution of strengths and limits across document components. Line-level behavior is now the most stable component, especially for MainZone:P and, with greater variability, for MainZone:Head. The main difficulties remain at block level, where over-segmentation continues to affect Signature, StampZone:Reception, and marginal areas. The next step is therefore selective ground-truth expansion on the archival zones that remain structurally fragile, together with possible extension of the current comparative framework through additional and more experimental standard metrics, including Document Layout Error Rate (DLER) (Vesalainen et al., 2024).

For ATR, the updated trajectory supports a more precise reading of the training sequence. Fine-tuning on in-domain material is clearly effective in the early and intermediate cycles, and the sequence reaches a strong consolidation point in ATRModel-5, which remains the most robust operational checkpoint. Subsequent runs do not surpass that checkpoint, but they clarify the limits of further fine-tuning under more rigorous experimental conditions. Evaluation therefore remains useful for checkpoint comparison, correction-effort estimation, further ground-truth selection, and longitudinal inspection of model behavior across versions.

A relevant outcome of this first cycle lies beyond the two core recognition modules. On the written side, the TEI workflow has moved from a stable baseline to a PAGE-informed prototypical phase in which Head, Body, and Signature are progressively converging toward a single documentary transcription layer. On the oral side, the project has established an inventory layer and completed a first exploratory ASR full pass on the primary recording of the first interview. The resulting transcript is currently treated as a first operational draft for post-check, thematic segmentation, and local reruns.

These developments reinforce a broader methodological point. The value of the workflow lies also in the explicit documentation of the mediations introduced at each stage, so that the digital environment can enrich consultation while keeping provenance, documentary stratification, and access constraints legible over time. Future work will proceed along four connected directions. On the written stream, segmentation will be refined on the most fragile archival zones, and ATR evaluation will be extended through additional

out-of-sample testing on previously unseen lots and conditions. The TEI branch will move toward a more stable integration of the main documentary branches within a unified transcription architecture. In parallel, the workflow will progress toward TEI encoding coupled with Named Entity Recognition (NER) to establish reusable access points across the corpus. Shared identifiers will support stable cross-links between written and oral materials while keeping provenance and access constraints explicit. On the oral side, the first full-pass transcript will support post-check, segment mapping, targeted reruns, and progressive revision under the stated ethical and descriptive standards. Finally, the publication layer will continue to be defined as an environment connecting surrogates, derived texts, TEI representations, and, where appropriate, the corresponding audio/video materials. At the same time, the project will strengthen its reproducibility package through model documentation, stable identifiers, split-defining artifacts, and selected ground-truth releases for dissemination through open repository workflows.

REFERENCES

Ballinger, P. (2003). History in exile : memory and identity at the borders of the Balkans / Pamela Ballinger. Princeton University Press.

Bastian, J. A. (2003). Owning memory: How a Caribbean community lost its archives and found its history. Libraries Unlimited. https://doi.org/10.5040/9798400694752

Benoit, E., & Eveleigh, A. (2019). Participatory archives. Facet Publishing. https://doi.org/10.29085/9781783303588

Bollini, A., Cortese, C., Groppo, E. A., & Mornati, S. (2017). DSpace-GLAM (Galleries, Libraries, Archives and Museums): Un'estensione di DSpace per un Digital Library Management System innovativo e sostenibile. *Bibelot: notizie dalle biblioteche toscane*, *23*(1), 46–55.

Borgman, C. L. (1999). What are digital libraries? Competing visions. *Information Processing & Management*, *35*(3), 227–243. https://doi.org/10.1016/S0306-4573(98)00059-4

Bubula, M., Baierer, K., Lehmann, J., Neudecker, C., Rezanezhad, V., & Škarić, D. (2025). How Scalable is Quality Assessment of Text Recognition? A Combination of Ground Truth and Confidence Scores. *Anthology of Computers and the Humanities*, *3*, 1286–1310. https://doi.org/10.63744/GR59c1iXu6Wj

Bucciero, A., Chirivì, A., Colella, R., Emara, M., Greco, M., Palamà, D. M., Pandurino, A., Taurino, F., & Zecca, D. (2024). H2IOSC Project: The Italian Federated Cluster for IoT-Based Monitoring and Digital Twinning of Cultural Heritage. *2024 9th International Conference on Smart and Sustainable Technologies (SpliTech)*, 1–6. https://doi.org/10.23919/SpliTech61897.2024.10612510

Burnard, L., & Larrousse, N. (2013, October). TEI as an archival format. *TEI (Text Encoding in the Web) Conference*. https://hal.science/hal-02153026

Caswell, M. (2014). Seeing Yourself in History. *The Public Historian*, *36*(4), 26–37. https://doi.org/10.1525/tph.2014.36.4.26

Caswell, M. (2016). 'The Archive' is Not an Archives: Acknowledging the Intellectual Contributions of Archival Studies. *Reconstruction*, *16*(1), 21–36. https://doi.org/10.1093/oxfordhb/9780198829324.013.0002

Chiffoleau, F. (2020). *Publication of my digital edition – Working with TEI Publisher* [Text/html]. https://doi.org/10.58079/NMYW

Chiffoleau, F. (2024). *Understanding the automatic text recognition process: Model training, ground truth and prediction errors* [These de doctorat, Le Mans]. https://theses.fr/2024LEMA3002

Damiani, C. (2021). Nomina nuda tenemus. Riformulare il senso archivistico. In L. Pezzica & F. Valacchi (Eds.), *Dimensioni archivistiche. Una piramide rovesciata*. Editrice Bibliografica. https://iris.unicampania.it/handle/11591/489069

Del Grosso, A., Bellandi, A., Giovannetti, E., Marchi, S., & Nahli, O. (2018). Scanning is Just the Beginning: Exploiting Text and Language Technologies to Enhance the Value of Historical Manuscripts. *2018 IEEE 5th International Congress on Information Science and Technology (CiSt)*, 214–219. https://doi.org/10.1109/CIST.2018.8596373

Di Marcantonio, G. (2025). Archivi, tecnologie semantiche e Intelligenza Artificiale: Conoscenza partecipata al servizio del territorio. In *Paesaggi e archivi della Val d'Orcia: Strategie per creare conoscenza: Atti del convegno: Castiglione d'Orcia, 6-7 ottobre 2023 / a cura di Giorgia Di*

Marcantonio e Claudia Maccari (pp. 37–44). EUM - Edizioni Università di Macerata. https://doi.org/10.48219/GDMCM_60569851_005

Echavarria, A., & Le Meur, M. (2025). *CER et WER: Métriques d'évaluation de modèles de reconnaissance automatique de manuscrits et d'imprimés anciens*. Consortium-HN ARIANE - Analyses, Recherches, Intelligence Artificielle et Nouvelles Editions numériques. https://hal.science/hal-05267873

Erll, A. (2011). Travelling Memory. *Parallax*, *17*(4), 4–18. https://doi.org/10.1080/13534645.2011.605570

Gabay, S., & Clérice, T. (2024). *CATMuS-Print [Large]*. Zenodo. https://doi.org/10.5281/ZENODO.10592716

Gabay, S., Camps, J.-B., Pinche, A., & Jahan, C. (2021, September). SegmOnto: Common vocabulary and practices for analysing the layout of manuscripts (and more). *1st International Workshop on Computational Paleography (IWCP@ICDAR 2021)*. https://hal.science/hal-03336528

Gilliland, A. J. (2017). Archival and Recordkeeping Traditions in the Multiverse and Their Importance for Researching Situations and Situating Research. In A. J. Gilliland, S. McKemmish, & A. J. Lau (Eds.), *Research in the Archival Multiverse* (pp. 31–82). Monash University Publishing.

Giuva, L., Vitali, S., & Zanni Rosiello, I. (2007). Il potere degli archivi: Usi del passato e difesa dei diritti nella società contemporanea. B. Mondadori.

Hedstrom, M. (2002). Archives, memory, and interfaces with the past. *Archival Science*, *2*(1–2), 21–43. https://doi.org/10.1007/BF02435629

Hedstrom, M. (2010). Archives and Collective Memory: More than a Metaphor, Less than an Analogy. In T. Eastwood & H. MacNeil (Eds.), *Currents of Archival Thinking* (pp. 163–180). Libraries Unlimited (Santa Barbara, United States).

Hoekstra, R., & Koolen, M. (2019). Data scopes for digital history research. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, *52*(2), 79–94. https://doi.org/10.1080/01615440.2018.1484676

Jurafsky, D., & Martin, J. H. (2023). Speech and Language Processing (3rd ed. draft). https://web.stanford.edu/~jurafsky/slp3/

Kiessling, B. (2019). *Kraken—A Universal Text Recognizer for the Humanities* [Application/pdf]. DataverseNL. https://doi.org/10.34894/Z9G2EX

Kiessling, B., Tissot, R., Stokes, P., & Stokl Ben Ezra, D. (2019). eScriptorium: An Open Source Platform for Historical Document Analysis. *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, 19–19. https://doi.org/10.1109/ICDARW.2019.10032

Lazzarich, D. (2021). Foiba: Genealogy of an untranslatable word. In S. Radstone & R. Wilson (Eds.), *Translating Worlds Migration, Memory, and Culture* (pp. 101–117). Routledge.

Lian, Z. (2017). A History of Archival Ideas and Practice in China. In A. J. Gilliland, S. McKemmish, & A. J. Lau (Eds.), *Research in the Archival Multiverse* (pp. 96–121). Monash University Publishing.

Liégeois, L., Etienne, C., Benzitoun, C., Parisse, C., & Chanard, C. (2015, October 28). *Using the TEI as a pivot format for oral and multimodal language corpora*. Text Encoding Initiative Conference and Member's meeting 2015. https://shs.hal.science/halshs-01345777

Long, J., Frings-Hessami, V., & Feng, H. (2025). Divergence and dialogue: Analyzing the linguistic turn of the archive in digital humanities research. *Archival Science*, *25*(1), 5. https://doi.org/10.1007/s10502-024-09473-7

Mazzagufo, L., Sichera, P., Bruno, D., Cristofaro, S., Del Grosso, A. M., & Spampinato, D. (2025). An Infrastructural Solution for Digital Publication starting from Automatic Layout and Text Recognition: Insights from Italian Literary Journals. *2025 IEEE 8th Congress on Information Science and Technology (CiSt)*, 494–499. https://doi.org/10.1109/CiSt65886.2025.11224217

McKemmish, S., & Piggott, M. (2013). Toward the Archival Multiverse: Challenging the Binary Opposition of the Personal and Corporate Archive in Modern Archival Theory and Practice. *Archivaria*, 76, 111–114.

Melaccio, D., Boschetti, F., & Monachini, M. (2025). Interfacing CLARIN with H2IOSC: Metadata Interoperability through Ontology-based Mediation. *CLARIN Annual Conference Proceedings*. CLARIN Annual Conference 2025, Vienna, Austria.

Michetti, G. (2015, marzo 27). Unneutrality of archival standards and processes. Re:inventing Information Science in the Networked Society (ISI 2015) (ISI 2015), Zadar. Croatia, https://doi.org/10.5281/zenodo.16421

Michetti, G. (2023). Il riuso come trasformazione del contesto. Una prospettiva archivistica. *DIGITALIA*, *18*(2), 99–112. https://dx.doi.org/10.36181/digitalia-00078

Pavone, C. (2009). Archivi e potere. *Contemporanea*, 12(1), 211–217.

Pletschacher, S., & Antonacopoulos, A. (2010). The PAGE (Page Analysis and Ground-Truth Elements) Format Framework. *2010 20th International Conference on Pattern Recognition*, 257–260. https://doi.org/10.1109/ICPR.2010.72

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). *Robust Speech Recognition via Large-Scale Weak Supervision* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2212.04356

Rothberg, M. (2009). *Multidirectional memory: Remembering the holocaust in the age of decolonization*. Stanford University Press.

Salarelli, A. (2017). International Image Interoperability Framework (IIIF): A panoramic view. *JLIS : Italian Journal of Library, Archives and Information Science = Rivista Italiana Di Biblioteconomia, Archivistica e Scienza Dell'informazione : 8, 1, 2017*, (8). https://doi.org/10.4403/jlis.it-12090

Tammaro, A. M. (2008). *Biblioteca digitale e scienze umane. Parte II: La Biblioteca digitale di ricerca per l'apprendimento*. ITA. https://air.unipr.it/handle/11381/1895225

Tavolo permanente per le fonti orali. (2023). *Vademecum per il trattamento delle fonti orali*. Ministero della Cultura. https://dgagaeta.cultura.gov.it/public/uploads/documents/Quaderni/65378110481ed.pdf

Valacchi, F. (2024). *L'archivio aumentato: Tempi e modi di una digitalizzazione critica*. Editrice Bibliografica.

Vesalainen, A., Tolonen, M., & Ruotsalainen, L. (2024). Document Layout Error Rate (DLER) metric to evaluate image segmentation methods. *Machine Learning with Applications*, *18*, 100606. https://doi.org/10.1016/j.mlwa.2024.100606

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18

Wilson, R., & Radstone, S. (Eds.). (2021). *Translating worlds: Migration, memory and culture*. Routledge.

A Workflow-Centred Approach to Managing Semantic Artifacts as Linked Open Data

MICHELE MALLIA, FAHAD KHAN, VALERIA QUOCHI
CNR Istituto di Linguistica Computazionale "A. Zampolli", (CNR-ILC), Italia, name.surname@ilc.cnr.it

ABSTRACT (ENGLISH)

Semantic Artifacts such as vocabularies and lexica are increasingly published as Linked Open Data, yet their management often relies on fragmented and ad-hoc practices. This paper presents a workflow-centred approach for the ingestion, validation, editing and publication of RDF-based Semantic Artifacts within a dedicated infrastructural pilot developed in the context of the H2IOSC project and integrated with the CLARIN-IT infrastructure. The proposed workflow integrates certified repositories, automated quality checks, controlled editing environments and semantic web publication services, supporting both SKOS vocabularies and OntoLex-Lemon lexica. By formalising each step of the lifecycle of a Semantic Artifact, the workflow ensures compliance with FAIR and Linked Data principles, supports versioning and traceability, and reduces technical barriers for Digital Humanities practitioners.The paper discusses how a workflow-oriented design, as implemented in the LLOD pilot, can improve sustainability, reusability and interoperability of semantic resources across research communities.
Keywords: Semantic Artifacts; Linked Open Data; RDF Workflow; Digital Humanities Infrastructure;

ABSTRACT (ITALIANO)

Un approccio incentrato sul flusso di lavoro per la gestione degli artefatti semantici come dati aperti collegati.
Gli artefatti semantici, quali vocabolari e lessici, sono sempre più spesso pubblicati come Linked Open Data; tuttavia, la loro gestione si basa ancora frequentemente su pratiche frammentate e ad hoc. Questo contributo presenta un approccio incentrato sul workflow per l'ingestione, la validazione, la modifica e la pubblicazione di artefatti semantici basati su RDF all'interno di un pilota infrastrutturale sviluppato nel contesto del progetto H2IOSC e integrato con l'infrastruttura CLARIN-IT. Il workflow proposto integra repository certificati, controlli automatici di qualità, ambienti di editing controllato e servizi di pubblicazione del Web Semantico, supportando sia vocabolari SKOS sia lessici conformi a OntoLex-Lemon. Formalizzando ciascuna fase del ciclo di vita di un artefatto semantico, il workflow garantisce la conformità ai principi FAIR e Linked Data, supporta il versionamento e la tracciabilità e riduce le barriere tecniche per i ricercatori nelle Digital Humanities. Il contributo discute inoltre come un design orientato al workflow, così come implementato nel pilota LLOD, possa migliorare la sostenibilità, la riusabilità e l'interoperabilità delle risorse semantiche all'interno delle comunità di ricerca.
Parole chiave: Artifatti semantici; Dati aperti collegati; Flusso di lavoro RDF; Infrastruttura per le Digital Humanities;

1. INTRODUCTION

Semantic Artifacts such as vocabularies, taxonomies, and lexica play a central role in the representation and organisation of knowledge within Digital Humanities research. By providing structured, machine-readable descriptions of domain concepts and their relations, these resources support interoperability across datasets and enable the integration of heterogeneous sources (Antopolsky, 2022). In recent years, the publication of Semantic Artifacts as Linked Open Data (LOD) has further enhanced their potential for reuse and cross-domain integration, by leveraging shared identifiers, standardised data models, and web-based access mechanisms (Bizer et al., 2023; Heath and Bizer, 2011). Within this context, the use of RDF and Semantic Web technologies is closely aligned with the FAIR principles, which promotes findability, accessibility, interoperability, and reusability of research data (Wilkinson et al., 2016).
Despite the availability of mature standards and widely adopted models—such as SKOS for knowledge organisation systems and OntoLex-Lemon for lexical resources—the practical management of Semantic Artifacts remains challenging. In many Digital Humanities projects, the publication of RDF resources, if any, is handled through heterogeneous and often ad-hoc practices. Validation, quality assessment,

editing, and versioning are frequently performed using disconnected tools or informal workflows, resulting in processes that are difficult to document, reproduce, or maintain over time. This fragmentation represents a significant barrier for research communities, particularly for users without advanced technical expertise, and negatively impacts the long-term sustainability and reusability of semantic resources.

Addressing these challenges requires moving beyond tool-centric solutions toward a workflow-centred approach. Explicit workflows can act as enabling mechanisms that structure the lifecycle of Semantic Artifacts into well-defined phases, clarifying responsibilities and transitions between deposition, validation, modification, and publication (Siefkes, 2012). By formalising these phases, workflows support data quality assurance, traceability of changes, and compliance with established best practices, while at the same time reducing technical barriers for non-specialist users involved in the creation and curation of semantic resources.

The objective of this paper is to present an operational workflow for publishing RDF-based Semantic Artifacts in Digital Humanities. The workflow is described through its main phases and discussed in terms of its ability to support FAIR compliance, technical sustainability, and reuse of semantic resources. The workflow is illustrated within the context of the LLOD (Linked Linguistic Open Data) publishing platform, developed within the H2IOSC project and integrated with the CLARIN-IT research infrastructure and the ILC4CLARIN repository.

2. SEMANTIC ARTIFACTS AND DESIGN REQUIREMENTS

In recent Digital Humanities research, Semantic Artifacts are increasingly understood as structured and machine-actionable knowledge resources that formalise domain concepts and their relationships in a way that supports interoperability and reuse. Vocabularies, taxonomies, thesauri, and lexica provide conceptual frameworks that enable consistent description and alignment of heterogeneous cultural and linguistic data across projects and platforms. Rather than functioning as ancillary documentation, these resources act as semantic reference layers that underpin data integration and interpretation in complex research ecosystems (Hyvönen, 2020).

A defining characteristic of Semantic Artifacts is their reliance on shared identifiers, explicit semantics, and formal data models that allow them to be processed computationally while remaining meaningful to human users. When expressed in RDF-based representations and linked to external resources, Semantic Artifacts contribute to the construction of knowledge graphs that support advanced discovery, reasoning, and cross-domain analysis (Hogan et al., 2021). In this perspective, their value lies not only in their internal structure but also in their capacity to function as reusable and extensible components within broader semantic infrastructures.

Managing Semantic Artifacts therefore requires more than basic data storage solutions. These resources are inherently dynamic: conceptual structures evolve, terminological choices are revised, and semantic relations are refined as scholarly understanding develops or as communities negotiate shared interpretations. Such evolution introduces the need for explicit lifecycle management, including controlled modification processes over time. Without these mechanisms, Semantic Artifacts risk becoming inconsistent, opaque, or difficult to maintain, undermining their reliability and long-term reuse (Cox et al., 2021).

The workflow presented in this paper is motivated by a set of design requirements that address these challenges. First, Semantic Artifacts must be managed according to FAIR-by-design principles, ensuring that openness and reusability are embedded from the earliest stages of their lifecycle (Jacobsen et al., 2020). Second, formal validation is required to guarantee both syntactic correctness and semantic coherence with recognised models. Third, a clear separation between the phases of deposition, validation, editing, and publication is necessary to preserve transparency and prevent uncontrolled changes from affecting publicly accessible resources. Finally, quality control and sustainability must be treated as integral aspects of Semantic Artifact management, rather than as post hoc interventions.

With respect to scope, the workflow focuses on RDF-based Semantic Artifacts compliant with widely adopted semantic models in the Digital Humanities, notably SKOS for knowledge organisation systems and OntoLex-Lemon for lexical resources. Restricting the workflow to these models allows concrete validation and management requirements to be addressed while preserving extensibility to additional representations. By articulating these conceptual and operational requirements, this section motivates the adoption of a workflow-centred approach, which is described in detail in the following section and

instantiated within the LLOD (Linked Linguistic Open Data) pilot developed in the context of the H2IOSC project and integrated with the CLARIN-IT research infrastructure.

3. THE LLOD PILOT: CONTEXT AND SCOPE

The workflow presented in this paper is implemented within the LLOD (Linked Linguistic Open Data) pilot, developed in the context of the H2IOSC project and integrated with the CLARIN-IT research infrastructure, which aims to support the creation, management, and dissemination of semantic resources for the Digital Humanities. Such initiatives address the growing need for shared services that facilitate the publication of linguistic and cultural data as Linked Open Data, while promoting interoperability, sustainability, and reuse across heterogeneous research communities (Khan et al., 2023). In this setting, the pilot (see Fig. 1) is introduced exclusively as an operational context in which the proposed workflow is instantiated, rather than as the primary object of discussion.
The infrastructural pilot does not aim to function as a generic RDF editor or as a data repository. Instead, it operates as a workflow orchestrator, coordinating a sequence of actions that span resource ingestion, validation, controlled modification, and publication. By explicitly separating these phases, the workflow enables a clear distinction between data deposition, quality assessment, semantic editing, and public access, supporting transparency and reproducibility.

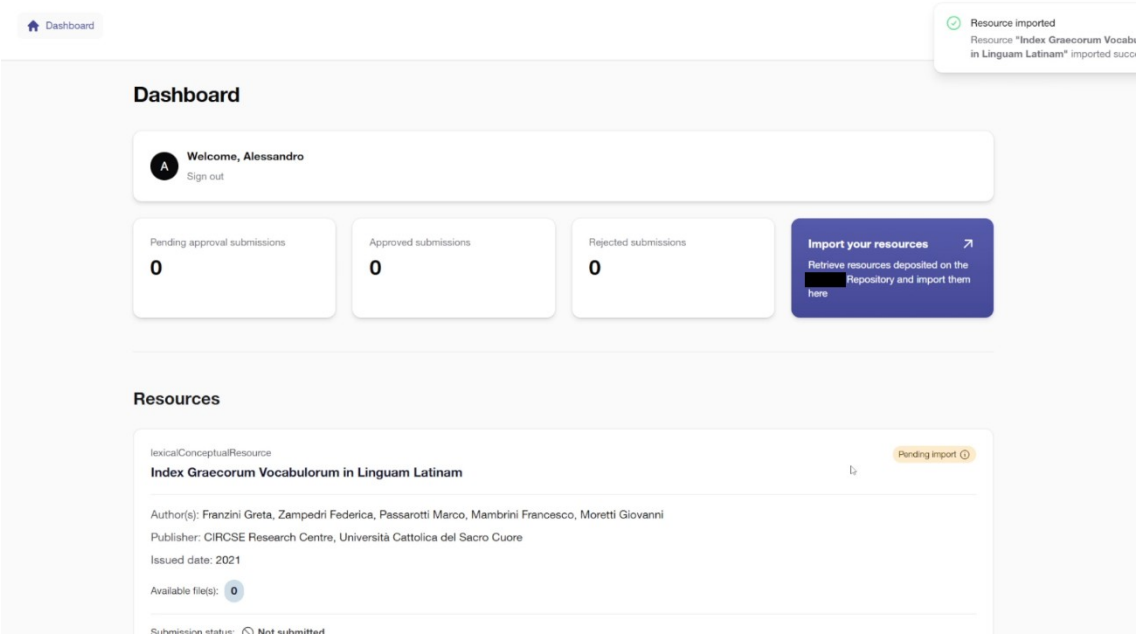


Figure 1: Overview of PLLOD Dashboard

The operational context relies on the interaction between three logically distinct environments. A certified repository—specifically the ILC4CLARIN repository—is responsible for long-term preservation and authoritative access to deposited resources and their versions. In this perspective, the repository is not only a preservation layer but also the entry point of the workflow, ensuring that all resources are managed from the outset as persistent and versioned research objects. A staging environment supports validation and controlled editing activities, enabling iterative refinement while preserving traceability across versions. Finally, a production environment hosts validated resources and exposes them through semantic web services, ensuring stable access and reuse.Within this scope, the pilot supports RDF resources compliant with established semantic models, specifically SKOS vocabularies and OntoLex-Lemon lexica. By constraining the scope to these resource types and by focusing on orchestration rather than on editing or storage functionalities, the pilot provides a controlled setting in which the proposed workflow can be applied, evaluated, and transferred across different Digital Humanities contexts.

4. WORKFLOW FOR MANAGING RDF SEMANTIC ARTIFACTS

The core contribution of this paper is a structured workflow designed to support the lifecycle management of RDF-based Semantic Artifacts, from initial ingestion to public access. Rather than

prescribing specific tools or implementations, the workflow focuses on roles, transitions, and constraints, ensuring that each phase is explicit, traceable, and reproducible.
The workflow, as implemented within the LLOD (Linked Linguistic Open Data) pilot, reflects these principles in a concrete infrastructural setting. The workflow begins with the ingestion of a Semantic Artifact that has been deposited in a certified repository, which is assumed to act as the authoritative source for both the resource and its associated descriptive metadata, including version information and persistent identifiers. In the current implementation, this role is fulfilled by the ILC4CLARIN repository. This design choice reflects the need to treat the certified repository as the authoritative source of record for both the resource and its associated metadata, ensuring persistent identification, versioning, and long-term preservation from the earliest stage of the workflow. By requiring deposition prior to validation, the workflow guarantees that all subsequent processing steps operate on a traceable and citable version of the artifact, in line with FAIR principles. This approach also supports reproducibility and provenance tracking, as even intermediate or non-final versions remain explicitly documented within the repository. Resources may be provided either as native RDF datasets or as collections containing multiple RDF serialisations packaged together. During ingestion, eligible RDF content is automatically retrieved from the repository and prepared for further processing, ensuring that the workflow operates on an authoritative and well-defined input. This phase is responsible for identifying and handling different standard RDF serialisation formats, as well as for resolving and associating the resource with its descriptive metadata. By explicitly linking the retrieved RDF content to the corresponding deposited version and metadata record, the ingestion phase establishes a clear and persistent connection between the original artifact and all subsequent workflow stages. This linkage provides the foundation for traceability, provenance tracking, and reproducibility throughout the lifecycle of the Semantic Artifact.
Following ingestion, the resource undergoes a series of automated validation and quality assessment procedures. These checks are designed to verify both the syntactic correctness of the RDF data and its semantic conformity to recognised models; in the latter case, the workflow makes use of SHACL, a constraint language developed for the Semantic Web. Validation at this stage ensures that the resource can be reliably processed and interpreted, and that it adheres to the formal constraints expected for its declared type.
In particular, the workflow distinguishes between resources compliant with SKOS and those following the OntoLex-Lemon model, applying model-specific validation rules written in SHACL where appropriate. This distinction allows the validation process to account for differences in conceptual structure, required properties, and modelling patterns associated with each model. Beyond model conformity, the workflow incorporates additional quality checks related to Linked Data and FAIR principles, such as the use of resolvable and persistent identifiers, consistency of namespaces, and overall structural coherence of the RDF graph.

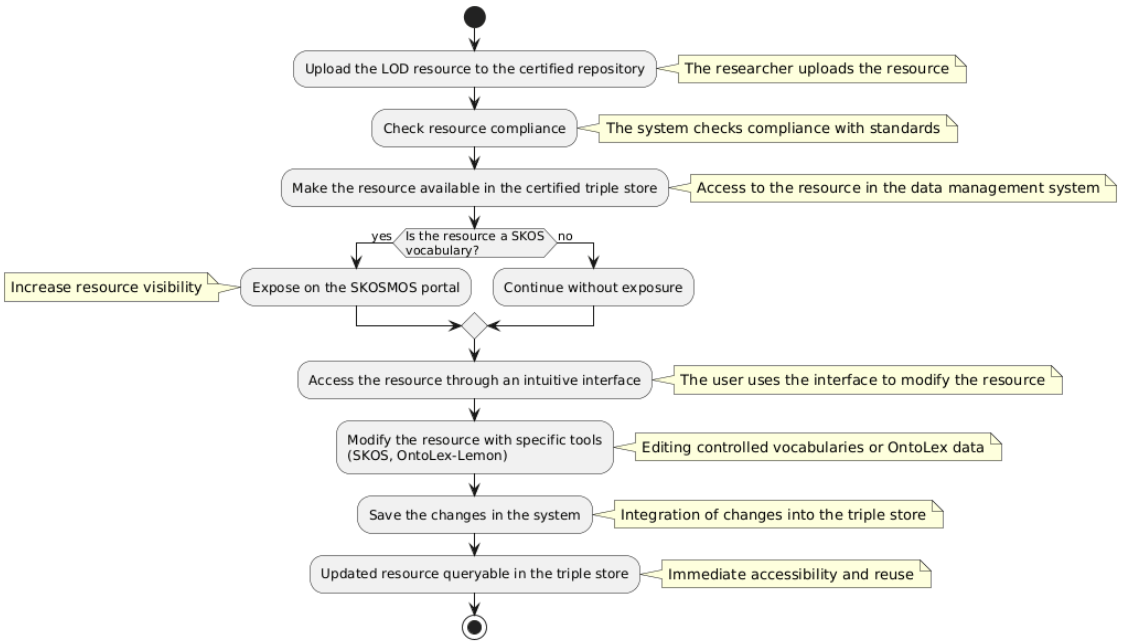


Figure 2: Example of a workflow

Importantly, validation is not conceived as a binary pass-or-fail mechanism. Instead, it is designed as an informative and iterative process that produces diagnostic feedback highlighting potential issues and areas for improvement. This feedback supports subsequent refinement of the resource, enabling users to address modelling or quality problems in a controlled manner before publication.

Resources that do not fully satisfy validation or quality requirements, or that require further semantic enrichment, may enter an iterative improvement phase. In the LLOD pilot, this phase takes place in a dedicated staging environment designed to support controlled editing activities, isolating in-progress modifications from publicly accessible resources. The staging environment allows changes to be applied incrementally while preserving the integrity of validated versions.

Typical interventions during this phase include correcting modelling errors identified during validation, refining conceptual alignments, enriching semantic relations, and normalising labels or metadata to improve consistency and interoperability. Editing activities are guided by the diagnostic feedback produced in the previous phase, ensuring that modifications are targeted and well documented.

To preserve traceability and ensure consistency with repository-based curation practices, any modification of a Semantic Artifact results in the creation of a new version. This revised version must be redeposited in the certified repository, namely the ILC4CLARIN repository, where it becomes subject to a new ingestion and validation cycle. The workflow thus enforces an explicit iterative loop of validation, editing, redeposition, and revalidation, supporting transparency, version control, and the long-term sustainability of Semantic Artifacts.

Once a resource has successfully completed validation and any required revisions, it can be approved for publication. Publication marks the transition from the staging environment to the production environment, where the Semantic Artifact becomes publicly accessible and can be reused by external users and systems. Within the LLOD pilot, this transition corresponds to the controlled publication of validated resources in the production environment integrated with the CLARIN-IT infrastructure. This transition represents a deliberate and controlled step in the workflow, ensuring that only resources that meet defined quality and conformity criteria are exposed.

Validated resources are made available through standard semantic web access mechanisms, including SPARQL endpoints that support structured querying of RDF data and, in the case of SKOS vocabularies, browsing interfaces that facilitate human-readable exploration of concepts and relations. In the current implementation, these access modalities support the publication and reuse of resources managed within the LLOD pilot. These access modalities support both computational reuse and exploratory analysis by human users, broadening the potential impact of published Semantic Artifacts.

By clearly separating publication from earlier workflow phases, the workflow prevents provisional or incomplete versions from being disseminated. This separation ensures that publicly accessible resources correspond to stable, validated versions, reinforcing trust, consistency, and long-term reliability of Semantic Artifacts within broader research ecosystems.

This final phase supports persistent access, reuse, and integration of Semantic Artifacts within broader research ecosystems, including digital research infrastructures and knowledge graph–based environments. By providing stable access points and standard interfaces, published resources can be reliably referenced, linked, and incorporated into downstream applications and analyses.

At the same time, the workflow maintains a clear provenance trail that explicitly links each published resource back to its deposited and validated counterparts. This traceability ensures transparency with respect to the origin, version, and evolution of the Semantic Artifact, supporting both scholarly accountability and long-term maintenance.

Through this structured and iterative workflow, Semantic Artifacts are managed as evolving infrastructural objects rather than static datasets released at a single point in time. As demonstrated in the LLOD pilot developed in the context of H2IOSC and integrated with the CLARIN-IT infrastructure, this perspective enables sustainable reuse across diverse Digital Humanities contexts, fostering interoperability, reproducibility, and cumulative knowledge building over time.

4.1 CONCLUSIONS AND FUTURE DIRECTIONS

This paper has presented a workflow-centred approach for managing RDF-based Semantic Artifacts in the context of Digital Humanities, focusing on the coordinated phases of ingestion, validation, controlled editing, and publication. By formalising these activities within a single operational workflow, the proposed

approach addresses recurring challenges related to data quality, sustainability, and reuse, while at the same time reducing technical complexity for non-specialist users involved in the creation and curation of semantic resources.
A key contribution of the workflow lies in the explicit separation between data deposition and semantic publication. By enforcing validation and versioning practices at clearly defined stages, the workflow ensures transparency, provenance tracking, and long-term traceability of changes. This design supports the treatment of Semantic Artifacts as evolving research objects whose lifecycle is explicitly managed, rather than as static datasets released at a single point in time. In doing so, the workflow aligns the management of Semantic Artifacts with FAIR and Linked Data principles, embedding best practices directly into the operational process.
As demonstrated within the LLOD (Linked Linguistic Open Data) pilot developed in the context of the H2IOSC project and integrated with the CLARIN-IT research infrastructure, the proposed approach proves effective in supporting the sustainable management and publication of semantic resources.
Future work will focus on extending the workflow to support additional semantic models and resource types, as well as on enhancing automation in validation and quality assessment procedures. Further refinements will also address user-facing components to better accommodate the needs of diverse disciplinary communities within the Digital Humanities and Social Sciences. Beyond the specific operational setting discussed in this paper, the workflow is intended to be reusable and adaptable across different research infrastructures, offering a general methodological framework for the sustainable management of Semantic Artifacts as Linked Open Data.

ACKNOWLEDGEMENTS

This work has been supported by the H2IOSC (Humanities and cultural Heritage Italian Open Science Cloud) project, funded by the European Union – NextGenerationEU under the National Recovery and Resilience Plan (NRRP), Mission 4 "Education and Research", Component 2 "From research to business", Investment 3.1 "Fund for the realization of an integrated system of research and innovation infrastructures", Action 3.1.1 "Creation of new research infrastructures and strengthening of existing ones and their networking for Scientific Excellence under Horizon Europe" (Project code IR0000029, CUP B63C22000730005), implemented by CNR.
The work has also been carried out within the framework of the CLARIN-IT infrastructure.

REFERENCES

Antopolsky, A.B. Linked Open Data in the Digital Humanities (Review of Publications). Sci. Tech. Inf. Proc. 49, 119–126 (2022). https://doi.org/10.3103/S014768822202006X

Bizer, C., Heath, T., & Berners-Lee, T. (2023). Linked data – The story so far. In Linking the world's information (pp. 115–143). ACM. https://doi.org/10.1145/3591366.3591378

Cox, S. J. D., Gonzalez-Beltran, A., Magagna, B., & Marinescu, M.-C. (2021). Ten simple rules for making a vocabulary FAIR. *PLoS Computational Biology*, *17*(6), e1009041. https://doi.org/10.1371/journal.pcbi.1009041

Heath, T., & Bizer, C. (2011). *Linked data: Evolving the web into a global data space*. Morgan & Claypool. https://doi.org/10.2200/S00334ED1V01Y201102WBE001

Hitzler, P., Janowicz, K., & Hyvönen, E. (2020). Using semantic web technologies for digital humanities: Shift from data publishing to data-analysis and serendipitous knowledge discovery. *Semantic Web*, *11*(1), 187–193. https://doi.org/10.3233/SW-190386

Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., de Melo, G., Gutierrez, C., Kirrane, S., Labra Gayo, J. E., Navigli, R., Neumaier, S., Polleres, A., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., & Zimmermann, A. (2021). Knowledge graphs. *ACM Computing Surveys*, *54*(4), Article 71. https://doi.org/10.1145/3447772

Jacobsen, A., de Miranda Azevedo, R., Juty, N., Batista, D., Coles, S., Cornet, R., Courtot, M., Crosas, M., Dumontier, M., Evelo, C. T., Goble, C., Guizzardi, G., Hansen, K. K., Hasnain, A., Hettne, K., Heringa, J., Hooft, R. W. W., Imming, M., Jeffery, K. G., … Mons, B. (2020). FAIR principles: Interpretations and implementation considerations. *Data Intelligence*, *2*(1–2), 10–29. https://doi.org/10.1162/dint_r_00024

Khan, A. F., Mallia, M., Quochi, V., Pedonese, G., Frontini, F., & Squadrito, E. (2025). A pilot project for promoting linguistic linked open data. In Proceedings of the 2025 IEEE International Conference on Cyber Humanities (IEEE-CH) (pp. 1–6). https://doi.org/10.1109/IEEE-CH65308.2025.11279386

Siefkes, M. (2012). The semantics of artefacts: How we give meaning to the things we produce and use. IMAGE: Journal of Interdisciplinary Image Science, 8(2), 67–102. https://doi.org/10.25969/mediarep/16556

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, *3*, Article 160018. https://doi.org/10.1038/sdata.2016.18

SEBASTIAN: uno strumento per l'organizzazione, l'analisi spaziale e la visualizzazione dei dati storici[1]

Christian Marcantonio[1], Giuseppe Consolo[2]

[1] DSWLab – Documenti Storici nel Web; ICARUS Italia – christianmarcantonio@gmail.com
[2] Department of Digital Humanities, University of Graz, Austria – giuseppe.consolo@uni-graz.at

ABSTRACT (ITALIANO)

Il contributo presenta SEBASTIAN, uno strumento per l'analisi e la visualizzazione di dati storici. Inserendosi nel dibattito delle Digital Humanities sul ruolo del digitale come elemento capace di incidere sulle pratiche interpretative della ricerca storica, il progetto si concentra sul passaggio dal testo al dato come momento metodologicamente rilevante, in cui le informazioni contenute nelle fonti vengono selezionate, scomposte e strutturate secondo un modello esplicito.
Il contributo assume come banco di prova un corpus di dati storici relativi a beni, luoghi, microtoponimi, usi del suolo ed elementi paesaggistici, estratti da fonti documentarie napoletane tra X e XII secolo. L'obiettivo è sperimentare un modello di organizzazione dei dati orientato all'analisi spaziale e alla rappresentazione del patrimonio documentario in una prospettiva diacronica. L'unità minima di analisi non è il documento nel suo complesso, ma tutte le entità attestate nelle fonti corrispondenti a beni e unità territoriali (e.g. appezzamenti di terra, abitazioni e strutture produttive), registrate all'interno di una struttura tabellare progettata *ad hoc*, che consente di mantenere un collegamento esplicito tra il dato, la fonte documentaria e il contesto archivistico. SEBASTIAN integra un ambiente operativo sviluppato in Python, basato su un database relazionale (SQLite) e dotato di interfacce grafiche per l'inserimento, la gestione e l'analisi dei dati e al contempo capace di generare mappe interattive in formato HTML. Particolare attenzione è dedicata ai temi dell'accessibilità, della tracciabilità e del riuso: ogni informazione visualizzata rimane riconducibile alla fonte di provenienza e lo strumento è concepito per essere reso disponibile in ambiente open source, contribuendo alla riflessione sulle descrizioni aumentate come pratiche di mediazione e di esplicitazione del processo interpretativo nella ricerca storica.
Parole chiave: Storia Digitale; Analisi Spaziale; Strutturazione dei Dati; Mappatura Storica; Patrimonio Documentario

ABSTRACT (ENGLISH)

SEBASTIAN: a tool for the organisation, spatial analysis and visualisation of historical data
This paper presents SEBASTIAN, a tool designed for analysing and visualising historical data. As part of the Digital Humanities debate on the role of digital technologies in reshaping interpretative practices in historical research, the project focuses on the text-to-data transition as a methodologically significant step, in which the information contained in sources is selected, deconstructed, and restructured according to an explicit model.
The study uses a corpus of historical data of properties, places, microtoponyms, land use, and landscape features, derived from Neapolitan documentary sources dating from the 10th to 12th centuries, as its testing ground. The aim is to test a data organisation model designed for spatial analysis and for the diachronic representation of the documentary heritage. The fundamental unit of analysis is not the document as a whole but rather the set of entities attested in the sources corresponding to properties and territorial units (e.g., plots of land, dwellings, and production facilities). These are recorded within a purpose-built tabular structure designed to preserve an explicit link between each data point, its documentary source, and its archival context.
SEBASTIAN integrates an operational environment developed in Python, based on a relational database (SQLite), and equipped with graphical interfaces for data entry, management, and analysis. At the same time, it enables the generation of interactive maps in HTML format. Particular attention is paid to accessibility, traceability, and reusability: every piece of information remains directly traceable to its

[1] Lo strumento è attualmente in fase di sviluppo. Una sua prima versione sarà resa disponibile come repository GitHub sul profilo https://github.com/xpi-an.

source, and the tool is conceived for open-source dissemination in accordance with FAIR principles. The project contributes to the broader reflections on augmented descriptions as practices of mediation and clarification of the interpretative process in historical research.
Keywords: Digital History; Spatial Analysis; Data Structuring; Historical GIS; Documentary Heritage

1. INTRODUZIONE

L'integrazione di metodologie digitali in ambito umanistico ha favorito un ripensamento dei processi di raccolta, strutturazione e restituzione dei dati, rendendo possibili interrogazioni più articolate delle fonti e ampliandone l'accesso anche al di fuori delle comunità di specialisti (Schöch, 2013; Alves, 2014). Il presente contributo si inserisce nel solco di una riflessione che considera il digitale non come un semplice strumento di supporto alla ricerca storica, ma come elemento capace di incidere sulle pratiche interpretative e sulle forme di mediazione del sapere storico, mostrando come la strutturazione dei dati in ambiente digitale possa contribuire alla comprensione e alla valorizzazione del patrimonio storico-documentario, nonché alla loro accessibilità e intelligibilità.
Presentiamo quindi SEBASTIAN, uno strumento *data-driven* per l'analisi spaziale e la visualizzazione dei dati, concepito come un modello progressivamente generalizzabile per l'organizzazione e l'interrogazione di informazioni provenienti da fonti archivistiche. Il progetto pone al centro il passaggio dal testo al dato: scomporre le informazioni contenute nelle fonti relative a beni, luoghi, microtoponimi, diritti, elementi paesaggistici e relazioni territoriali, normalizzarle ove possibile ed estrarle in forma strutturata, così da renderle interrogabili, computabili e visualizzabili in ambienti cartografici esplorabili.
Il lavoro si pone in relazione con le ricerche condotte nell'ambito del *SS. Pietro e Sebastiano Monastery Project*, che ha progressivamente ricostruito e valorizzato la documentazione pervenuta attraverso l'antico archivio monastico di un monastero femminile napoletano di particolare rilievo per la storia medievale di Napoli, ed è tuttora oggetto di ricerche, pubblicazioni, tesi, attività didattiche e sperimentazioni digitali (Ambrosio, 2003; Ambrosio, 2006; Ambrosio, 2016; Ambrosio, 2020; Consolo, 2025; Ambrosio & Consolo, 2026). Tale contesto documentario, frammentario e dispersivo, rappresenta un banco di prova per affrontare il problema della strutturazione e dell'organizzazione di dati storici relativi al patrimonio e alla sua dimensione territoriale nel lungo periodo.
Il contributo propone un modello di organizzazione dei dati orientato alla loro rappresentazione spaziale, in cui le visualizzazioni cartografiche interattive costituiscono parte integrante del processo di analisi. Il lavoro si colloca inoltre nel dibattito sulle *Digital Humanities* come possibilità di rendere dati storici complessi più ampiamente accessibili, esplorabili e riutilizzabili, mentre, nel contesto napoletano, si pone in uno spazio di intersezione ancora poco esplorato, in cui analisi storico-documentaria, rappresentazione cartografica e strutturazione digitale dei dati vengono assunte come momenti profondamente connessi.

2. STATO DELL'ARTE

Negli ultimi anni, la riflessione storiografica e metodologica ha mostrato con crescente chiarezza come il patrimonio documentario medievale costituisca un'importante risorsa per ricostruire forme di proprietà, gestione dello spazio, rapporti giuridici, lessici del paesaggio e pratiche di organizzazione territoriale. In contesti come quello napoletano, caratterizzati da lunga durata istituzionale, continuità abitativa e forte complessità urbana, la ricchezza della documentazione evidenzia la necessità di strumenti in grado di trasformare tale patrimonio in dati strutturati, interrogabili e comparabili. In questa prospettiva, il patrimonio documentario si configura come l'osservatorio privilegiato per l'analisi storica del territorio, delle forme di possesso e delle pratiche di organizzazione dello spazio.
Questa impostazione si inserisce nel panorama storiografico articolato che comprende gli studi sull'evoluzione urbana di Napoli tra tarda antichità e medioevo e le indagini sulla storia agraria e sulle strutture del territorio, le quali hanno evidenziato la complessità delle dinamiche fondiarie, la diffusione di modelli contrattuali e la presenza di molteplici attori sociali (Arthur, 2002; Di Meglio, 2003; Vitolo & Di Meglio, 2003; Di Meglio, 2005; Martin, 2006; Feniello, 2011; Vitolo, 2014). Da tali ambiti di studio emerge un quadro ricco ma disomogeneo, in cui la densità informativa della documentazione si accompagna a limiti strutturali di dispersione, di accessibilità e di leggibilità del materiale pervenutoci e di sistematizzazione di quest'ultimo.
È proprio a partire da tali limiti che nasce l'interesse per nuovi approcci orientati alla formalizzazione digitale delle informazioni desunte dalle fonti, mantenendo al contempo il fondamentale legame con il

contesto documentario e archivistico. È in tale quadro, a seguito di un'ampia ricognizione delle fonti, oggetto di una tesi magistrale nell'ambito del corso di laurea magistrale di Scienze Storiche (Marcantonio, 2024), che l'analisi è stata sviluppata a partire da specifici elementi individuati nella documentazione, utili all'interpretazione dei dati relativi agli aspetti paesaggistici del territorio napoletano nell'arco cronologico compreso tra il X e il XII secolo.
Parallelamente, la riflessione sulla cartografia storica ha evidenziato come le rappresentazioni spaziali del passato siano il risultato di una costruzione interpretativa, condizionata dalle fonti disponibili, dalle categorie adottate e dagli obiettivi della ricerca. In questo senso, la mappa non può essere intesa come uno strumento di restituzione neutro, bensì come un autentico dispositivo che seleziona, ordina e gerarchizza le informazioni (Smail, 2025). La traduzione di tali dati in una rappresentazione spaziale richiede quindi un'esplicita riflessione metodologica, capace di rendere visibili le scelte operate e la relazione tra il dato visualizzato e la fonte che lo sostiene. Lo sviluppo degli *Historical GIS* nelle Digital Humanities ha aperto la possibilità di nuove analisi spaziali, integrando dati testuali, cronologici e geografici in ambienti digitali complessi (Gregory & Healey, 2007). La trasformazione delle informazioni in dati *machine-readable* costituisce un momento indispensabile nell'organizzazione del sapere storico in ambienti digitali, essa rende possibile integrare *corpora* eterogenei e sottoporli a interrogazioni cronologiche e spaziali (Perreaux, 2021), ed è al tempo stesso la fase in cui si rendono esplicite le scelte interpretative e si costruisce la base per ulteriori ricerche sui dataset.
In questa direzione si collocano le riflessioni emerse dal confronto tra ricerca storica e modellazione dei dati, sviluppate a partire dall'analisi del corpus documentario napoletano già richiamato. Tale corpus ha costituito un banco di prova per la definizione di un modello concepito con duplice finalità: da un lato, come strumento aderente al contesto documentario specifico e, dall'altro, come dispositivo trasferibile. Ne deriva una concezione insieme dimostrativa e operativa, in cui la selezione degli attributi rilevanti per l'analisi storica del paesaggio rende computabili alcuni aspetti del dato senza recidere il legame con il contesto documentario.

3. METODI E STRUTTURA DELLO STRUMENTO

SEBASTIAN nasce dalla volontà di organizzare e rendere interrogabili i dati relativi a beni, luoghi, microtoponimi ed elementi paesaggistici ricavati da fonti documentarie medievali. Il suo presupposto metodologico consiste nel passaggio dal testo al dato, inteso come momento interpretativo centrale della ricerca storica in cui le informazioni vengono selezionate, normalizzate e rese comparabili (Schöch, 2013), non come operazione meramente tecnica. Lo scopo di questo progetto non è automatizzare l'interpretazione, ma fornire uno strumento che supporti il lavoro di selezione, organizzazione e interrogazione dei dati, rendendo possibile individuare e mettere in relazione le informazioni rilevanti in base alle specifiche esigenze di ricerca.
Il sistema è progettato assumendo come unità minima di analisi i singoli elementi attestati nelle fonti, distinguendo tra le informazioni spaziali e temporali ed i caratteri paesaggistici associati. SEBASTIAN si configura come uno strumento *data-driven* in cui la strutturazione delle informazioni estratte dai documenti costituisce il presupposto per le operazioni analitiche e di visualizzazione, per cui la mappatura rappresenta l'esito di operazioni computazionali applicate a dati storici organizzati secondo un modello esplicito. In questa prospettiva, il documento non è trattato come unità analitica primaria, bensì come contenitore di più attestazioni. Le informazioni vengono quindi scomposte in singole unità di analisi, corrispondenti agli elementi materiali e territoriali menzionati nelle fonti. Questa scomposizione consente di analizzare distintamente ciascuna entità, evitando di trattare il documento come un unico blocco informativo.
Sul piano operativo, il modello è implementato in una struttura tabellare progettata *ad hoc*, in cui ciascuna riga corrisponde a un'entità. La tabella è organizzata in modo da distinguere chiaramente diversi livelli informativi.
Una prima sezione è dedicata all'identificazione della fonte documentaria e della sua data, attraverso i campi *doc_id* e *date*, che consentono di mantenere un collegamento esplicito tra ciascuna entità e il documento di provenienza.
La parte centrale della struttura è invece dedicata alla descrizione dell'entità, includendo la denominazione attestata (*entity*), la sua tipologia (*entity_type*), i riferimenti toponomastici, articolati tra il livello della testimonianza documentaria (*toponym_attested*) e quello della georeferenziazione moderna con il nome attuale del luogo oppure, nei casi in cui non sia possibile un'identificazione puntuale, l'area o

il contesto territoriale di riferimento (*modern_place*), eventualmente accompagnata da coordinate geografiche (*latitude*, *longitude*).
Un ulteriore livello descrittivo è rappresentato dal campo dedicato agli elementi paesaggistici intesi come *features* territoriali (e.g. uso del suolo, strutture produttive, elementi idrografici) che descrivono il paesaggio e l'organizzazione materiale dello spazio (*landscape_category*). Il sistema consente inoltre di introdurre nuove categorie qualora quelle già presenti non risultino adeguate, permettendo agli utenti di adattare il modello alle specificità del proprio contesto geografico e cronologico e di riutilizzare tali configurazioni in fasi successive della ricerca.
La colonna *notes*, infine, ha il compito di conservare osservazioni e informazioni non riconducibili alle categorie strutturate.
Un esempio della struttura dei campi e della loro organizzazione nell'interfaccia di inserimento è riportato in figura 1. Si tratta di una prima versione di SEBASTIAN, utilizzata per testarne le funzionalità su un dataset ristretto. Il lavoro sul tool è tuttora in corso e prevede un progressivo affinamento grafico e dei campi descrittivi, al fine di restituire in modo più adeguato la complessità delle fonti.

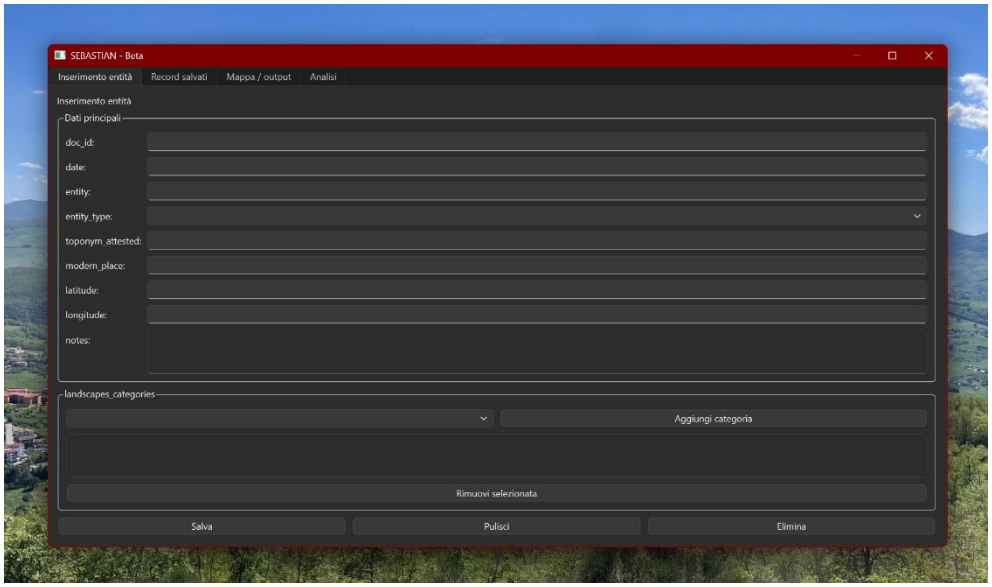


Figura 1. Esempio di interfaccia per l'inserimento dei dati.

Questa organizzazione è stata progettata per rendere integrabili i dati secondo diverse prospettive, mantenendo al contempo il legame con il contesto documentario e la complessità delle fonti.
L'interfaccia grafica consente l'inserimento, la modifica e l'esplorazione dei dati, oltre all'esecuzione di analisi preliminari (distribuzioni per luogo, categoria ed epoca) come riportato in figura 2, rendendo lo strumento accessibile anche a utenti privi di competenze di programmazione.

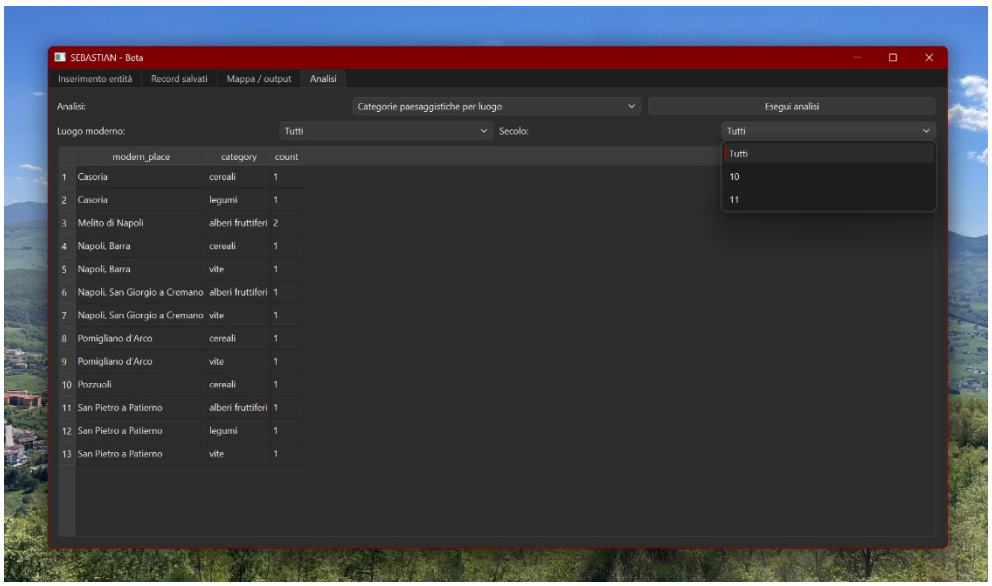


Figura 2. Interfaccia di analisi con dataset di prova.

La visualizzazione spaziale, in fase di sviluppo, prevede la generazione di mappe interattive in formato HTML (figura 3). Questa implementazione rende trasparente l'intero workflow, dalla raccolta dei dati alla loro interrogazione, e costituisce la base per future evoluzioni dello strumento in direzione di una maggiore accessibilità.



Figura 3. Esempio di mappa HTML generata da SEBASTIAN con panel informativo minimo del luogo selezionato.

4. RISULTATI ATTESI

L'applicazione di SEBASTIAN a un corpus di dati territoriali estratti da fonti documentarie prodotte a Napoli mira a un duplice obiettivo: da un lato, impiegare lo strumento per condurre un'analisi concreta di dati relativi a beni, luoghi, microtoponimi, relazioni spaziali ed elementi paesaggistici attestati nella documentazione; dall'altro, il progetto mira a verificarne l'applicabilità su un insieme di fonti eterogenee e discontinue, sia sul piano cronologico sia su quello documentario. Nella fase attuale, il corpus di riferimento è costituito da fonti documentarie risalenti al periodo tra X e XII secolo, ma il modello è concepito per essere esteso ad archi cronologici più ampi e a ulteriori nuclei documentari. È in questa prospettiva di estensione che può essere testata la generalizzabilità del modello, messo alla prova dal processo di strutturazione e normalizzazione di un dataset proveniente dal patrimonio documentario cittadino, stratificato e conservato in forma non uniforme.

Sul piano analitico, i risultati attesi riguardano la possibilità di interrogare in modo sistematico la quantità, la copertura e la qualità dei dati raccolti, osservandone la distribuzione spaziale e diacronica. In particolare, SEBASTIAN dovrà consentire di aggregare e confrontare la tipologia delle entità attestate nelle fonti, le caratteristiche paesaggistiche ad esse associate, la ricorrenza e la localizzazione delle tipologie contrattuali, nonché la presenza e l'evoluzione dei toponimi e dei microtoponimi attestati nella documentazione. In questo modo sarà possibile osservare sia le componenti materiali del patrimonio paesaggistico sia le modalità con cui il paesaggio viene descritto nelle fonti, evidenziando persistenze, slittamenti e ambiguità nella denominazione dei luoghi e delle loro caratteristiche.
Coerentemente con questa impostazione, SEBASTIAN intende offrire anche una base replicabile per confronti con altri contesti documentari e territoriali e, in prospettiva, per condurre analisi relazionali tra luoghi, forme di possesso e processi di trasformazione del patrimonio, osservabili attraverso la ricorrenza delle stesse entità in documenti diversi, le variazioni nelle loro caratteristiche e le modifiche nelle relazioni spaziali e giuridiche nel tempo.
Dal punto di vista degli output, i risultati attesi consistono nella produzione di visualizzazioni cartografiche in formato HTML, organizzate in layer informativi in funzione delle classi di dati analizzate, e di grafici che restituiscano in forma strutturata gli esiti delle interrogazioni. Questi output sono concepiti non solo come restituzione finale, ma anche come strumenti di esplorazione e di supporto all'analisi, capaci di rendere osservabili relazioni spaziali e temporali e di orientare nuove domande sul rapporto tra patrimonio, territorio e fonti nel lungo periodo. Nel loro insieme, questi risultati attesi mirano a mostrare come la strutturazione digitale dei dati storici renda le fonti documentarie degli oggetti interrogabili, definendo al tempo stesso le condizioni di accesso, tracciabilità e riuso dei dati e delle visualizzazioni.

5. ACCESSIBILITÀ, TRACCIABILITÀ E RIUSO DEI DATI

Un aspetto centrale del progetto riguarda l'accesso, la tracciabilità ed il riuso della struttura, dei dati prodotti e delle visualizzazioni generate attraverso SEBASTIAN. Fin dalla fase di progettazione, lo strumento è concepito come un sistema informativo orientato alla trasparenza del processo di ricerca (Atzenhofer-Baumgartner et al., 2024), in cui ogni informazione visualizzata rimane costantemente riconducibile alla documentazione da cui è stata estratta. In questo modo, la trasformazione delle fonti in dati mantiene esplicito il rapporto tra informazione, interpretazione e documento, preservando la verificabilità e l'affidabilità dell'analisi storica. La visualizzazione cartografica si configura come un'interfaccia di accesso ai dati, dalla quale sarà possibile risalire alla fonte, valutarne il contesto di produzione e interrogare criticamente le informazioni visualizzate. Dal punto di vista dell'accessibilità, la creazione di mappe interattive in formato HTML è pensata per una fruizione esplorabile, sia in ambito di ricerca sia in contesti formativi o di *public engagement*, senza richiedere competenze tecniche avanzate. La distinzione tra il livello dei dati strutturati e quello della visualizzazione consente di modulare la complessità dell'output in funzione di pubblici differenti, mantenendo invariata la base informativa e il riferimento alle fonti.
SEBASTIAN è concepito per essere reso disponibile in ambiente open source, insieme ai modelli di dati, al codice e alla documentazione necessaria al suo utilizzo, favorendo la riusabilità dello strumento e la replicabilità dei workflow in diversi contesti di ricerca. Il progetto segue i principi FAIR (*findable, accessible, interoperable, reusable*): l'intero applicativo, con codice e documentazione necessaria, sarà reso pubblico su GitHub, mentre è prevista anche la pubblicazione dei dataset su Zenodo, al fine di garantirne l'accesso e l'utilizzo nel tempo.
Con questo contributo intendiamo contribuire ad una riflessione più ampia sul ruolo delle descrizioni aumentate nelle *Digital Humanities*, intese come dispositivi capaci di coniugare accessibilità, controllo metodologico e verificabilità delle informazioni, trasformando la visualizzazione storica da output a parte integrante della ricerca storica.

RINGRAZIAMENTI

Giuseppe Consolo fa parte del progetto DiDip, finanziato dall'European Research Council (ERC) e attivo presso l'Università di Graz. L'ideazione e lo sviluppo di SEBASTIAN sono stati realizzati da Christian Marcantonio, mentre Giuseppe Consolo ha fornito supporto e consulenza per la parte tecnica; entrambi gli autori hanno contribuito alla definizione delle tipologie di dati incluse nello strumento. Christian Marcantonio è autore delle sezioni 3-4-5; Giuseppe Consolo è autore delle sezioni 1-2. Gli autori

desiderano ringraziare l'European Research Council, Georg Vogeler e il DSW Lab dell'Università degli Studi di Napoli Federico II, in particolare Antonella Ambrosio, per i preziosi consigli e la collaborazione.

BIBLIOGRAFIA

Alves, D. (2014). Introduction: Digital methods and tools for historical research. International Journal of Humanities and Arts Computing, 8, 1–12. https://doi.org/10.3366/ijhac.2014.0116

Ambrosio, A. (2003). Il monastero femminile domenicano dei SS. Pietro e Sebastiano di Napoli. Regesti dei documenti dei secoli XIV-XV, Documenti per la storia degli ordini mendicanti nel Mezzogiorno 1, Salerno, Carlone Editore.

Ambrosio, A. (2006). SS. Pietro e Sebastiano. Monastero domenicano. In A. Valerio (Ed.), I luoghi della memoria. Istituti religiosi femminili a Napoli dal IV al XVI secolo. vol. I, Napoli: Voyage Pittoresque, pp. 367-371,

Ambrosio, A. (2016). Open the shrine. The South Italian charters preserved at the Biblioteca della Società Napoletana di Storia Patria are now accessible for all on Monasterium.net in European Archival Blog.

Ambrosio, A. (2020). La Diplomatica e il digitale. Il Fondo della Biblioteca della Società Napoletana di Storia Patria online, in Research Trends in Humanities Education & Philosophy, pp. 1-15

Ambrosio, A., & Consolo, G. (2026). Dai registri medievali alle tecnologie digitali: Un'edizione critica per nuove prospettive di ricerca. In R. Di Meglio & F. Senatore (Eds.), Topografia delle attività economiche nelle città campane del basso Medioevo,UniorPress, pp. 173–200.https://doi.org/10.6093/978-88-6719-321-9

Arthur, P. (2002). Naples, from Roman Town to City-State: An Archaeological Perspective. London: British School at Rome.

Atzenhofer-Baumgartner, F., Geiger, B. C., Vogeler, G., & Kowald, D. (2024). Value Identification in Multistakeholder Recommender Systems for Humanities and Historical Research: The Case of the Digital Archive Monasterium.net. https://doi.org/10.48550/arXiv.2409.17769

Consolo, G. (2025). L'edizione critica digitale della documentazione del monastero dei SS. Pietro e Sebastiano di Napoli nel medioevo [Tesi di Dottorato]. Università degli Studi di Napoli "Federico II".

Di Meglio, R. (2003). Il convento francescano di S. Lorenzo di Napoli. Regesti dei documenti dei secoli XIII-XV, Salerno: Carlone Editore.

Di Meglio, R. (2005). Ordini mendicanti e città: l'esempio del convento di S. Lorenzo Maggiore di Napoli. In Romano, S. & Bock, N. (Eds.), Le chiese di San Lorenzo e San Domenico: gli ordini mendicanti a Napoli, Napoli: Selecta Napoli, pp. 15-26.

Feniello, A. (2011). Napoli: Società ed economia : 902-1137. Nella sede dell'Istituto Palazzo Borromini.

Gregory, I. N., & Healey, R. G. (2007). Historical GIS: Structuring, mapping and analysing geographies of the past. Progress in Human Geography, 31(5), 638–653. https://doi.org/10.1177/0309132507081495

Marcantonio, C. (2024). Paesaggi e prospettive digitali: un'indagine sui documenti napoletani tra i secoli X e XII (tutor: Prof. Rosalba Di Meglio, Prof. Antonella Ambrosio). [Tesi di Laurea Magistrale] Università degli Studi di Napoli "Federico II", CdLM in Scienze Storiche, a.a. 2023-2024.

Martin, J. M. (2006). I contratti agrari altomedievali di area campana. In Contratti agrari e rapporti di lavoro nell'Europa medievale, Bologna: CLUEB, 1-26.

Perreaux, N. (2021). Possibilities, Challenges and Limits of a European Charters Corpus (Cartae Europae Medii Aevi—CEMA). https://doi.org/10.48550/arXiv.2105.00932

Schöch, C. (2013). Big? Smart? Clean? Messy? Data in the Humanities. Journal of Digital Humanities, 2(3), 2–13.

Smail, D. L. (2025). Cartografie immaginarie. Mappare il possesso e l'identità nella Marsiglia bassomedievale. Firenze University Press. (Florence). https://doi.org/10.36253/979-12-215-0606-8

Vitolo, G. (2014). L'Italia delle altre città: Un'immagine del Mezzogiorno medievale. Liguori editore.

Vitolo, G., & Di Meglio, R. D. (2003). Napoli angioino-aragonese: Confraternite, ospedali, dinamiche politico-sociali. Salerno: Carlone.

Raccontando e valorizzando il patrimonio culturale. Dalle prime esperienze ai progetti futuri del CNR-ISEM

MARIA GRAZIA MELE R.[1], SAMUEL ALDO IACOLINA[2], LUIGI SERRA[3], GIOVANNI SERRELI[4]

[1]CNR ISTITUTO DI STORIA DELL'EUROPA MEDITERRANEA (CNR-ISEM), ITALIA - MELE@ISEM.CNR.IT
[2]CENTRO DI RICERCA, SVILUPPO E STUDI SUPERIORI IN SARDEGNA (CRS4)- SAMUEL@CRS4.IT
[3]CNR CNR ISTITUTO DI STORIA DELL'EUROPA MEDITERRANEA (CNR-ISEM) - SERRA@ISEM.CNR.IT),
[4]CNR CNR ISTITUTO DI STORIA DELL'EUROPA MEDITERRANEA (CNR-ISEM) - SERRELI@ISEM.CNR.IT

ABSTRACT (ITALIANO)

Fin dalle origini l'Istituto di Storia dell'Europa Mediterranea del Consiglio Nazionale delle Ricerche (CNR-ISEM) ha dedicato particolare attenzione alle attività di divulgazione delle ricerche in ambito storico, già prima che la cosiddetta Terza Missione fosse istituzionalizzata. Al principio furono mostre e pubblicazioni destinate a larghe fasce di popolazione; dopo il primo decennio del Duemila gli storici dell'Istituto, in collaborazione con altri specialisti, si sono dedicati anche ad attività di divulgazione e comunicazione scientifica attraverso allestimenti multimediali e musealizzazioni.
La filosofia che ha animato queste esperienze è stata quella del coinvolgimento sensoriale e immersivo del fruitore, attraverso installazioni multimediali interattive nelle quali personaggi, monumenti e testimonianze materiali si raccontano in prima persona. Nella realizzazione dei prodotti multimediali e degli allestimenti sono state coinvolte comunità e compagnie di rievocazione storica locali.
Tali esperienze hanno permesso la valorizzazione dei beni storico-culturali di alcune realtà locali (il castello medievale di Marmilla a Las Plassas; l'epigrafe cinquecentesca sull'incursione del corsaro Barbarossa a Gonnostramatza; la torre costiera di Chia a Domus de Maria) e sono diventate un modello di approccio alla valorizzazione del patrimonio culturale materiale e immateriale. Ci sono state anche criticità dovute alla difficoltà di aggiornamento dei prodotti e alla mancanza di continuità progettuale e amministrativa.
Facendo tesoro di ciò, i ricercatori dell'ISEM sono ora impegnati in altri due progetti di valorizzazione del patrimonio culturale: la digitalizzazione di importanti relazioni cinquecentesche sulle coste del Regno di Sardegna attraverso la realtà aumentata e l'AI; la valorizzazione del patrimonio storico del compendio naturalistico e industriale del Parco Regionale Naturale Molentargius-Saline, in collaborazione con ASC e CRS4.
Parole chiave: immersive experience; digital humanities; valorizzazione storico-culturale; digital storytelling; human computer interaction.

ABSTRACT (ENGLISH)

Telling and promoting cultural heritage. From early experiences to future projects of the CNR-ISEM
Since its inception, the Institute of History of Mediterranean Europe of the National Research Council (CNR-ISEM) has devoted particular attention to the dissemination of historical research, even before the so-called Third Mission was formally institutionalized. In its early years, the Institute focused on exhibitions and publications aimed at broad audiences; after the first decade of the 2000s, its historians, in collaboration with other specialists, also engaged in scientific dissemination and communication through multimedia installations and museum displays.
The guiding philosophy behind these initiatives has been to create sensorial and immersive experiences for visitors, using interactive multimedia installations in which historical figures, monuments, and material testimonies narrate their stories in the first person. Local communities and historical reenactment groups were involved in the development of these multimedia products and installations.
These experiences have contributed to the valorization of the historical and cultural heritage of several local contexts (the medieval castle of Marmilla in Las Plassas; the sixteenth-century epigraph commemorating the corsair Barbarossa's incursion in Gonnostramatza; the coastal tower of Chia in Domus de Maria) and have become a model for approaches to enhancing both tangible and intangible cultural heritage. Some critical issues have also emerged, including difficulties in updating the products and a lack of continuity in project and administrative support.

Building on these lessons, ISEM researchers are now engaged in two further cultural heritage valorization projects: the digitization — from the perspective of the original authors — of significant sixteenth-century reports on the coasts of the Kingdom of Sardinia, using augmented reality and AI; and the enhancement of the historical heritage of the naturalistic and industrial complex of the Molentargius-Saline Regional Natural Park, in collaboration with ASC and CRS4.
Keywords: digital scholarly editing; Linked Open Data; historical-cultural valorization; immersive engagement; interaction.

1. INTRODUZIONE

L'ISEM, erede dell'Istituto sui Rapporti Italo-Iberici (IRII), accanto alla tradizionale attività di ricerca storica, ha progressivamente sviluppato, nel corso degli ultimi decenni, una linea strategica orientata alla valorizzazione, alla disseminazione e al trasferimento dei risultati scientifici. Tale indirizzo si colloca nel quadro delle politiche nazionali ed europee sulla Terza Missione e sull'impatto sociale della ricerca, con l'obiettivo di rendere le conoscenze prodotte accessibili, riutilizzabili e funzionali ai processi di crescita culturale, economica, sociale e occupazionale.
Già prima che la Terza Missione venisse istituzionalizzata, l'Istituto aveva dedicato parte delle proprie attività a iniziative volte a far conoscere a un pubblico ampio i risultati delle ricerche condotte, come nel caso della mostra di testi rari e carte geografiche della Biblioteca Universitaria di Cagliari Descubrir el Levante por el Poniente, accompagnata da un catalogo multimediale e svoltasi a Cagliari nel dicembre 2001 (Gallinari, 2002).
In un contesto nazionale caratterizzato da un patrimonio culturale di eccezionale ampiezza, densità e stratificazione, l'ISEM promuove modelli integrati di produzione e comunicazione del sapere storico, finalizzati a favorire il dialogo tra ricerca, istituzioni territoriali, sistema produttivo e società civile, contribuendo alla costruzione di strumenti conoscitivi utili alla tutela, alla gestione e alla valorizzazione delle risorse culturali e paesaggistiche.
Parallelamente, si avverte con crescente evidenza l'esigenza di integrare nello studio interpretativo delle fonti le più recenti metodologie e infrastrutture digitali, capaci di supportare efficacemente i processi di analisi, correlazione e comparazione di insiemi documentari complessi e di grande consistenza.
La progressiva estensione dell'archeologia a fasi cronologiche più recenti — in particolare all'età post-medievale — insieme al consolidarsi di pratiche di collaborazione multidisciplinare, ha ampliato gli ambiti di indagine, includendo in modo sistematico le testimonianze materiali superstiti. Il loro studio richiede oggi un quadro metodologico integrato e multilivello, fondato sull'interazione tra competenze diverse e coerente con gli indirizzi scientifici nazionali in materia di integrazione dei saperi.
Conoscere e far conoscere la nostra storia è sempre più necessario; con la crescita dell'offerta turistica diventa ancor più imprescindibile garantire che i nostri monumenti, le testimonianze materiali e il territorio siano adeguatamente studiati e raccontati. Coinvolgendo tutte le istituzioni interessate nei processi di tutela e valorizzazione, gli enti di ricerca e le università devono assicurare che la diffusione della conoscenza sia scientificamente corretta e trasmessa in modo appropriato agli enti locali, agli operatori e agli imprenditori del settore turistico e culturale. È fondamentale considerare fin dall'inizio i beni culturali e paesaggistici nella loro interezza, mettendo a disposizione le competenze necessarie affinché siano conosciuti, compresi, valorizzati e, di conseguenza, tutelati (Delli Noci, Gabellone, Giannotta, Lombardo, Matarrese, 2001; Antinucci, Cultraro, Sabbioni, Siano, 2010; Antinucci, 2014; Cultraro, 2022).

2. METODOLOGIA

Esistono diversi approcci per presentare un tema, un monumento o un personaggio storico. Il metodo più tradizionale si basa su testi scritti, esposti in pannelli all'interno di mostre temporanee o musei; quello più tecnologico, invece, ha privilegiato l'uso dell'immagine — bidimensionale o tridimensionale — delle ricostruzioni, delle scansioni digitali, della realtà aumentata o diminuita, accompagnate da testi esplicativi esteticamente curati e/o da una voce narrante in sottofondo.
Oggi più che mai, soprattutto per raggiungere i più giovani, la comunicazione è essenzialmente multimediale e, quando possibile, interattiva. Gli adolescenti digitali hanno un'attenzione selettiva che si affida principalmente alla vista e all'udito; diventa quindi necessario sviluppare metodologie capaci di coinvolgere anche le nuove generazioni — o comunque un pubblico ampio — senza rinunciare al rigore

scientifico dei contenuti, trasmettendo anche il dubbio e dunque la possibilità di continuare la ricerca. L'obiettivo è comunicare la Storia come avremmo voluto apprenderla da bambini, affinché rimanesse più impressa: attraverso la memoria visiva e uditiva, le sensazioni, i disegni, gli audio, i filmati.
Per raggiungere questo risultato è fondamentale che storici, geografi, geologi e archeologi lavorino insieme a un regista multimediale, individuando peculiarità e vissuti di un territorio, mettendo in evidenza testimonianze materiali e figure storiche: saranno loro a permettere una narrazione più naturale e immediata.
Dare la parola a personaggi e monumenti consente di far rivivere i protagonisti del passato nella loro storia e nel loro contesto. Incontrarli e ascoltarli direttamente rende l'esperienza più immersiva e coinvolgente.
Con questa metodologia, abbandonando la scheda tecnica museale e il pannello espositivo tradizionale, si passa alla selezione di luoghi e personaggi significativi capaci di dialogare con l'utente. Per far rivivere un monumento o un personaggio è necessario trasformare il testo scientifico in una sceneggiatura che permetta all'oggetto, al monumento o al protagonista storico di accogliere e guidare il visitatore, riportandolo indietro nel tempo attraverso filmati e contenuti audio. Un attore interpreta il personaggio, ne racconta le vicende e le spiega con l'ausilio di mappe e disegni: il discorso diventa chiaro perché collocato visivamente in un luogo definito, e si arricchisce di emotività grazie alla musica composta ad hoc e al tono della voce.
Personaggi, oggetti e monumenti rivivono così attraverso la multimedialità, raccontando coralmente le loro storie. La trascrizione o il regesto di documenti d'archivio si trasformano in una giornata della loro vita, contribuendo a costruire la scenografia di una narrazione significativa e coinvolgente (Serreli, 2016; Mele, 2020).

3. ESPERIENZE PASSATE

Da alcuni decenni l'ISEM sperimenta e implementa questa metodologia comunicativa, aprendosi costantemente a nuove sfide che favoriscono una migliore interazione tra competenze umanistiche e informatiche.
Le installazioni finora realizzate seguono un principio comune: nelle musealizzazioni e negli allestimenti non è previsto nulla da leggere; il visitatore entra invece con naturalezza in un viaggio immersivo nel passato, dove personaggi, monumenti e oggetti d'uso quotidiano interagiscono tra loro. Anche un piccolo e apparentemente insignificante frammento ceramico — un coccio, un pezzo di ciotola o di piatto — diventa narratore delle proprie vicende e della storia dei traffici commerciali.
Nell'installazione del MudA di Las Plassas, ad esempio, i visitatori possono ascoltare la ceramica valenzana che, con accento catalano, racconta dove fu prodotta, chi la vendette, chi la utilizzò e come venne rinvenuta dagli archeologi nel sito del castello. Possono assistere a una giornata in cui il castello ospita il re e il suo seguito, lasciandosi coinvolgere dall'arte culinaria del cuoco. I bambini possono giocare con i personaggi del passato e apprendere la storia in modo spontaneo. Sagome, disegni, proiezioni e audio popolano lo spazio espositivo, senza alcun pannello da leggere.
Anche la scheda tecnica di una torre costiera di età moderna, tradizionalmente accompagnata da un pannello esplicativo, può trasformarsi grazie alla multimedialità. Nel progetto Torri multimediali, la torre — un tempo elemento di un sistema difensivo della Monarchia ispanica — mantiene la sua funzione di nodo di collegamento: non trasmette più segnali di fuoco o di fumo, ma veicola informazioni sui beni culturali del territorio. Un tempo difendeva il paesaggio circostante, oggi può raccontarlo e contribuire alla sua tutela (Mele, Serreli, Casu, Merone 2007a, pp. 81-82; Mele, Serreli, Casu, Merone 2007b, p. 24; Mele, G. Serreli, Multimedia reuse of medieval monuments, in Le tecnologie del CNR per i beni culturali, Roma, 2010, p. 21).
Il primo allestimento del progetto è stato realizzato all'interno della Torre di Chia (Domus de Maria, CA) situata su un promontorio costiero dalla vista straordinaria. L'interno ospita un allestimento essenziale ma di grande impatto, composto da alcuni monitor — metafora delle antiche feritoie — e da un database territoriale consultabile tramite touch-screen. Con la vista e il suono del mare come sottofondo, la torre inizia a raccontare il proprio vissuto, accompagnata da disegni e musiche composte ad hoc, spiegando chi la fece costruire, quando e perché.
Il progetto Torri multimediali può essere esteso a molte altre torri costiere, con installazioni differenziate in base al territorio, alle attività economiche locali e alle vicende storiche che le hanno caratterizzate.

Ogni torre, pur appartenendo a un sistema unitario, si racconterà in modo diverso e offrirà nuove informazioni.
Negli allestimenti di questo genere siamo andati ancora oltre. Un'epigrafe e un retablo, pur noti da tempo agli specialisti e valorizzati in modo tradizionale, sono stati utilizzati — insieme ai frammenti di maiolica e in collaborazione con un archeologo — per rinnovare un intero museo, raccontando la storia del Regno di Sardegna e del Mediterraneo nel XVI secolo.
Nel Museo Turcus e Morus di Gonnostramatza, l'epigrafe, il retablo cinquecentesco, la distruzione e la cattura degli abitanti di Uras e le vicende del villaggio superstite di Sèrzela costituiscono i cardini del percorso espositivo. Ad accogliere i visitatori è lo stesso Barbarossa (su schermo touch-screen), incorniciato come un quadro d'altri tempi, che con fierezza narra la propria storia.
Comunicare un personaggio storico è un'operazione complessa: occorre collocarlo nel suo contesto, trasmetterne il carattere attraverso l'aspetto, l'abbigliamento, i gesti, il modo di parlare. Quando la collaborazione tra ricercatori, regista e attore funziona al meglio, la comunicazione diventa immediata e il personaggio rimane impresso nel visitatore, che lo ascolta rapito. Nel racconto di Barbarossa abbiamo trasmesso tutto l'orgoglio di un uomo che si è fatto da sé: nato figlio di un vasaio, divenne pirata, poi corsaro, fino a raggiungere il grado di ammiraglio della potente flotta ottomana. Il personaggio è carismatico, astuto, capace di rapportarsi agli altri e di presentare la propria versione dei fatti. È importante sottolineare che le sue narrazioni si basano sui più recenti risultati storiografici dedicati ai protagonisti della frontiera mediterranea di età moderna.
Barbarossa introduce il percorso che il visitatore compie all'interno del museo, immergendosi nelle vicende del Mediterraneo del XVI secolo: osserva le immagini del retablo in una grande proiezione che ne evidenzia ogni dettaglio; incontra tre manichini — un pirata, uno schiavo e un torriere — che prendono vita nei rispettivi filmati e si raccontano; scopre i pop-up dei personaggi chiave dell'epoca: Carlo V, Solimano il Magnifico, Filippo II, Andrea Doria e altri. La chiesa campestre di San Paolo, dove è murata l'epigrafe, insieme all'abitato abbandonato di Sèrzela e ai materiali ceramici rinvenuti, introduce le vicende successive alla distruzione di Uras: la vita del villaggio superstite e la costruzione del sistema di torri costiere per l'avvistamento del nemico. Il museo si inserisce così nel progetto Torri multimediali, contribuendo ad arricchirne ulteriormente i contenuti.
In conclusione, al di là delle nuove tecnologie di musealità virtuale, la comunicazione efficace che proponiamo richiede un lungo e rigoroso lavoro scientifico su oggetti, siti e personaggi, un'analisi accurata dei contesti e una successiva elaborazione dei dati, affinché possano essere trasformati in una narrazione diretta e coinvolgente. Le parole dei personaggi e degli oggetti comunicano ciò che la ricerca ha scoperto fino a quel momento, veicolano una corrente storiografica, possono lasciare spazio al dubbio, esprimere coralità o divergenze di punti di vista e consentono un'ampia libertà interpretativa.

4. EVOLUZIONI. SULLA ROTTA DI MARCO ANTONIO CAMÓS: COAST VIEW CON GOOGLE

Ma gli esempi possono andare ben oltre quanto è stato finora realizzato. Nelle prossime installazioni si potrebbe, ad esempio, "circumnavigare" la Sardegna seguendo lo stesso percorso compiuto dal capitano Marco Antonio Camós negli anni Settanta del Cinquecento. Su ordine del re Filippo II, Camós effettuò un periplo dell'isola per verificare le torri costiere già esistenti e individuare i siti in cui costruirne di nuove. Sarebbe una sfida affascinante: ripercorrere via mare quel tragitto, cercando di ricreare — attraverso realtà aumentata e realtà diminuita — l'esperienza di quasi quattro secoli e mezzo fa, catturando con immagini e filmati l'intera fascia costiera descritta nel documento antico.
Alla base vi è la ricerca scientifica tradizionale: una nuova edizione della relazione di Camós e di altre preziose fonti, rese più coinvolgenti dall'idea di far rivivere la Storia arricchendola di nuovi contenuti. Tecnicamente, il progetto prevede la ripresa della costa dal mare con una tecnologia simile allo "Street View" di Google installata su un'imbarcazione che segue la rotta originaria. Oggi ciò è realizzabile anche con tecnologie OpenStreetMap, mediante riprese sferiche a 360°, per simulare la vista che Camós e il suo equipaggio — così come pirati e corsari — ebbero delle coste sarde. Eventuali ricostruzioni in Realtà Virtuale, aumentata o diminuita, permetterebbero di simulare lo stato dei luoghi al tempo di Camós, restituendo l'aspetto che presumibilmente avevano allora. La spedizione effettuò sia una ricognizione marina sia una terrestre nei pressi dei siti individuati; viste integrate marine, terrestri e aeree potrebbero essere restituite tramite riprese con fotocamere 360° a spalla e droni nei pressi dei monumenti (Mele, Serra, Serreli, 2015; Serra, 2016; Di Marcantonio, 2024; Ciletti, 2025).

In un altro progetto (Ideha) sarà possibile far conoscere la ricostruzione virtuale di una città e dei suoi punti nevralgici, colti in un preciso momento storico o in un evento significativo: la visita dell'imperatore, una seduta parlamentare, una riunione del consiglio cittadino, un'assemblea di quartiere. Le proposte possono diventare sempre più interattive, coinvolgendo direttamente i visitatori. Nuove tecnologie ci attendono grazie alle future collaborazioni con il CRS4.

5. EVOLUZIONI. CONFRONTO DELLE FONTI STORICHE MEDIANTE LLM, RAG ED EVENTUALE FINE TUNING COMPORTAMENTALE

Tra le proposte di ricerca sperimentale da avviare si colloca anche l'analisi comparativa di fonti storiche omogenee ed eterogenee mediante tecniche di NLP basate su LLM, RAG ed eventuale fine-tuning comportamentale, con l'obiettivo di affinare la ricerca dei contenuti all'interno di tali fonti e, di conseguenza, ottenere risposte coerenti a quesiti specialistici. L'esigenza nasce dalla necessità di confrontare, tra le altre, due fonti storiche sincroniche di età moderna e di analizzare i toponimi in esse presenti.
L'attività di normalizzazione dei contenuti, una volta trascritte le relazioni, e l'inserimento di annotazioni interne o metadati per il tagging dei toponimi nelle loro diverse varianti toponomastiche e linguistiche, potrà facilitare la ricerca integrata tra fonti eterogenee che riportano gli stessi luoghi. Tra queste rientrano, ad esempio, le carte IGM e la cartografia storica opportunamente indicizzate, ma non solo. La metodologia proposta sarebbe inoltre utile allo studio dei porti e degli approdi, recentemente affrontato come unità di ricerca di un PRIN PNRR, attraverso l'analisi testuale di portolani, diari di bordo e relazioni di viaggio per localizzare e caratterizzare i siti descritti.
Queste contestualizzazioni non rappresentano gli unici ambiti applicabili: una volta configurato il sistema, esso potrebbe essere declinato su una molteplicità di fonti, generando risposte molto accurate grazie all'integrazione tra LLM e RAG. Tale approccio mira a colmare le note limitazioni dei modelli linguistici, che possono produrre allucinazioni dovute alla mancanza di ground truth e ai bias introdotti consapevolmente — attraverso scelte progettuali, criteri di selezione e strategie di allineamento nelle fasi di training e fine-tuning — o inconsapevolmente, come conseguenza della composizione e dello sbilanciamento dei dati di addestramento.

6. NUOVI PROGETTI: VALORIZZAZIONE STORICO-CULTURALE DEL PARCO REGIONALE NATURALE DI MOLENTARGIUS-SALINE

Un'ulteriore occasione per mettere in pratica ed evolvere le esperienze maturate riguarda l'idea progettuale per la valorizzazione storico-culturale del patrimonio materiale e immateriale del Parco Regionale Naturale di Molentargius-Saline. Quella delle ex saline di Molentargius è una zona umida situata nell'area metropolitana di Cagliari, tra le più importanti d'Europa ai sensi della Convenzione di Ramsar (1971), dove nidificano stabilmente numerose colonie di fenicotteri. Con la L.R. 5 del 1999 è stato istituito il Parco Naturale Regionale Molentargius-Saline (<https://www.parcomolentargius.it>), uno straordinario compendio naturalistico frequentato tanto dai turisti quanto dai residenti.
Il Parco è però anche un complesso storico-industriale sfruttato dall'uomo per millenni: dal Neolitico fino agli anni Ottanta del Novecento è stato un luogo di produzione e raccolta del sale marino e dei suoi derivati. Le saline furono proprietà demaniale fin dal Medioevo e divennero Monopolio di Stato nel Settecento. Fino al 1985, anno della dismissione industriale, le Saline di Stato furono un centro di competenze e saperi di alto livello, producendo sale marino in numerose formulazioni. Di questo stratificato patrimonio industriale restano oggi molte vestigia: alcune recuperate, altre in fase di recupero, altre ancora da valorizzare anche in funzione di nuove attività economiche.
Per la ricchezza della sua storia industriale, testimoniata dai monumenti e dalla documentazione conservata soprattutto presso l'Archivio di Stato di Cagliari, la valorizzazione degli aspetti storico-culturali e sociali legati alla natura del Parco potrebbe offrire un valore aggiunto alla già notevole attrattività turistica, affidando all'Ente un ruolo importante nella trasmissione della storia istituzionale, culturale, industriale e sociale del territorio. È quindi necessario che questo immenso patrimonio materiale e immateriale sia reso accessibile al grande pubblico anche attraverso tecnologie innovative, capaci di offrire al visitatore un'immersione completa nella storia, raccontata in modo coinvolgente.
In questa prospettiva, e grazie all'esperienza del CNR-ISEM e alla disponibilità del fondo archivistico delle Saline di Stato, nel 2021 l'Ente Parco ha stipulato convenzioni con i due istituti culturali e finanziato sia il riordino della documentazione sia una ricerca biennale che ha dimostrato la sostenibilità della

valorizzazione storico-culturale come valore aggiunto all'offerta turistica del compendio. Tale valorizzazione permetterebbe di trasformare il Parco in un attrattore culturale oltre che naturalistico, fungendo da porta d'ingresso e vetrina per le ricchezze dell'intera isola e da volano economico legato alla gestione e manutenzione delle infrastrutture.
Il CNR-ISEM, in collaborazione con l'ASC, ha quindi coordinato un gruppo di esperti del Centro di Ricerca, Sviluppo e Studi Superiori in Sardegna (CRS4) e dell'Università di Cagliari. Con il convegno La valorizzazione del Parco Regionale Naturale di Molentargius-Saline: l'offerta di servizi e contenuti nell'ambito del turismo storico-culturale (11 ottobre 2024), il gruppo di lavoro ha trasformato l'idea in una bozza progettuale; la Regione Sardegna ha per ora finanziato la progettazione esecutiva.
La proposta — per la quale è stata valutata anche la sostenibilità economica — prevede l'allestimento di un percorso che faccia rivivere l'attività estrattiva nelle sue diverse fasi storiche, a partire dall'edificio Sali Potassici. L'attuazione del progetto è prevista in tre anni: i ricercatori del CNR-ISEM e delle Università di Cagliari e di Salerno svilupperanno le ricerche storiche e i contenuti, che saranno poi sceneggiati da un regista multimediale, mentre gli informatici del CRS4 realizzeranno software e hardware per gli allestimenti.

BIBLIOGRAFIA

Antinucci, F., Cultraro, M., Sabbioni, C., Siano, S. (2010), Le tecnologie del CNR per i beni culturali, Roma, CNR.

Antinucci, F. (2014). Comunicare nel Museo, Bari, Laterza.

Ciletti, M. (2025). Retrieval-augmented generation systems for enhanced access to digital archives. In S. Rebora, M. Rospocher, & S. Bazzaco (Eds.), Diversity, Equity, and Inclusion: Challenges and Opportunities for Digital Humanities in the Age of Artificial Intelligence: Proceedings del XIV Convegno Annuale AIUCD (Verona, 11–13 giugno 2025) (pp. 149–155). Associazione per l'Informatica Umanistica e la Cultura Digitale (AIUCD).

Cultraro, M. (2022). Un Mondo di Mezzo? L'esperienza narrativa dei Musei Virtuali tra reale e immaginario, in Arizza M. (ed.), Trattamento e Restituzione del Patrimonio Culturale, Roma, CNR Edizioni, pp. 275-286.

Delli Noci, F., Gabellone, F., Giannotta, M. T., Lombardo, M., Matarrese, M. (2001). Beni Culturali, impresa turistica e nuove tecnologie: strumenti e percorsi tra conoscenza e innovazione territoriale, in "Le reti di innovazione e lo sviluppo territoriale. Analisi di una esperienza: il progetto LINK", Atti dei lavori 16/17 gennaio, Città di Castello, Scuola Sup. S. Anna – Pisa, pp. 233-239.

Di Marcantonio, G. (2024). Artificial intelligence, LLM and RAG supporting library and information science: Challenges and opportunities. In Conferenza GARR 2024 – Navigare la complessità – Selected Papers (pp. 24–28). Consortium GARR. https://doi.org/10.26314/GARR-Conf24-proceedings-04

Gallinari, L. (ed.). 2002. Descubrir el Levante por el Poniente. I viaggi e le esplorazioni attraverso le collezioni della Biblioteca Universitaria di Cagliari. Atti del Convegno internazionale di studi (Villanovaforru, 1 dicembre 2001), Cagliari, Istituto sui rapporti italo iberici del CNR.

Mele, M. G. R., Serra, L., & Serreli, G. (2015). Coast View: sulla rotta di Marco Antonio Camos. RiMe. Rivista dell'Istituto di Storia dell'Europa Mediterranea, 14, 205–215. https://doi.org/10.7410/1168

Mele, M. G. R. (2020), Diamo la parola a personaggi e monumenti, in Bertini, A., Colesanti, G. T., Caruso, I., Vitolo, T., Cultura in transito. Ricerca e tecnologie per il Patrimonio Culturale, Roma, L'ERMA di Bretschneider, pp. . ISBN: 9788891320827.

Mele, M.G.R. (2020). Diamo la parola a personaggi e monumenti, in Cultura in transito. Ricerca e tecnologie per il patrimonio culturale, A. Bertini, I. Caruso, G.T. Colesanti, T. Vitolo (eds.), Roma Bristol, L'Erma di Bretschneider, pp. 91-105;

Serra, L. (2016). A document of the XVI century about the coastal defense of the Kingdom of Sardinia and a proposal for its multimedia development: Coast View with Google™. In G. Verdiani (Ed.), Defensive architecture of the Mediterranean XV to XVIII centuries: Proceedings of the International Conference on Modern Age Fortifications of the Mediterranean Coast (FortMed 2016), Vol. IV (pp. 309–316). DIDAPRESS/Dipartimento di Architettura, Università degli Studi di Firenze. ISBN 978-8896080603

Serra, L. (2020). Camós e Coast View con Google: strumenti tecnologici per il "trasferimento umanistico", in Cultura in transito. Ricerca e tecnologie per il patrimonio culturale, A. Bertini, I. Caruso, G.T. Colesanti, T. Vitolo (eds.), Roma Bristol, L'Erma di Bretschneider, pp. 107-121.

Serreli, G. (ed.). 2002. MudA. Museo Multimediale del regno di Arborèa. Catalogo, Las Plassas, MudA.

Boschetti, F. (2008). Alignment of Variant Readings for Linkage of Multiple Annotations. Proceedings of the ECAL 2007 Electronic Corpora of Ancient Languages, Prague 16–17 November 2007. Zemánek, P. (Ed.). 11-24. http://usj.ff.cuni.cz/system/files/Boschetti-Ch-2007.pdf.

Buzzetti, D., & McGann, J. (2006). Electronic Textual Editing: Critical Editing in a Digital Horizon. Electronic Textual Editing. Burnard L. et al. (Eds.), Modern Language Association of America. http://www.tei-c.org/About/Archive_new/ETE/Preview/mcgann.xml.

Ciula, A., & Stella, F. 2007. Digital Philology and Medieval Texts. Pisa: Pacini Editore.

Del Grosso, A. M., Capizzi, E., Cristofaro, S., De Luca, M. R., Giovannetti, E., Marchi, S., Seminara, G. & Spampinato, D. (2019, dicembre). Bellini's Correspondence: A Digital Scholarly Edition for a Multimedia Museum. Umanistica Digitale, 7, 23-47.

Del Grosso, A. M., Capizzi, E., Cristofaro, S., Seminara, G., & Spampinato, D. (2019). Promoting Bellini's Legacy and the Italian Opera by Scholarly Digital Editing His Own Correspondence. Poster presented at the TEI Conference and Member's Meeting. What is Text, really? TEI and beyond. Graz, Austria. https://zenodo.org/badge/DOI/10.5281/zenodo.3461673.svg.

Del Grosso, A. M., & Spampinato, D. (Eds.). 2023. Bellini Digital Correspondence. CNR Edizioni, 2023. ISBN: 978-88-8080-562-5.

Mastandrea, P. (2015). Archivi Elettronici Di Poesia Latina e Opzioni Multiple Di Ricerca Intertestuale. Semicerchio, 53, 60–69.

Pierazzo, E. (2015). Digital Scholarly Editing: Theories, Models and Methods. Farnham, Surrey: Ashgate.

Salvatori, E., & Simi, M. (2010). Beni Culturali e Ambienti Virtuali: Sfide e Problemi Di Un Passato a Tre Dimensioni. Memoria e Ricerca, 34, 171–86.

Vagionakis, Irene (2021), Cretan Institutional Inscriptions Dataset, ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli", National Research Council, in Pisa, http://hdl.handle.net/20.500.11752/OPEN-548

La Digital Library e l'Ecosistema digitale per la cultura come opportunità di *public engagement*

FEDERICO MESCHINI[1], BIANCAMARIA HERMANIN[2], ANTONELLA NEGRI[3], GIULIANO ROMALLI[4]

[1] UNIVERSITÀ DEGLI STUDI DELLA TUSCIA, ITALIA - FMESCHINI@UNITUS.IT–
[2] ISTITUTO CENTRALE PER LA DIGITALIZZAZIONE DEL PATRIMONIO CULTURALE – DIGITAL LIBRARY, ITALIA - BIANCAMARIA.HERMANINDEREICHENFELD@CULTURA.GOV.IT–
[3] ISTITUTO CENTRALE PER LA DIGITALIZZAZIONE DEL PATRIMONIO CULTURALE – DIGITAL LIBRARY, ITALIA – ANTONELLA.NEGRI@CULTURA.GOV.IT
[4] ISTITUTO CENTRALE PER LA DIGITALIZZAZIONE DEL PATRIMONIO CULTURALE – DIGITAL LIBRARY, ITALIA – GIULIANO.ROMALLI@CULTURA.GOV.IT

ABSTRACT (ITALIANO)

Il presente contributo esplora le criticità presenti nella realizzazione di una biblioteca digitale che intende proporsi come un ambiente collaborativo e distribuito, analizzandone la natura di sistema complesso in cui dimensioni tecnologiche, culturali e sociali si intrecciano in modo dinamico. Partendo dalla distinzione tra complessità essenziale e accidentale, l'articolo evidenzia come la modellazione del dominio di conoscenza sia il fulcro per superare le criticità dei sistemi informativi tradizionali. In questo scenario viene presentato Ecomic, l'Ecosistema digitale per la cultura promosso dall'Istituto Centrale per la Digitalizzazione del Patrimonio Culturale, concepito come un'infrastruttura abilitante che trasforma il dato in servizio per favorire il *public engagement*. L'architettura si fonda sull'integrazione di piattaforme dedicate alla digitalizzazione, gestione e analisi dei dati, ponendo al centro i grafi di conoscenza di dominio e cross-dominio. Questi strumenti permettono di superare la frammentazione tra archivi, biblioteche e musei, creando una rete semantica relazionale che abilita nuove forme di partecipazione cognitiva. Attraverso l'uso dell'intelligenza artificiale per l'arricchimento dei dati e dei modelli di cooperazione applicativa, Ecomic promuove un accesso collettivo al patrimonio culturale. In conclusione, il lavoro ribadisce come la centralità della modellazione dei dati e l'integrazione cross-dominio rappresentino un avanzamento cruciale nelle *Digital Humanities*, capace di generare valore conoscitivo inedito attraverso la connessione di risorse storicamente isolate, massimizzando così il riuso virtuoso dell'informazione culturale.
Parole chiave: biblioteca digitale; grafi di conoscenza; partecipazione pubblica

ABSTRACT (ENGLISH)

The Digital Library and the Digital Ecosystem for Culture as Opportunities for Public Engagement
This contribution illustrates the critical issues present in the creation of a digital library system that intends to present itself as a collaborative and distributed environment, analyzing its nature as a complex system in which technological, cultural and social dimensions intertwine dynamically. Starting from the distinction between essential and accidental complexity, the article highlights how knowledge domain modeling is central to overcoming the limitations of traditional information systems. In this context, Ecomic, the Digital Ecosystem for Culture promoted by the National Institute for the Digitization of Cutural Heritage, is presented as an enabling infrastructure designed to transform data into a service to foster public engagement. The architecture is based on the integration of platforms dedicated to data digitization, management, and analysis, with a focus on domain and cross-domain knowledge graphs. These tools overcome the fragmentation between archives, libraries, and museums, creating a relational semantic network that enables new forms of cognitive participation. Through the use of artificial intelligence for data enrichment and application cooperation models, Ecomic promotes non-hierarchical access to cultural heritage. In conclusion, the work reiterates how the centrality of data modeling and cross-domain integration represent a crucial advancement in Digital Humanities, capable of generating unprecedented cognitive value by connecting historically isolated resources, thereby maximizing the virtuous reuse of cultural information.
Keywords: digital library; knowledge graph; public engagement

1. BIBLIOTECHE DIGITALI E COMPLESSITÀ

Nonostante il termine "biblioteca digitale" sia da oramai trent'anni progressivamente entrato sempre più frequentemente nel linguaggio comune, il suo utilizzo continua a nascondere una complessità sottostante dalla quale sembra impossibile sottrarsi. Ciò è indubbiamente dovuto ai fattori eterogenei che ne vanno a costituire la natura multidimensionale. Sebbene quest'ultima sia propria di un qualsiasi sistema informativo (Zachman, 1987), in questa particolare tipologia sono presenti dei tratti distintivi che rendono di fatto inattuabile la possibilità di una definizione chiusa, rendendo un problema aperto qualsiasi discussione sulla natura, le funzionalità e il ruolo di una biblioteca digitale (Gonçalves *et al*., 2004: 270-271). In tutto ciò ritroviamo la continua interazione dinamica e non lineare delle suddette componenti eterogenee, in quella tipica interconnessione che caratterizza i sistemi complessi in cui "everything depends on everything else" (Rickles *et al*., 2007): uno stesso risultato può essere raggiunto attraverso diverse modalità e al tempo stesso una stessa configurazione iniziale può produrre risultati diversi. La componente deterministica delle biblioteche digitali, in quanto architetture informatiche, tende a mettere in secondo piano la loro natura di sistemi complessi, sia a livello di processo – le scelte critiche effettuate ad ogni passo della loro realizzazione – sia di prodotto finale. Va ricordato come "All software construction involves essential tasks, the fashioning of the complex conceptual structures that compose the abstract software entity (Brooks, 1987: 10). La complessità accidentale della dimensione tecnologica, fattore tutt'altro che trascurabile e su cui torneremo in seguito, il più delle volte nasconde quella essenziale, relativa alla modellazione del dominio di conoscenza cui si fa riferimento nello sviluppare un particolare sistema informativo. Nel caso delle biblioteche digitali il principale punto di riferimento è l'ambito delle istituzioni della memoria e, nello specifico, quello biblioteconomico, caratterizzato a sua volta da una natura multidisciplinare in cui componenti tecniche, culturali e sociali s'intrecciano senza soluzione di continuità (Ranganathan, 1931). Ad un livello astratto, qualsiasi azione legata al dominio bibliografico, fosse pure lo sfogliare la pagina di un libro, implica nel mondo digitale un'interazione continua tra diversi livelli, tra cui la logica operativa, i dati strutturati e non strutturati e l'interfaccia utente, ed ognuno di questi aspetti afferisce a settori diversi delle discipline computazionali, come l'algoritmica e la programmazione, la modellazione e la gestione dei dati e l'interazione uomo-computer. Lo scopo ultimo di una biblioteca digitale è di essere un'architettura conoscitiva in grado di favorire l'accesso e la fruizione – nonché in prospettiva la conservazione – di oggetti informativi eterogenei. Date queste premesse, diventa perciò fondamentale l'utilizzo ad alto livello di modelli formali per gestire la complessità essenziale (Gonçalves *et al*., *Ibid.*; Candela & Castelli, 2009). Ciò serve a controbilanciare la complessità accidentale, che da un lato offre sempre maggiori possibilità e funzionalità, dall'altro però risulta essere altamente confusiva, andando a costituire un labirinto in continuo mutamento fatto di sigle e concetti a prima vista oscuri. Anche solo prendendo il *World Wide Web* come punto di riferimento, nei pochi decenni si è potuto assistere ad un'evoluzione inarrestabile che ha ciclicamente cambiato i paradigmi di questo ambiente digitale. Si è passati infatti da un web statico ad una sempre maggiore dinamicità e interoperabilità delle pagine web, divenute da documenti a vere e proprie applicazioni. Ciò è avvenuto tramite la programmazione sia lato client sia lato server strutturatasi mano a mano in architetture stratificate basate su di una pletora di tecnologie eterogenee, come i *framework* di pubblicazione basati sul paradigma *Model View Controller* a livello di pubblicazione e XML, i Web Services, il Semantic Web e i Linked Data a livello di dati, per arrivare alla recente diffusione delle reti neurali e dell'Intelligenza Artificiale. Tutto questo senza considerare le varie componenti specifiche di dominio caratterizzate da una complessità verticale, come ad esempio quella relativa alla metadatazione in ambito culturale (Riley, 2010).

2. INFRASTRUTTURE ED ENGAGEMENT

Sulla base di queste non semplici premesse, va considerato come la già citata evoluzione e diffusione del web abbia portato questo ambiente ad essere partecipato da un sempre maggior numero di persone. Ciò non poteva non coinvolgere anche le biblioteche digitali, che, va ricordato, nella maggior parte dei casi sono infrastrutture pubbliche e destinate al pubblico. Questa costante maggiore accessibilità a strumenti digitali sempre più avanzati ha determinato la necessità stringente di ridefinire ruoli e obiettivi della ricerca di ambito umanistico, evidenziando altresì le opportunità offerte dal coinvolgimento strutturale e mirato di un pubblico eterogeneo. Andando a considerare l'ambito delle *Digital Humanities*, infatti, il *public engagement* costituisce punto di intersezione tra ricerca scientifica e infrastrutture digitali,

trasformando le vecchie modalità di condivisione unidirezionali in nuove prospettive di partecipazione biunivoca. La creazione di nuovi ambienti ed ecosistemi digitali sta progressivamente rendendo sempre più permeabili i confini tra i settori della ricerca scientifica e il pubblico, specialistico e non, favorendo forme di partecipazione che vanno dal semplice accesso alle risorse all'interazione e alla co-produzione di nuovo sapere. In questo scenario emerge con particolare evidenza il ruolo delle infrastrutture digitali pubbliche che, garantendo standard condivisi, interoperabilità e sostenibilità a lungo termine, costituiscono le candidate privilegiate ad una partecipazione pubblica strutturata, consapevole e duratura. In tale quadro si inserisce l'operato dell'Istituto Centrale per la digitalizzazione del Patrimonio culturale-Digital Library che attraverso la creazione di Ecomic, l'Ecosistema digitale per la cultura, mira a fornire gli strumenti basilari per la creazione e co-creazione di nuove conoscenze partecipate e condivise. Il presente contributo propone pertanto una riflessione sul *public engagement* nelle *Digital Humanities* attraverso l'analisi delle opportunità offerte dalle infrastrutture realizzate da Digital Library grazie all'investimento PNRR M1C3 1.1 "Strategie e piattaforme digitali per il patrimonio culturale".

3. ECOMIC

Ecomic, l'ecosistema digitale per la cultura, è un ambiente digitale distribuito e collaborativo promosso dall'Istituto centrale per la digitalizzazione del patrimonio culturale – Digital Library a supporto della trasformazione digitale del patrimonio culturale (Cerullo & Negri, 2023; Cerullo *et al*., 2024). Il termine ecosistema digitale non è inteso come un semplice insieme di strumenti tecnologici o di infrastrutture digitali, ma richiama una rete dinamica e interconnessa di soggetti, dati, processi, tecnologie e pratiche, orientata alla generazione di valore pubblico condiviso. Questo concetto è strettamente legato alla logica del dato come servizio – *Data as a Service, Knowledge as a Service* - secondo la quale la cultura digitale diventa un ambiente generativo, capace di produrre nuova conoscenza, innovazione sociale, contenuti creativi e modelli educativi. Ecomic adotta e interpreta questo paradigma configurandosi come uno spazio digitale e organizzativo in cui convivono infrastrutture tecnologiche, standard comuni, dati e modelli di cooperazione tra soggetti diversi. Attraverso l'adozione di tecnologie interoperabili e la promozione di pratiche collaborative, Ecomic è un ecosistema capace di adattarsi, crescere e innovare nel tempo. Ecomic non impone un unico modello di gestione del patrimonio digitale, ma offre un contesto abilitante plurale, dove ciascun Attore – pubblico o privato, istituzionale o comunitario – può riconoscersi, partecipare e contribuire secondo le proprie competenze e missioni.
Al centro dell'ecosistema si trova un nucleo tecnologico che fornisce servizi e strumenti digitali abilitanti. Il primo ambiente di Ecomic è D.PaC (Digitalizzazione per il patrimonio culturale), la piattaforma che supporta in modo strutturato ed efficiente la gestione, il monitoraggio e il controllo delle campagne di digitalizzazione dei beni culturali. Gli asset digitali sono poi conferiti in I.PaC (Infrastruttura e servizi digitali per il patrimonio culturale), nucleo centrale dell'ecosistema, che integra il patrimonio culturale digitale e ne rende possibile la gestione avanzata e la piena valorizzazione. Infine, la piattaforma D.PaaS *(Data product as a service*) è lo spazio di sperimentazione dove analizzare dati culturali per creare nuovi elaborati, ottenuti attraverso processi di trasformazione dei dati stessi.
Cuore delle possibilità di disseminazione della conoscenza e del conseguente coinvolgimento civile degli utenti, chiamati ad essere attori dell'Ecosistema, è I.PaC, che attraverso i suoi diversi moduli consente una elaborazione e un arricchimento dei dati in grado di aumentarne il valore, nell'ottica di un miglioramento continuo. Infine, va sottolineato come la natura distribuita e collaborativa di Ecomic renda l'ecosistema inserito a buon diritto nelle maggiori iniziative europee, quali ad esempio il *Common European Data Space for Cultural Heritage*[1] e l'*European Collaborative Cloud for Cultural Heritage* (ECHOES)[2]. Grazie all'adozione di standard condivisi e modelli di governance aperti, Ecomic consente l'interoperabilità con l'infrastruttura di Europeana, dando così la possibilità ai soggetti aderenti di proiettare i propri asset digitali in una dimensione europea. In questo modo diventa possibile costruire uno spazio comune dove il patrimonio culturale può diventare un motore di innovazione sociale ed economica, superando i confini nazionali per raggiungere un pubblico globale.

[1] https://www.dataspace-culturalheritage.eu/ (cons. 20/04/2026)
[2] https://www.echoes-eccch.eu/ (cons. 20/04/2026)

4. I GRAFI DI CONOSCENZA DI DOMINIO E CROSS DOMINIO

All'interno di I.PaC, i servizi di arricchimento svolgono una funzione centrale nel trasformare le risorse digitali da elementi isolati a elementi attivi in un sistema di conoscenza relazionale, costituendo le condizioni infrastrutturali che rendono possibile l'interazione con il pubblico. Tali servizi operano su più livelli complementari. Il primo è quello tecnico, finalizzato a rendere fruibili le risorse digitali attraverso l'elaborazione automatizzata dei contenuti. Attraverso processi di trasformazione dei contenuti, una singola risorsa può essere adattata a modalità diverse di consultazione e impiego, costituendo una condizione preliminare per il *public engagement*, in quanto consente a pubblici diversi di accedere ai contenuti più appropriati sulla base di esigenze e competenze eterogenee. Su questo piano si innesta un secondo livello, di natura semantica, che consente di costruire vere e proprie reti di conoscenza mediante l'inserimento delle risorse in grafi di conoscenza che ne ampliano il contesto interpretativo. I beni digitali in tal modo non sono più descritti solo in relazione alle proprie caratteristiche interne, ma vengono collegati ad altre informazioni secondo le logiche semantiche proprie di ciascun ambito culturale: archivistico, bibliografico, museale, multimediale. Un'opera d'arte, in questo modo, sarà agevolmente collegata al suo autore, all'epoca storica in cui è stata prodotta, alla tecnica esecutiva, al luogo di conservazione etc. Allo stesso tempo, l'integrazione tra grafi appartenenti a domini diversi apre alla dimensione cross-dominio, un'innovazione senza precedenti nel panorama italiano della digitalizzazione culturale. Il grafo cross- dominio costituisce infatti lo strumento attraverso il quale archivi, biblioteche, musei e altri contesti culturali possono dialogare in un'unica infrastruttura semantica, superando la storica frammentazione disciplinare della conoscenza. In questa prospettiva i servizi di arricchimento e i grafi di conoscenza non si limitano a supportare la ricerca settoriale e specialistica, ma incidono direttamente sulle modalità di coinvolgimento dei pubblici. Il patrimonio culturale digitale diventa così uno spazio condiviso e dinamico nel quale produrre nuove forme di conoscenza attraverso l'accesso e il riuso delle risorse. l grafi di conoscenza, in questo modo, rendono visibili e utilizzabili le relazioni tra entità culturali, trasformando il patrimonio da insieme di risorse a sistema relazionale. Il coinvolgimento del pubblico non è quindi legato esclusivamente alla possibilità di accedere alle risorse, ma soprattutto alla possibilità di esplorarle, interrogarle e interpretare le relazioni tra entità culturali. Grazie ai grafi di dominio e soprattutto di cross-dominio, dunque, si creano forme di partecipazione cognitiva declinabili secondo differenti punti di vista e il collegamento tra le diverse entità consente di leggere il patrimonio secondo prospettive differenti, valorizzando la diversità dei pubblici e delle competenze e permettendo a utenti con livelli di conoscenza diversi di accedere al patrimonio per interpretarlo secondo traiettorie inedite.

I grafi di conoscenza sono strutture dati flessibili e navigabili, progettate per organizzare le informazioni in entità (nodi) e relazioni (archi), secondo una logica simile a quella con cui l'essere umano collega concetti, persone, luoghi, oggetti ed eventi. Ogni nodo rappresenta un'entità univoca del patrimonio culturale; ogni arco descrive il tipo di relazione che intercorre tra due entità. Questo approccio consente la ricostruzione di scenari complessi e favorisce la comprensione profonda delle connessioni culturali, storiche e concettuali tra le entità, offrendo un'esperienza di fruizione interdisciplinare e intermediale. È particolarmente utile per i sistemi di accesso e i portali che vogliono proporre percorsi tematici, territoriali o cronologici, mettendo in relazione oggetti apparentemente distanti ma semanticamente connessi. I sistemi possono interagire con i grafi di I.PaC attraverso due livelli di cooperazione applicativa.

Il Livello 1 (L1) consente ai sistemi informativi cooperanti con Ecomic (sistemi integrati, federati o di accesso) di interrogare, esplorare e recuperare porzioni di grafo tramite servizi di API o *widget*. La navigazione può avvenire nodo per nodo, oppure attraverso template di ricerca predefiniti, utili per mappare rapidamente le relazioni tra entità. È il modello di cooperazione ideale per chi desidera esplorare i grafi senza modificare i propri sistemi esistenti.

Il Livello 2 (L2) consente ai sistemi integrati o federati di scrivere in tempo reale sui grafi di dominio sottoscritti, contribuendo all'aggiornamento e all'arricchimento dei dati. Questo modello è adatto a chi intende partecipare attivamente alla costruzione e manutenzione del patrimonio informativo condiviso, con maggiore controllo e integrazione.

Le informazioni recuperate dai grafi possono essere rese disponibili tramite API, *widget*, o *dataset* personalizzati, da integrare nei portali degli enti o da utilizzare come input per *middleware* di ricerca. La visibilità e la protezione dei dati sono sempre regolate tramite "Profili di Visibilità" (PdV) e "Profili di Protezione" (PdP), configurabili dagli enti aderenti, nel rispetto dei criteri di sicurezza e di *governance* dell'infrastruttura. Questi grafi offrono ai sistemi aderenti a I.PaC la possibilità di interrogare, esplorare e

arricchire le conoscenze disponibili nel grafo, contribuendo attivamente a un processo condiviso di valorizzazione e arricchimento semantico del patrimonio. L'Intelligenza Artificiale supporta la pulizia, il miglioramento e l'arricchimento dei dati descrittivi presenti nel grafo di I.PaC. Le elaborazioni principali comprendono la riconciliazione con vocabolari controllati, la bonifica e normalizzazione di dati disomogenei, la clusterizzazione di entità duplicate, l'arricchimento dei profili descrittivi tramite nuovi attributi o relazioni. Tutti i dati generati dall'IA vengono sempre identificati come tali e non sovrascrivono i dati inseriti dal catalogatore, che rimangono invariati e disponibili.
La cooperazione dei sistemi informativi con Ecomic e con la sua infrastruttura centrale, I.PaC, permette l'accesso ai servizi di arricchimento dell'ecosistema. I sistemi informativi di diversi enti pubblici aderiscono a I.PaC per migliorare l'accessibilità del proprio patrimonio digitale e valorizzarlo tramite portali e siti istituzionali. Una volta certificata la cooperazione tra i sistemi esterni di produzione dei dati e l'infrastruttura I.PaC e sottoscritta la convenzione, è possibile, per i sistemi esterni, richiamare i servizi di *content processing* avanzato (CPA). Tra questi, è possibile menzionare alcuni casi d'uso collaudati. Ad esempio, l'applicazione dell'elaboratore di immagini basata su IA, permette di ottenere, in modo completamente automatizzato, l'identificazione del soggetto principale e delle ulteriori entità riconoscibili; la riconciliazione semantica delle entità estratte con vocabolari esterni; la descrizione testuale delle immagini, corredata dalla relativa riproduzione audio; suggerimenti sulle possibili relazioni con beni culturali analoghi o affini. Nel caso di un sistema informativo di una Regione cooperante con l'Ecosistema, l'applicazione di tale flusso (personalizzato per rispondere alla specifica esigenza dell'ente) ad esempio al bene digitale di un manoscritto miniato di fine Trecento, caratterizzato da preziose miniature iconografiche, ha generato l'opportunità di fruire di testi esplicativi delle scene bibliche e narrazioni audio che accompagnano la visualizzazione del manoscritto, aumentando il valore interpretativo del bene e amplificandone la fruibilità anche per pubblici con esigenze particolari (es. ipovedenti, dislessici, anziani). In un altro caso, l'applicazione dei servizi di CPA a immagini provenienti da campagne fotografiche dedicate ai teatri storici ha permesso il riconoscimento automatico di elementi architettonici e contestuali distintivi (es. scenografia, decorazioni, illuminazione, statue, lampadari, etc.) e la riconciliazione di entità e concetti significativi con il Nuovo Soggettario di Firenze[3], sottoposta successivamente alla validazione degli operatori del sistema di catalogazione. Le immagini pubblicate sul portale istituzionale della Regione sono state così arricchite da descrizioni complete e altamente informative, oltre che da *tag* coerenti e standardizzati che migliorano l'efficacia della navigazione e della ricerca. L'utilizzo dei servizi di I.PaC ha quindi contribuito in modo significativo al perseguimento degli obiettivi regionali di promozione e accessibilità del patrimonio, senza alcun investimento tecnologico aggiuntivo. I risultati dell'elaborazione avanzata sono confluiti anche nel grafo di conoscenza di dominio, contribuendo ad aumentare la base informativa per tutti i partecipanti all'ecosistema. Sempre nel contesto delle Regioni aderenti a I.PaC, l'utilizzo dei grafi della conoscenza ha aperto l'opportunità di offrire agli utenti dei siti istituzionali un'esperienza di fruizione avanzata, che permette non solo di consultare il patrimonio digitale della Regione, ma anche di esplorare e navigare entità correlate. Per esempio, l'interrogazione del grafo a partire dalle immagini di manifesti cinematografici ha permesso di estrarre una serie di risultati pertinenti provenienti da altri sistemi aderenti, come per esempio la colonna sonora del film (es. spartiti, registrazioni audio, copertine di dischi), risorse bibliografiche sull'autore del romanzo da cui è tratto il film, fotografie e materiali di scena del set cinematografico. È stato quindi possibile integrare una nuova sezione "Scopri di più" sui siti istituzionali per l'esplorazione di informazioni e beni correlati da parte degli utenti. Il caso dimostra la capacità di I.PaC di trasformare i beni digitali in risorse informative connesse, aprendo la strada a esperienze di fruizione trasversale e multidimensionale del patrimonio culturale.

5. CONCLUSIONI

Se nel primo paragrafo è stata data una definizione di biblioteca digitale come sistema inevitabilmente complesso – insieme per quanto possibile ad una spiegazione dei motivi sottostanti questa l'inevitabilità –, in quelli successivi è stata descritta la risposta insita nella progettazione e realizzazione di Ecomic relativa alla gestione e consapevolezza di suddetta complessità. È evidente come lo scopo principale sia quello di rendere disponibili all'utente finale contenuti culturali strutturati, facilmente fruibili e, soprattutto, riutilizzabili tramite funzionalità avanzate, così da massimizzare la partecipazione degli utenti nell'ottica del *public engagement* basato sulle infrastrutture digitali e mettere in moto dei circoli virtuosi

[3] https://thes.bncf.firenze.sbn.it/ (cons. 20.04.2026).

di riuso dell'informazione. Ciò spiega altresì il focus della relazione sui dati e sul relativo livello di modellazione, in quanto – sempre tenendo in considerazione l'aspetto sistemico e di interdipendenza di tali sistemi – giudicato come la parte più cruciale dell'infrastruttura, un concetto oramai centrale da diversi anni nella teoria e nella pratica delle *Digital Humanities* (Buzzetti, 2002). I dati, difatti, si collegano direttamente all'aspetto conoscitivo e, soprattutto, di potenziale interrelazione tra i diversi domini della conoscenza, aspetto quest'ultimo alla base del concetto stesso di ipertesto (Bush, 1945), e va da sé come quello di biblioteca digitale in un ambiente *web* ne costituisca una specializzazione. Nello specifico e per concludere, la parte relativa al grafo di conoscenza cross-dominio costituisce un concreto passo in avanti nel superare la frammentazione attualmente presente nelle istituzioni della memoria, in quanto integra archivi, biblioteche e musei in una ricca rete di relazioni e rende il loro patrimonio esplorabile attraverso una ampia rete di relazioni. Si vanno così a generare quei vantaggi conoscitivi di cui parlava Lyotard alla fine del XX secolo, grazie alla creazione di collegamenti tra dati ritenuti fino a quel momento indipendenti (Lyotard, 1979: 85).

BIBLIOGRAFIA

Brooks, F. P. Jr. (1987). No Silver Bullet – Essence and Accident in Software Engineering. Computer, 20(4), 10-19, doi: 10.1109/MC.1987.1663532.

Bush, V. (1945). As We May Think. The Atlantic Monthly, July 1945.

Buzzetti, D. (2002). Digital Representation and the Text Model. New Literary History 33(1), 61-88, <http://www.jstor.org/stable/20057710>.

Candela, L., & Castelli, D. (2009). Una teoria fondazionale per le Biblioteche Digitali: il DELOS Digital Library Reference Model. Digitalia 4(1), 44-82.

Cerullo, L., & Negri, A. (2023). L'infrastruttura software per il patrimonio culturale (ISPC) come abilitatore di un Ecosistema digitale nazionale del patrimonio culturale. DigItalia, 18(1), 38–50, doi: 10.36181/digitalia-00059.

Cerullo, L., Bartoli, M., Coppola, L. A., Landino, C., Madonna, A. D., Negri, A., Pescarmona, G., Fauda Pichet , C., Porena, M., & Rossetti, V. (2024). Verso la creazione di un Ecosistema digitale nazionale per la Cultura. DigItalia, 19(2), 11–48, doi: 10.36181/digitalia-00101.

Gonçalves, M. A., Fox, E. A., Watson, L. T., & Kipp, N. A.(2004) Streams, Structures, Spaces, Scenarios, Societies (5S): A Formal Model for Digital Libraries. ACM Transactions on Information Systems, 22(2), 270-312, doi: 10.1145/984321.984325.

Lyotard, J. F. (1979). La condition postmoderne. Rapport sur le savoir. Paris: Les Éditions de Minuit.

Rickles D., Hawe, P., & Shiell, A. (2007). A simple guide to chaos and complexity. Journal of Epidemiology and Community Health, 61(11), 933–937, doi: 10.1136/jech.2006.054254.

Riley, J. (2010). Seeing Standards. Jenn Riley Digital Librarian. <http://jennriley.com/metadatamap/>.

Ranganathan, S. R. (1931). The five laws of library science. Madras: The Madras Library Association.

Zachman, J. A. (1987). A framework for information systems architecture. IBM Systems Journal, 26(3), 454-70.

Strategie di engagement nella creazione e sostenibilità di un parco letterario: i risultati del progetto CHANGES

SARA OBBISO[1], NICOLA MARINIELLO[2], DAVIDE BAGNARESI[3], RICCARDO STRACUZZI[4], GIULIANA BENVENUTI[5], ALESSANDRO IANNUCCI[6]

[1]ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA, ITALIA - SARA.OBBISO2@UNIBO.IT
[2]ENGINEERING INGEGNERIA INFORMATICA S.P.A. - NICOLA.MARINIELLO@ENG.IT
[3]ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA, ITALIA - DAVIDE.BAGNARESI2@UNIBO.IT
[4]ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA, ITALIA - RICCARDO.STRACUZZI@UNIBO.IT
[5]ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA, ITALIA - GIULIANA.BENVENUTI2@UNIBO.IT
[6]ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA, ITALIA - ALESSANDRO.IANNUCCI@UNIBO.IT

ABSTRACT (ITALIANO)

Il progetto CHANGES (2022-2025) ha offerto l'occasione per sperimentare un approccio innovativo alla valorizzazione digitale del patrimonio culturale nel contesto dei "Parchi Letterari", grazie alla progettazione di tecnologie immersive e alla co-creazione di narrazioni e percorsi attraverso diverse strategie di *public engagement*. Il coinvolgimento di numerosi e diversificati attori ha richiesto un grande sforzo di collaborazione tra università, aziende tecnologiche, amministrazioni locali e società civile. Ciò ha portato ad affrontare molteplici sfide legate alla gestione di *team* multidisciplinari o al divario digitale, ma nello stesso tempo ha consentito di individuare soluzioni concrete e condivise riguardo problematiche come l'accessibilità ai luoghi dei parchi e alle informazioni sul loro patrimonio culturale, l'inclusione o la sostenibilità del progetto nel tempo. Scopo del contributo è quello di presentare e discutere i risultati del lavoro svolto, nello specifico sui casi-studio dei Parchi Letterari "Grazia Deledda" a Galtellì e "Carlo Levi" ad Aliano, che si configurano come modelli simili ma metodologicamente distinti. I due approcci metodologici hanno visto la creazione di contenuti digitali e multimediali fruibili tramite *touchscreen*, *web app* e mappe interattive. In entrambi i casi l'integrazione di tecniche di acquisizione tradizionali (2D, fotogrammetria) e innovative (*Gaussian Splatting*, modellazione 3D immersiva, *tour* virtuali interattivi) ha permesso di creare esperienze coinvolgenti e accessibili sia *on-site* che da remoto. Gli obiettivi di scalabilità e sostenibilità del progetto nel tempo sono stati perseguiti attraverso la condivisione e l'utilizzo di buone pratiche, linee guida, strumenti *open-source* e principi FAIR. Di conseguenza, il progetto può porsi come modello replicabile per la valorizzazione digitale del patrimonio culturale e letterario, con prospettive future di estensione ad altri contesti museali simili.
Parole chiave: Parchi Letterari; tecnologie immersive; co-creazione; sostenibilità; inclusione

ABSTRACT (ENGLISH)

Engagement strategies in the creation and sustainability of a literary park: the results of the CHANGES project.
The CHANGES project (2022–2025) provided an opportunity to experiment with an innovative approach to the digital enhancement of cultural heritage in the context of "Literary Parks," through the design of immersive technologies and the co-creation of narratives and routes using various public engagement strategies. The involvement of numerous and diverse stakeholders required a significant collaborative effort among universities, technology companies, local administrations, and civil society. This led to addressing multiple challenges related to managing multidisciplinary teams and the digital divide, while at the same time enabling the identification of concrete, shared solutions to issues such as accessibility to park sites and to information about their cultural heritage, inclusion, or the long-term sustainability of the project. The aim of this contribution is to present and discuss the results of the work carried out, specifically focusing on the case studies of the "Grazia Deledda" Literary Park in Galtellì and the "Carlo Levi" Literary Park in Aliano, which can be seen as similar models yet methodologically distinct. Both methodological approaches involved the creation of digital and multimedia content accessible via touchscreens, web apps, and interactive maps. The integration of traditional acquisition techniques (2D, photogrammetry) and innovative ones (Gaussian Splatting, immersive 3D modeling, interactive virtual tours) made it possible to create engaging and accessible experiences both on-site and remotely. The goals of scalability and long-term sustainability were pursued through the sharing and use of best practices, guidelines, open-source tools, and FAIR principles. Consequently, the project can serve as a

replicable model for the digital enhancement of cultural and literary heritage, with future prospects for extension to other similar museum contexts.

Keywords:
Literary Parks; immersive technologies; co-creation; sustainability; inclusion

1. LA CO-CREAZIONE E VALORIZZAZIONE DIGITALE DEI PARCHI LETTERARI NEGLI OBIETTIVI DEL PROGETTO CHANGES TRA ACCADEMIA, INDUSTRIA E COMUNITÀ LOCALI

L'idea di fondare dei parchi letterari in Italia nasce prima di tutto come progetto editoriale già verso la fine degli anni '80 grazie all'iniziativa di Stanislao Nievo (Nievo, 1990), che intendeva preservare la memoria di Castello di Colloredo in Friuli, dove il suo avo Ippolito Nievo scrisse le *Confessioni di un italiano*. Successivamente, la sua concreta realizzazione come strategia di sviluppo territoriale locale, rivolta principalmente ad aree interne e marginali, ha preso forma grazie all'impulso dell'Unione Europea verso la fine degli anni '90, finché nell'estate 2009 la costituzione e il coordinamento dei parchi letterari furono affidati alla società Paesaggio Culturale Italiano S.r.l., creata per promuovere i parchi e le iniziative correlate per una loro attiva e vitale fruizione, all'interno di una rete diffusa in Italia e in Norvegia (De Marsanich, 2009). L'obiettivo della costituzione dei Parchi Letterari® fu dunque quello di incentivare un turismo specifico – di tipo culturale, sostenibile e responsabile – in grado di offrire opportunità concrete di sviluppo alle comunità e imprese locali all'interno di una rete nazionale funzionale ed efficiente.

Nel 2022 la valorizzazione attraverso tecnologie digitali dei Parchi Letterari di Aliano (MT) e Galtellì (NU) è rientrata come caso di studio negli obiettivi del Partenariato esteso CHANGES (*Cultural Heritage Active Innovation for Nex-Gen Sustainable Society*), in particolare all'interno dello *Spoke* 4, gruppo di ricerca guidato dall'Università di Bologna e orientato all'area tematica delle "Virtual technologies for museums and art collections" (Genovese et al., 2025). Tale Spoke ha preso in esame diverse tipologie di musei tra cui, appunto, anche il caso-studio dei Parchi Letterari, individuati all'interno del *type D: sites museums with tangible and intangible heritage and landscapes* (Bagnaresi et al., 2025).

Il lavoro dello Spoke 4, diviso a sua volta in cinque sottogruppi per la gestione di cinque ben distinti *work packages* (dati, stato dell'arte, linee guide di acquisizione, implementazione e disseminazione), è stato orientato verso un approccio partecipativo alla gestione di iniziative culturali, basato sull'idea che i progetti culturali possano essere co-progettati e valutati con il coinvolgimento attivo di *stakeholders*, comunità e pubblico (Piber et al., 2018). In particolare, il coinvolgimento, e di conseguenza la collaborazione, si è fondato su quattro componenti della società: *Academia*, costituita dai ricercatori e le ricercatrici delle Università partner; *Industry*, di cui hanno fatto parte alcune aziende partner e altre esterne vincitrici di bandi a cascata; *Government*, nel nostro caso le amministrazioni locali dei due Parchi Letterari coinvolti; *Civil Society*, tra cui i cittadini dei due paesi, partecipanti attivi nella narrazione dei loro luoghi.

Un parco letterario può essere pensato come un particolare tipo di museo diffuso, non limitato in un preciso perimetro ma esteso su un'area geografica e culturale ampia, corrispondente ai luoghi di un autore e/o della sua opera letteraria (Gouchan, 2023). È la voce letteraria, infatti, la chiave per attraversare i luoghi e il testo letterario diventa il principale motore del viaggio, nonché guida per itinerari turistici e base per la mappatura di beni e luoghi di interesse.

Più nello specifico, il Parco Letterario dedicato a Grazia Deledda (1871–1936) celebra la scrittrice nuorese, vincitrice del Premio Nobel per la Letteratura nel 1926. Le sue opere, profondamente radicate in Sardegna, esplorano temi come tradizione, famiglia, senso di colpa e rapporto con la natura, rendendola una figura centrale sia nella letteratura italiana sia in quella internazionale. Nel borgo di Galtellì, Deledda ambienta uno dei suoi romanzi più importanti, *Canne al vento* (1913), oltre a diversi racconti brevi dedicati alle tradizioni locali. Il Parco Letterario dedicato a Carlo Levi (1902–1975) rende omaggio al pittore, scrittore e medico torinese che, tra il 1935 e il 1936, fu confinato dal regime fascista in un piccolo paese della Lucania, Aliano. Da quell'esperienza nacque *Cristo si è fermato a Eboli* (1945), un'opera letteraria e antropologica che ha trasformato Aliano in un simbolo della resistenza al fascismo e dell'attenzione al tema della marginalizzazione del Mezzogiorno d'Italia.

2. PROCESSI E TECNICHE NELLA CREAZIONE DEI CONTENUTI DIGITALI

All'inizio del progetto i due parchi non utilizzavano o non disponevano di risorse tecnologiche o strumenti di fruizione digitale, una situazione che rendeva difficoltosa l'accessibilità alle informazioni e ad alcuni luoghi solitamente chiusi o faticosamente raggiungibili, come le rovine del Castello di Pontes a Galtellì o la casa-confino di Carlo Levi ad Aliano. Nell'ambito del progetto CHANGES l'obiettivo era dunque quello di rinnovare e migliorare la fruizione e l'accessibilità del patrimonio di questi due parchi, valorizzandone l'aspetto paesaggistico e culturale attraverso l'uso di nuove tecnologie digitali immersive, come l'uso di fotografie panoramiche annotate per costruire *virtual tours*, individuati come strumenti in grado di influenzare significativamente i flussi nel turismo culturale, migliorando il coinvolgimento e l'intenzione di visita (Sinha et al., 2025).

L'attività di sviluppo tecnologico per i Parchi Letterari è stata guidata dall'azienda Engineering S.p.A., in collaborazione con le aziende esterne per lo sviluppo dei software e l'acquisizione dei contenuti digitali, oltre che con un *team* multidisciplinare dell'Università di Bologna dedicato principalmente ad attività di *engagement* della comunità e ad aspetti di scrittura, traduzione e comunicazione delle informazioni sul patrimonio storico-culturale e letterario raccolto.[1] Altri *team* universitari sono stati coinvolti nel progetto: tra questi, il gruppo degli archeologi del Dipartimento di Storia e Culture (DISCI) dell'Università di Bologna, che hanno mappato e analizzato i più importanti siti archeologici del territorio attorno al paese di Galtellì, e un *team* dell'Università di Bari, che ha svolto laboratori educativi con le comunità dei due paesi, rivolti in particolare agli studenti, visti come piccoli *storyteller* coinvolti direttamente nella narrazione delle loro tradizioni familiari.[2]

Le attività di digitalizzazione hanno riguardato principalmente l'acquisizione di molteplici tipologie di contenuti: fotografie 2D, video acquisiti in alcuni casi anche con drone, audio, rilievi fotogrammetrici e scansioni per la creazione di modelli 3D di luoghi e oggetti, fotografie panoramiche 360° di tutti i punti di interesse, la produzione di *tour* virtuali per la loro esplorazione e la creazione di contenitori digitali per la loro fruizione. La creazione di alcuni modelli 3D è stata un'attività di particolare sperimentazione tecnica, poiché sono state eseguite sia tecniche tradizionali come *Surface Reconstruction* da *point cloud* ottenuti con rilievi *laser scan*, oppure tecniche di fotogrammetria classica. Sono state, inoltre, sperimentate nuove tecniche Nerf, che hanno tuttavia il limite di elevati tempi di elaborazione e hardware sofisticato, oppure ancora la più recente *Gaussian Splatting*, che offre il vantaggio di eliminare l'utilizzo di *laser scanner* e di non richiedere hardware specializzati, garantendo copertura e illuminazione uniforme, nonché un effetto particolarmente realistico dell'oggetto acquisito (Kerbl et al. 2023).[3] Per la visualizzazione e la gestione degli ambienti, dei modelli e delle ricostruzioni tridimensionali è stato scelto di utilizzare come *framework* di riferimento lo strumento *open-source* ATON, che ha reso possibile creare ambienti tridimensionali che replicassero i luoghi reali, e inoltre, a partire dai diversi ambienti immersivi, costruire i *tour* virtuali interattivi all'interno dei quali visualizzare i contenuti multimediali (Fanini et al. 2023).[4]

Tutti gli ambienti immersivi e i contenuti digitali sono stati integrati in un contenitore digitale web adatto ad essere fruito *on-site* tramite un monitor *touchscreen* in cui si trovano le mappe interattive, le gallerie multimediali, alcuni itinerari selezionati e i *QR code* che permettono di accedere ad applicazioni PWA per le visite guidate fruibili anche via *smartphone*.

[1] L'azienda esterna selezionata per il lavoro sul Parco di Galtellì è stata Risviel S.r.l., che si è servita della collaborazione di MammutFilm S.r.l. per la realizzazione delle video-interviste alla comunità. L'azienda vincitrice per il Parco di Aliano è risultata essere NAIS-Nextant Application Innovative Solutions S.r.l., in collaborazione con Forum Engineering S.r.l.. L'azienda Engineering S.p.A. si è occupata della supervisione tecnica e dell'integrazione multi-layer delle diverse componenti, come la gestione dei contenuti, l'interfaccia touchscreen e le mobile PWAs.

[2] I risultati di questo lavoro sono raggiungibili sulla piattaforma https://izi.travel/it, cercando i nomi dei due paesi nell'apposita barra di ricerca (ultimo accesso: 27 gennaio 2026).

[3] I processi di acquisizione, in particolare in riferimento agli oggetti tridimensionali, sono stati uniformati e ricondotti alle linee guida e alle best practices prodotte da un sub-team dello Spoke 4 chiamato WP3, a partire dal lavoro di ricerca sul digital twin della mostra di Ulisse Aldrovandi (Barzaghi et al., 2024).

[4] Tuttavia, alcune aziende come Risviel S.r.l. si sono discostate da questa linea di indirizzo, preferendo altri strumenti come la libreria open-source Three.js (Danchilla, 2012).

3. DUE SOLUZIONI METODOLOGICHE DISTINTE E COMPLEMENTARI

Nell'approccio metodologico le due aziende hanno adottato soluzioni parzialmente differenti in base alle specificità dei due luoghi, la cui analisi comparativa si può articolare attorno a quattro dimensioni:
Content delivery, che nel caso di Galtellì segue un modello *standalone*, quindi l'applicazione principale viene eseguita direttamente sul *touchscreen* garantendo un'ottima velocità di fruizione e contenuti in loco, ma in caso di necessità di aggiornamenti software o contenutistici è richiesto un intervento *on-site* per essere applicati; Aliano usufruisce, invece, di un modello *server-based*, che ha una fruizione vincolata maggiormente alla qualità della rete, ma tutti gli aggiornamenti software o contenutistici sono eseguiti da remoto e visualizzabili in tempo reale.
Stack tecnologico, che per Galtellì ha determinato la scelta di un sistema integrato, basato sulla piattaforma commerciale Wordpress con un plugin sviluppato *ad hoc* per aggiungere graficamente aree interattive e pulsanti multimediali alle foto 360°, chiamato GisDoc.[5] Il sistema è stato costruito per risultare il più possibile usabile e intuitivo nell'utilizzo sia dell'interfaccia *front-end* sia nel *back-end*, da cui è possibile accedere in un unico ambiente per la gestione autonoma di pressoché tutte le componenti del *touchscreen*, tuttavia con delle limitazioni nelle modifiche grafiche. Aliano, al contrario, usufruisce di una soluzione basata totalmente su strumenti *open source*, come ATON e Blender per la gestione delle immagini immersive e interattive. La struttura dei *tours*, i punti di interesse e i percorsi sono configurabili manualmente su file in formato *.json* e gli ambienti immersivi e i contenuti sono configurabili attraverso interfaccia web. Questa alternativa, quindi, da un lato presenta il vantaggio di essere totalmente aperta, ma dall'altro può risultare un po' meno accessibile ad utenti e operatori non specializzati, poiché richiede una certa conoscenza nella configurazione di file in formato .json, per poter eseguire modifiche di *tours* o dei contenuti multimediali.
Approccio alla modellazione 3D, applicato a Galtellì su larga scala, come nel caso del Castello di Pontes (sviluppato con tecnica Gaussian Splatting) e finalizzata ad una sua fruizione e accessibilità anche da remoto (Mariniello et al., 2025)[6] e della ricostruzione virtuale 3D (sviluppata con tecnica di modellazione tradizionale) di "Casa Nieddu", descritta nel romanzo deleddiano *Canne al vento* come "la casa delle dame Pintor", negli anni sottoposta a molti cambiamenti e suddivisioni di proprietà, nonché attualmente inaccessibile per motivi di sicurezza.[7] Per contro ad Aliano la modellazione si è focalizzata maggiormente su singoli oggetti o statue (ad esempio la statua di Carlo Levi nella piazza del paese) e su alcuni interni museali.
Utilizzo della gamification, che l'azienda Risviel a Galtellì ha trattato autonomamente dal resto, scegliendo di sviluppare insieme al resto del *team* un'applicazione apposita che sfrutta la presenza delle targhe letterarie già installate in paese, per incentivare un'esplorazione fisica dei luoghi attraverso un'attività ludica. Nel caso di Aliano, invece, il gioco si presenta all'interno di un'*app mobile* sottoforma di quiz integrato con uno dei due itinerari guidati di visita, quello che riguarda la storia culturale del paese.

4. PRATICHE E STRUMENTI DI COINVOLGIMENTO, ACCESSIBILITÀ E RIUSO DEL PATRIMONIO CULTURALE DEI PARCHI LETTERARI

In entrambi i casi di studio i due *team* hanno lavorato tanto da remoto quanto più volte sul campo a stretto contatto con la popolazione locale, tra cui lo staff delle amministrazioni locali, gli operatori degli spazi culturali come musei e biblioteche, i gestori di attività enogastronomiche, i membri delle associazioni, le guide escursionistiche o quelle dedicate a percorsi culturali e/o letterari e tanti altri

[5] La documentazione sul framework GisDoc è consultabile ai seguenti link: https://risviel.com/it/article/34 e https://parcodeledda.it/galtelli/site/it/ (ultimo accesso 27 gennaio 2026).
[6] Il modello è raggiungibile ed esplorabile tramite il seguente link: https://aton-changes.net:8083/container_mediacontents_gallery_galtelli/GALTELLI/3Dmodels/castello-pontes-splat.html (ultimo accesso 27 gennaio 2026).
[7] La ricostruzione, che costituisce una versione iniziale non definitiva, si è basata principalmente sull'analisi delle immagini dello sceneggiato televisivo Canne al vento, pubblicato dalla Rai nel 1958. Nella prima fase sono stati estratti e rielaborati fotogrammi significativi che fungono da riferimento documentario per l'elaborazione grafica preliminare. Successivamente, con Autodesk Revit è stata modellata la struttura architettonica dell'edificio acquisita tramite drone, prestando particolare attenzione agli elementi strutturali e spaziali. La fase successiva, effettuata con Blender, ha permesso di integrare i dettagli dei materiali e definire le texture, nonché inserire arredi e oggetti tipici della cultura sarda dei primi del '900, per garantire un contesto storico e ambientale coerente.

componenti della comunità di Aliano e Galtellì. Questa densa e costante attività di coinvolgimento ha consentito di sperimentare costi e benefici di un lungo processo collaborativo, la cui importanza, in quanto parte integrante della ricerca stessa, risiede nella possibilità di fondare nuove relazioni professionali e personali in cui università e comunità co-producono nuovi saperi e nuove modalità di comprensione e azione sul patrimonio culturale (Facer and Enright, 2016).
Il coinvolgimento della comunità o di altri stakeholders è avvenuto principalmente attraverso la mediazione dell'amministrazione comunale dei due paesi, attraverso *meetings* organizzati in diverse fasi durante tutta la durata del progetto. Le amministrazioni locali hanno così potuto partecipare costantemente alla supervisione generale dei lavori e al loro aggiornamento, avanzare dubbi o suggerimenti e intervenire nella modifica di alcuni contenuti, con la possibilità di prendere parte attiva al processo decisionale in modo condiviso.
La collaborazione è avvenuta, inoltre, nella messa a punto di strategie di accessibilità e inclusione ai punti di interesse e alle informazioni. Tra queste, ricordiamo:
La traduzione multilingua dei contenuti in italiano e in inglese, nonché la predisposizione per eventuali implementazioni future, compresa l'aggiunta di altre lingue e dialetti locali come l'alianese e il sardo-galtellinese;
L'arricchimento dei contenuti testuali con corrispettivi file audio;
La definizione di targhe con *QR code* da installare all'esterno dei principali punti di interesse, in modo da assicurarne la disponibilità delle informazioni, seppure in forma virtuale, nel caso in cui questi luoghi fossero trovati dal visitatore chiusi o inaccessibili. Le targhe sono provviste della traduzione in lingua inglese e nel sistema Braille. Quelle prodotte dall'azienda Risviel, inoltre, recano dei tag NFC (Near Field Communication), tecnologia di comunicazione wireless a corto raggio basata su campi elettromagnetici, che può consentire alle persone cieche di accedere a contenuti audio semplicemente avvicinando il proprio dispositivo smartphone attivo ad un tag NFC passivo (Ceipidor et al., 2013).
L'acquisizione con immagini 360° o la ricostruzione tridimensionale di luoghi difficilmente accessibili come il Castello di Pontes o la casa delle dame a Galtellì, oppure la casa-confino di Levi o le grotte e i calanchi ad Aliano.
Lo sviluppo di applicazioni PWA con percorsi guidati e contenuti testuali, audio, immagini 360° e riferimenti topografici, che il visitatore può portare sempre con sé durante il percorso.

Infine, le amministrazioni sono state coinvolte nel prendersi carico del prodotto finito, curandone aggiornamenti e manutenzione. Tale processo è stato agevolato da costanti attività di spiegazione del funzionamento degli hardware e dei software prodotti, nonché dal rilascio di specifici manuali d'uso (Mariniello 2025). In questo modo, tramite le proprie amministrazioni, la comunità dei due paesi è invitata a intervenire direttamente nella narrazione dei propri luoghi, a vantaggio di un arricchimento e aggiornamento continuo del senso e della memoria.

5. SFIDE E OPPORTUNITÀ, CRITICITÀ E STRATEGIE PER LA SOSTENIBILITÀ E REPLICABILITÀ DEL PROGETTO

Il progetto CHANGES applicato al caso dei Parchi Letterari ha messo in luce alcune difficoltà sfidanti: tra queste, ad esempio, il gap informatico e tecnologico che interessa i piccoli comuni di aree interne o marginali come i paesi di Aliano e Galtellì, dove manca spesso l'accesso ad una rete internet veloce e stabile.
Per venire incontro al divario in termini di competenze digitali si è cercato di progettare interfacce semplici e intuitive, con soluzioni come la suddivisione dei contenuti in sezioni ben definite, la categorizzazione delle informazioni e l'inserimento di immagini visivamente coinvolgenti e ben descritte, nonché di altri espedienti consolidati negli studi sulla *user experience* applicata agli strumenti di fruizione del patrimonio culturale (Potenziani et al. 2025).
I feedback riguardanti l'esperienza utente sono stati raccolti sia in modo informale sia con modalità formali e strutturate, in questo caso attraverso la somministrazione di venti questionari per ciascun Parco, basati sulle linee guida create ad uso interno dallo Spoke di progetto. Il questionario è stato suddiviso in diverse sezioni: la prima era finalizzata alla raccolta di dati significativi degli utenti (età, genere, livello di istruzione) e di informazioni sull'esperienza pregressa nell'uso di dispositivi quali i touchscreen, in particolare in contesti museali e/o turistici e di visita culturale (vd. Fig 1).

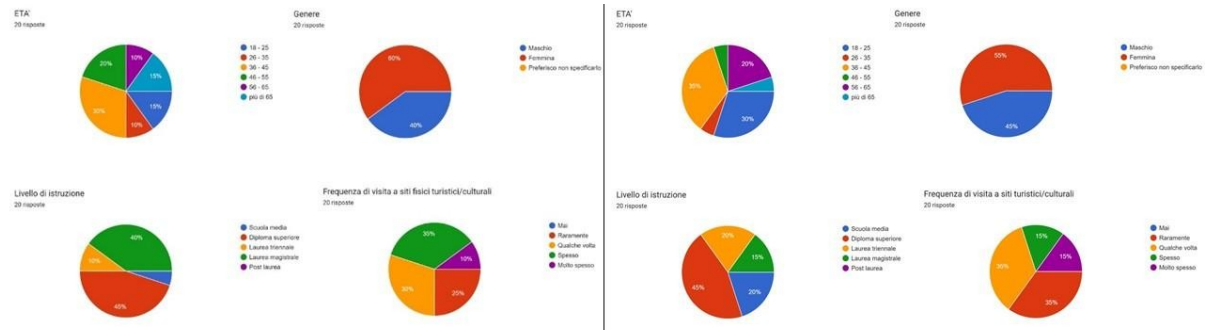


Figura 1: alcuni dati della sezione 1 del questionario svolto a Galtellì (sinistra) e Aliano (destra)

La seconda, incentrata sull'esperienza complessiva dell'utente rispetto ai prodotti proposti, affrontava aspetti quali l'accessibilità, la comprensibilità, la facilità d'uso, l'efficienza, la chiarezza, la capacità di suscitare interesse, l'innovazione e il loro potenziale di supporto (vd. Fig. 2).

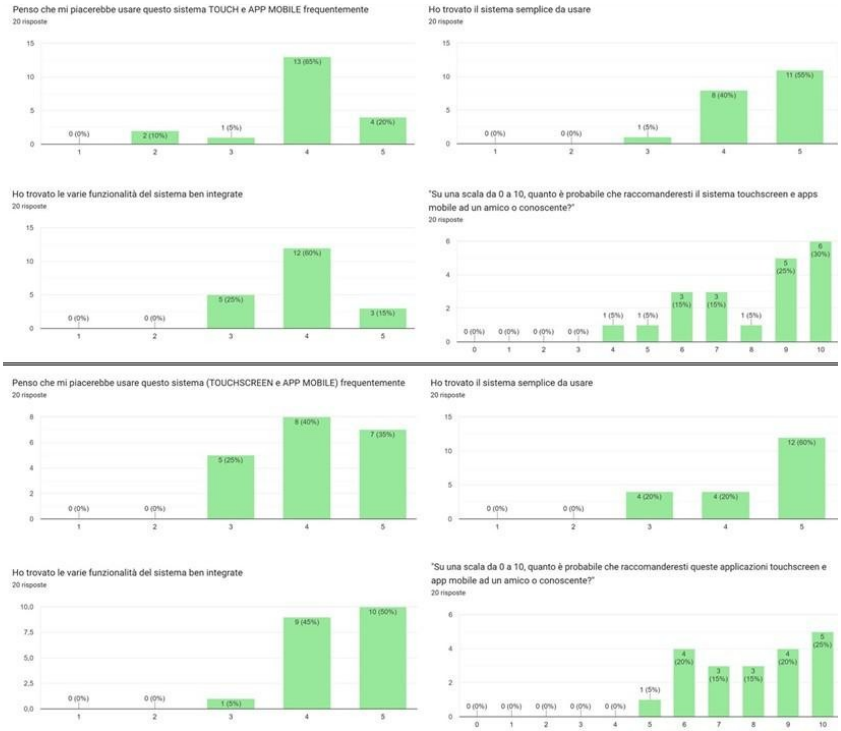


Figura 2: alcuni dati della sezione 2 del questionario svolto a Galtellì (sopra) e Aliano (sotto)

L'indagine comparativa basata su questi dati preliminari suggerisce che il sistema digitale ad Aliano potrebbe fungere da ponte per un pubblico ancora poco aduso a questo tipo di valorizzazione: con il 70% di visitatori che frequenta siti culturali solo "raramente" o "qualche volta" , l'elevata semplicità d'uso riscontrata nelle tecnologie proposte risulta fondamentale per abbattere le barriere d'accesso e rendere la visita inclusiva per i meno esperti; a Galtellì, invece, il sistema può essere maggiormente visto come strumento di approfondimento per un pubblico più colto e abituato e il più elevato numero di promotori (NPS 55%) suggerisce, inoltre, che gli utenti con un *background* più solido possano valorizzare maggiormente la tecnologia come complemento alla loro esperienza culturale, trasformandosi in fruitori attivi del parco.
Infine la terza sezione, dedicata alle specificità tecniche del *touchscreen* e delle applicazioni mobili, era volta a individuare per queste ultime eventuali criticità, aree di miglioramento e preferenze degli utenti attraverso domande a risposta aperta, che hanno permesso poi successivamente di costruire una preliminare analisi SWOT. È stata confermata l'efficacia delle soluzioni immersive, evidenziando punti di forza (usabilità, chiarezza narrativa, integrazione multimediale) e aree di miglioramento (ottimizzazione audio, qualità dei modelli 3D, design grafico). È emersa una generale risposta positiva all'integrazione di

sistemi *touchscreen* e *app mobile* in entrambi i contesti, sebbene con sfumature demografiche e di coinvolgimento differenti.
Una più completa analisi, derivante da ulteriori sessioni di raccolta feedback su UX, soddisfazione d'uso e valutazione complessiva dell'esperienza nel tempo, potrà rivelarsi utile non solo per le fasi di manutenzione e implementazione necessarie in futuro per la sostenibilità del progetto, ma anche in ottica di replicabilità in altri contesti simili, come altri parchi letterari, parchi culturali, siti archeologici o musei diffusi e a cielo aperto, dove il presente caso-studio potrà porsi come valido modello.
Il progetto, conclusosi nel mese di febbraio 2026, ha infine affrontato nella sua ultima fase i *deliverable* dedicati all'*exploitation*, all'analisi e alla disseminazione dei risultati, nonché al rilascio in *open source* delle informazioni sui materiali digitali e sui software sviluppati all'interno del progetto, comprensivi dei loro metadati e manuali. Quest'attività, che vedrà sviluppi aggiuntivi focalizzati particolarmente su questioni di diritto d'autore e riproducibilità digitale dei beni culturali, si fonda su quanto delineato nel *deliverable* dedicato al *Data Management Plan*, prodotto dal WP1 dello Spoke (Cavallo Perin et al., 2025): il documento invita a descrivere le categorie di dati prodotti nell'ambito del progetto e illustrarne la gestione in modo trasparente e in conformità ai principi FAIR della scienza aperta. Il rilascio dei materiali disponibili, la maggior parte dei quali in licenza aperta, è gestito tramite la piattaforma di condivisione Zenodo, dove saranno pubblicati man mano anche tutti i relativi aggiornamenti[8].

RINGRAZIAMENTI

Per il sostegno alle fasi operative del progetto si esprime un sentito ringraziamento alla Fondazione CHANGES, a tutto il *team* dello Spoke 4, al Presidente dell'Associazione Parchi Letterari® Stanislao De Marsanich e ai sindaci Franco Solinas del Comune di Galtellì e Luigi De Lorenzo del Comune di Aliano, insieme allo staff delle rispettive amministrazioni.

BIBLIOGRAFIA

Bagnaresi, D., Benvenuti G., Bruno, I., Capotorto, G., Iannucci, A., Inglese, A., Grinello, M., Mariniello, N., Marsili, G., Obbiso, S., Peroni, S., Pomero, M. E. & Stracuzzi, R. (2025). Implementation of pilot studies – Type D. Deliverable D14 v1.0, Project CHANGES - Spoke 4. Bologna: University of Bologna. doi:10.5281/zenodo.17238011.

Cavallo Perin, R., Gualandi, B., Paruzzo, F., & Peroni, S. (2025). Data Management Plan: Final version (Deliverable No. D10 v1.0; Project CHANGES - Spoke 4). Bologna: University of Bologna. https://doi.org/10.5281/zenodo.14955816.

Ceipidor, U.B., Medaglia, C.M., Volpi, V., Moroni, A., Sposato, S., Carboni, M. & Caridi, A. (2013). NFC technology applied to touristic-cultural field: A case study on an Italian museum. 5th International Workshop on Near Field Communication (NFC). Zurich: Switzerland, 2013, pp. 1-6, doi: 10.1109/NFC.2013.6482445.

Danchilla, B. (2012). Three.js Framework. In: Beginning WebGL for HTML5. Berkeley, CA: Apress. https://doi.org/10.1007/978-1-4302-3997-0_7.

De Marsanich, S. (2009). I parchi letterari, spazi geografici e suggestioni poetiche. Quaderni del '900, IX, 13-30. Roma-Pisa: Fabrizio Serra Editore.

Facer, K. & Enright B. (2016), Creating living knowledge. The connected community programme, community-university relationship and the participatory turn in the production of knowledge. Bristol: University of Bristol.

Fanini, B., Ferdani, D., Demetrescu, E., Berto, S., & d'Annibale, E. (2021). ATON: An Open-Source Framework for Creating Immersive, Collaborative and Liquid Web-Apps for Cultural Heritage. Applied Sciences, 11(22), 11062. https://doi.org/10.3390/app112211062.

Genovese, G., Heibi, I., Peroni, S., & Pescarin, S. (2025, luglio 17). Leveraging virtual technologies to enhance museums and art collections: insights from project CHANGES. DH2025 — the annual

[8] Per l'insieme dei materiali digitali prodotti per il Parco Letterario di Galtellì vd. Iannucci et al., 2026 e Iannucci, Mellara e Rossi; 2026; per il Parco Letterario di Aliano vd. Mariniello et al., 2026. Per quanto riguarda le componenti software del sistema prodotto per Galtellì vd. Iannucci e Sicilia, 2026 e Mariniello 2026; nel caso di Aliano vd. Mariniello, Dore e Modugno 2026. Infine per l'applicazione *game* sviluppata per Galtellì vd. Iannucci e Sicilia, 2026.

ADHO conference on "Building Access and Accessibility, Open Science to all Citizens" (DH 2025), Lisbon, Portugal. Zenodo. https://doi.org/10.5281/zenodo.16093001.

Gouchan, Y. (2023). Literary Parks. In R. Baleiro, G. Capecchi & J. Arcos-Pumarola (Eds.). E-Dictionary of Literary Tourism. University for Foreigners of Perugia.

Iannucci, A., Benvenuti, G., Obbiso, S., Mariniello, N., Stracuzzi, R., Bagnaresi, D., Giacomini, F., Ulderico, S., Santopuoli, N., Cattani, M., D'Angelo, F., Bonacini, E., & Corriero, M. (2026). CHANGES SPOKE 4 WP4 - Comprehensive Dataset of the Galtellì Literary Park [Data set]. Zenodo. https://doi.org/10.5281/zenodo.19108336

Iannucci, A., Mellara, M., & Rossi, A. (2026). CHANGES - SPOKE 4, WP4 - Galtellì Literary Park - MammutFilm audio&video [Data set]. Zenodo. https://doi.org/10.5281/zenodo.19108150

Iannucci, A., & Sicilia, U. (2026). CHANGES SPOKE 4 WP4 - Galtellì Literary Park - risviel gisdoc core. Zenodo. https://doi.org/10.5281/zenodo.19068459.

Iannucci, A., & Ulderico, S. (2026). CHANGES SPOKE 4 WP4 - Galtellì Literary Park - risviel game. Zenodo. https://doi.org/10.5281/zenodo.19108556.

Kerbl, B., Kopanas, G., Leimkühler, T., & Drettakis, G. (2023). 3D Gaussian Splatting for Real-Time Radiance Field Rendering, Published in ACM Transactions on Graphics 26 July 2023, Computer Science.

Mariniello N. (2025). CHANGES - SPOKE 4, WP4, Costruzione tours immersivi (D3.8.b - Manuale di costruzione TOURS immersivi con Aton). Zenodo. https://doi.org/10.5281/zenodo.19569288.

Mariniello N., Giacomini F., Obbiso S., Sicilia F., Sanna A., Iannucci A. & Benvenuti G. (2025). Digital technologies for the ''Grazia Deledda'' Literary Park in Galtellì (NU): the 3D virtual reconstruction of Pontes Castle. The Eurographics Association. https://doi.org/10.2312/dh.20253033.

Mariniello, N. (2026). CHANGES - SPOKE 4, WP4 - Galtellì Literary Park - ENG software (1.0). Zenodo. https://doi.org/10.5281/zenodo.19571951.

Mariniello, N., Benvenuti, G., Iannucci, A., Bagnaresi, D., Stracuzzi, R., Obbiso, S., Monteleone, A., Dore, N., Gattulli, V., Corriero, M., & Bonacini, E. (2026). CHANGES - SPOKE 4, WP4, Comprehensive Dataset of the Aliano Literary Park [Data set]. Zenodo. https://doi.org/10.5281/zenodo.18504899

Mariniello, N., Dore, N., & Modugno, M. (2026). CHANGES - SPOKE 4, WP4 - Aliano Literary Park - software (1.0). Zenodo. https://doi.org/10.5281/zenodo.19144745.

Nievo, S. (1990). I parchi letterari. Roma: Abete.

Piber M., Demartini P. & Biondi L. (2018). The management of partecipatory cultural initiatives: learning from the discourses on intellectual capital, "Journal of management and governance", v. 23: 435-458.

Potenziani, M., Callieri, M. & Cignoni, P. (2025). Designing and Developing Interactive Kiosks for Cultural Heritage: Best Practices and Common Issues. In: Liarokapis, F., Shehade, M., Aristidou, A., Chrysanthou, Y. (eds) Interactive Media for Cultural Heritage. Springer Series on Cultural Computing. Springer, Cham. https://doi.org/10.1007/978-3-031-61018-9_21.

Sebastian, B., Bordignon, A., Collina, F., Fabbri, F., Fanini, B., Ferdani, D., Gualandi, B., Heibi, I., Mariniello, N., Massari, A., Massidda, M., Moretti, A., Peroni, S., Pescarin, S., Rega, M. F., Renda, G., & Sullini, M. (2024). A reproducible workflow for the creation of digital twins in the cultural heritage domain. Zenodo. https://doi.org/10.5281/zenodo.14170838.

Sinha, N., Kapoor, G. & Kataria, N. (2025). Customer Engagement in Virtual Tourism: Embracing Immersive Experience of Culture and Heritage. In International Journal of Tourism Research 27, no. 4: e70094. https://doi.org/10.1002/jtr.70094.

Lo studio integrato del patrimonio materiale e immateriale per la ricostruzione delle pratiche storiche di gestione delle risorse ambientali nell'Alpujarra (Sierra Nevada): il progetto Sinergy

ALESSANDRO PANETTA

CNR ISTITUTO DI STORIA DELL'EUROPA MEDITERRANEA, CONSIGLIO NAZIONALE DELLE RICERCHE (CNR-ISEM), ITALIA-
ALESSANDRO.PANETTA@ISEM.CNR.IT

ABSTRACT (ITALIANO)

Il progetto Sinergy, sviluppato congiuntamente tra il CNR-ISEM e l'Università di Granada, è dedicato allo studio delle società rurali della Sierra Nevada (Alpujarra, Barranco del Poqueira) nei secoli recenti, attraverso la documentazione integrata del patrimonio materiale e immateriale legato alla gestione storica delle risorse ambientali. L'area conserva un ricco patrimonio di manufatti legati a colture irrigue (*acequias*, cisterne, canalizzazioni), sistemazioni di versante (terrazzamenti, muri), edifici e viabilità, oggi spesso abbandonati o rifunzionalizzati
La ricerca adotta un metodo regressivo e combina ricognizioni estensive con registrazione di Unità Topografiche (UT), rilievi fotogrammetrici 3D di strutture selezionate e raccolta di fonti orali semi-strutturate. I dati prodotti comprendono schede di UT, punti GPS, perimetrazioni e vettorializzazioni GIS, fotografie, modelli 3d, registrazioni audio, trascrizioni *verbatim* segmentate e annotate. Le interviste sono trascritte mediante prima bozza automatica con Whisper e successiva revisione umana, quindi segmentate per cambio di tema o episodio/pratica e annotate per collegare pratiche, spazi, attori sociali, strutture materiali e temporalità.
Bubión costituisce il principale laboratorio di documentazione sul campo già avviato; La Cebadilla, villaggio abbandonato legato alla modernizzazione idroelettrica del secondo Novecento, amplia invece la diacronia della ricerca e rappresenta il contesto previsto per la costruzione di una *deep map* in ambiente QGIS.
Il progetto mira così a costruire un'infrastruttura di conoscenza riusabile e interrogabile, in cui evidenze materiali, segmenti di intervista, dati spaziali e documentazione archivistica siano messi in relazione reciproca in una lettura multilivello del paesaggio culturale.
Parole chiave: archeologia del paesaggio; fonti orali; fotogrammetria 3D; deep mapping; gestione storica delle risorse ambientali.

ABSTRACT (ENGLISH)

An integrated study of tangible and intangible heritage for the reconstruction of historical practices in environmental resource management in the Alpujarra (Sierra Nevada): The Sinergy project.
The Sinergy project, developed jointly by CNR-ISEM and the University of Granada, is dedicated to the study of rural communities in the Sierra Nevada (Alpujarra, Barranco del Poqueira) in recent centuries, through the integrated documentation of the tangible and intangible heritage linked to the historical management of environmental resources. The area preserves a rich heritage of artefacts linked to irrigated farming (acequias, cisterns, canals), slope stabilisation works (terracing, walls), buildings and road networks, which are today often abandoned or repurposed
The research adopts a retrospective approach and combines extensive field surveys with the recording of Topographic Units (UTs), 3D photogrammetric surveys of selected structures, and the collection of semi-structured oral accounts. The data produced include TU records, GPS points, GIS perimeter measurements and vectorisations, photographs, 3D models, audio recordings, and segmented and annotated verbatim transcripts. Interviews are transcribed using an initial automatic draft generated by Whisper followed by human review, then segmented by change of topic or episode/practice and annotated to link practices, spaces, social actors, material structures and temporalities.
Bubión constitutes the main field documentation laboratory already underway; La Cebadilla, an abandoned village linked to the hydroelectric modernisation of the late 20th century, broadens the

diachronic scope of the research and represents the intended context for the construction of a deep map within the QGIS environment.
The project thus aims to build a reusable and queryable knowledge infrastructure, in which material evidence, interview segments, spatial data and archival documentation are interrelated within a multi-level reading of the cultural landscape.
Keywords: landscape archaeology; oral history; 3D photogrammetry; deep mapping; historical management of environmental resources

1. INTRODUZIONE

Questo contributo presenta il progetto Sinergy, realizzato congiuntamente tra il CNR-ISEM e l'Università di Granada e in collaborazione con l'Università di Genova (giugno 2025-dicembre 2026). L'area studiata è la Sierra Nevada, in particolare l'Alpujarra e il Barranco del Poqueira, dove si conservano tracce materiali e conoscenze relative alle pratiche storiche di gestione delle risorse ambientali: agricoltura irrigua, allevamento, uso dei boschi, sistemazioni di versante, modi di abitare e reti di percorrenza, oggi in gran parte abbandonate o trasformate e al centro di un processo diffuso di patrimonializzazione (Martín Civantos, 2014; García-Pulido, 2025; López Osorio et al., 2020).
L'obiettivo del progetto è documentare queste testimonianze archeologiche e orali per ricostruire storicamente i modi di gestione delle risorse ambientali e le loro trasformazioni, ma anche per conservare e rendere interrogabili conoscenze tecniche, lessici, pratiche e strutture materiali a esse connesse. Alla base del progetto vi è la ricostruzione delle dinamiche storiche che hanno interessato la regione montuosa dell'Alpujarra e il Barranco del Poqueira in età post-medievale e contemporanea (Martín Civantos et al., 2022). Al centro della ricerca si trovano i sistemi di gestione delle risorse ambientali e le trasformazioni delle comunità a essi collegate, osservati non solo come insieme di strutture funzionali, ma come intreccio storico tra pratiche, infrastrutture, spazi e attori sociali.
L'area è fortemente segnata dalla presenza storica di comunità irrigue, caratterizzate da pratiche e manufatti connessi alla gestione delle acque per le attività agro-silvo-pastorali, la cui traccia più evidente e caratterizzante è costituita dai reticoli di canali di irrigazione, le *acequias*. L'abbandono o la trasformazione di questi sistemi è accompagnato oggi dal degrado, dalla rifunzionalizzazione o dalla patrimonializzazione selettiva delle strutture e delle aree a essi collegate. Ciò è particolarmente evidente nei versanti storicamente riservati alle attività agricole e pastorali, dove la leggibilità archeologica di strutture e percorsi è ancora alta ma rischia di ridursi rapidamente.
I canali di irrigazione ancora in uso sono gestiti da *comunidades de regantes*, responsabili della distribuzione dell'acqua, della gestione dei turni, della risoluzione di conflitti e dell'organizzazione della pulizia e manutenzione dei canali (Ruiz-Ruiz & Martín Civantos, 2017; Abellán Santisteban et al., 2025)[1]. Questo rende il loro ruolo centrale non solo sul piano funzionale ma anche nella conservazione dei saperi tecnici associati alla costruzione, manutenzione e uso delle strutture e dei sistemi. Queste conoscenze, tramandate oralmente per generazioni, sopravvivono oggi perlopiù nella memoria di testimoni anziani e di pochi utilizzatori, a causa dell'abbandono di molte pratiche e dei recenti radicali cambiamenti economici e sociali.
Il progetto si colloca quindi in un punto di intersezione tra documentazione archeologica, fonti orali, gestione territoriale e patrimonializzazione. In questo quadro la sostenibilità non compare come formula generica, ma in rapporto a quattro questioni concrete: gestione territoriale delle acque e dei versanti, conservazione dei *saperi* locali, riuso scientifico e pubblico dei dati prodotti, pressione turistica su paesaggi storici fragili. Il patrimonio immateriale non è trattato come semplice cornice narrativa del patrimonio materiale, ma come insieme di informazioni necessarie per interpretarne funzionamento, temporalità e trasformazioni. Allo stesso modo la componente digitale non vuole essere qui proposta come semplice supporto alla documentazione, ma come ambiente di ordinamento, confronto e messa in relazione di dati topografici, modelli 3D, trascrizioni annotate e, in prospettiva, fonti cartografiche e archivistiche (Boyd, 2013; Roberts, 2016; Vrachliotou & Papatheodorou, 2024).

[1] La pulizia e manutenzione dei canali di irrigazione è spesso oggetto di iniziative di archeologia pubblica e comunitaria che coinvolgono l'Università di Granada in collaborazione con associazioni e comunità locali.

2. OBIETTIVI, CASI STUDIO E SCALA DELL'ANALISI

Concentrarsi solo sull'aspetto formale e funzionale delle *acequias* impedirebbe di cogliere il contesto economico, sociale e topografico più ampio in cui esse erano storicamente inserite, comprendente agricoltura, gestione del bosco, pascolo, allevamento, mobilità e forme dell'abitare. Il lavoro sul campo ha quindi riguardato la documentazione delle tracce materiali delle pratiche legate alla gestione delle risorse ambientali e la ricostruzione della loro dimensione non solo economico-produttiva, ma anche sociale e storica. Attraverso la fonte archeologica, integrata con le fonti orali, il progetto mira a ricostruire come diverse attività e infrastrutture si intersecassero concretamente nello spazio locale in determinati momenti.

Il progetto è focalizzato anche sull'abitare, inteso come forma storica di organizzazione degli spazi di vita e di lavoro: topografia delle aree abitate, edifici coinvolti nelle varie fasi produttive, strutture legate alla gestione delle risorse idriche, rapporti tra villaggi e spazi esterni, reti infrastrutturali idriche e stradali. Il duplice obiettivo di Sinergy è, da un lato, documentare e ricostruire le tracce materiali — manufatti, sistemazioni di versante, edificato, viabilità — e immateriali — terminologie, pratiche, *saperi* e tecniche — coinvolte nella gestione delle risorse ambientali; dall'altro, fornire una base documentaria e interpretativa utile a leggere criticamente i processi contemporanei di patrimonializzazione e trasformazione del paesaggio rurale.

Il progetto si articola in due casi di studio che assolvono differenti funzioni nell'economia generale. L'area di Bubión costituisce il principale laboratorio di documentazione sul campo già avviato, soprattutto per quanto riguarda i versanti soprastanti l'abitato, dove si conserva una forte densità di strutture ancora leggibili archeologicamente e presenti nella memoria orale locale. La sua rilevanza dipende proprio da questa condizione intermedia: manufatti e pratiche sono in parte usciti dall'uso, ma non ancora del tutto scomparsi dalla sfera materiale e della memoria. Il caso studio di Bubión consente quindi di osservare il passaggio da un sistema produttivo in uso a un sistema patrimonializzato, o comunque a un paesaggio in cui il valore testimoniale tende a prevalere su quello funzionale.

Il caso studio centrato sull'area de La Cebadilla, villaggio costruito negli anni Cinquanta per ospitare i lavoratori della centrale elettrica di Capileira, oggi abbandonato, consente invece di ampliare lo sguardo del progetto verso la fase della modernizzazione industriale, della sua crisi e della successiva risemantizzazione dei luoghi. Questa fase cruciale di elettrificazione rurale, oggi oggetto di una prima storicizzazione a livello provinciale[2], costituisce uno snodo fondamentale per l'area. In questo senso, il caso studio non è una semplice appendice contemporanea, ma un contesto utile per leggere la rottura e la ridefinizione dei modi di produrre, abitare e sfruttare le risorse ambientali e del conflitto su questi temi tra autorità centrali e comunità locali. Proprio per questo La Cebadilla rappresenta il contesto previsto per la costruzione di una *deep map* in ambiente QGIS, pensata non come semplice visualizzazione cartografica, ma come spazio di organizzazione e relazione tra fotografie storiche, documentazione archivistica, trascrizioni annotate e dati topografici (Roberts, 2016).

3. FONTI, DATI E METODI

La ricerca adotta un metodo regressivo, a partire dalle tracce di pratiche ancora in uso o in stato di recente abbandono per lo studio delle attività storicamente svolte lungo i versanti del Barranco del Poqueira. Le classi di evidenze attualmente documentate comprendono *acequias*, cisterne, *albercas*, terrazzamenti, muri di terrazzamento (*balates*), *cortijos*, *corrales*, edifici, viabilità e rimboschimenti a pineta. Al momento i dati concreti riguardano soprattutto Bubión, in particolare i versanti circostanti l'abitato; per La Cebadilla la raccolta sistematica dei dati è ancora in fase iniziale e riguarda, per ora, la predisposizione dell'integrazione fra documentazione topografica, fonti orali, cartografia e futura ricerca d'archivio.

Tra giugno e dicembre 2025 il progetto ha visto svolgere due campagne di ricognizione nell'area di studio, condotta con metodo estensivo (non per file equidistanti ma coprendo integralmente le porzioni di territorio selezionate) e ha utilizzato come unità minima spaziale di documentazione la Unità Topografica (UT). Per ciascuna UT sono stati raccolti dati di localizzazione (punti GPS, perimetrazione di poligoni,

[2] Per un inquadramento sul processo di modernizzazione energetica e sul patrimonio industriale dell'area, si veda la mostra digitale Historia de la electricidad en la provincia de Granada (Diputación de Granada): https://www.dipgra.es/servicios/areas/cultura/centros-culturales/casa-molino-de-angel-ganivet/exposiciones/micro-electricidad/

vettorializzazione di elementi lineari quali viabilità, canalizzazioni e corsi d'acqua) e descrittivi (schede, fotografie d'insieme e di dettaglio). Sono già stati effettuati campionamenti pollinici e geoarcheologici; eventuali sondaggi archeologici sono invece, allo stato attuale, previsti per l'autunno 2026.
La documentazione tridimensionale è stata avviata su strutture selezionate: due *acequias*, di cui una con ramificazioni secondarie nel campo irrigato, *un'alberca* e una *balata*. I modelli sono stati prodotti con *Zephyr*, scelto anche per verificare la spendibilità di un software gratuito in un contesto di ricerca applicata. In questa fase i modelli sono impiegati per documentazione, osservazione metrica, confronto morfologico, monitoraggio conservativo e visualizzazione, con l'obiettivo di costruire un abaco tipologico delle strutture legate alla gestione delle risorse ambientali del Barranco del Poqueira. Le acquisizioni fotografiche sono state effettuate mediante copertura completa dei manufatti e sovrapposizione fra immagini, secondo i protocolli di base comunemente adottati nello *Structure from Motion* (Remondino & El-Hakim, 2006; Green et al., 2014; Magnani et al., 2020). I modelli, al momento conservati in locale, non sono ancora stati georeferenziati o scalati sistematicamente per l'intero corpus; questo passaggio costituisce una delle integrazioni metodologiche previste per le prossime campagne.
La ricostruzione del patrimonio immateriale avviene attraverso la raccolta di fonti orali. Finora sono state realizzate tre interviste, rivolte a un *regante* attivo e a due anziani del villaggio, selezionati in base alla disponibilità, all'età e alla riconoscibilità locale come depositari di conoscenze relative a pratiche, luoghi, strutture e lessici tecnici. Si tratta di interviste semi-strutturate, con margini di adattamento al dialogo, dedicate a pratiche, riti, manufatti, modi di abitare, relazioni sociali, usi del monte e trasformazioni del paesaggio. La registrazione è avvenuta in formato audio, accompagnata da note di campo e fotografie contestuali.
Per la trascrizione si utilizza una prima bozza automatica prodotta con Whisper, sempre sottoposta a revisione umana mediante riascolto da parte dell'autore e dei collaboratori locali del progetto (Radford et al., 2022). La trascrizione conservata è *verbatim*; la predisposizione di una versione sintetica e ripulita è prevista per la disseminazione conclusiva del progetto, nell'ambito del quale vengono comunque mantenute le versioni integrali per la fase di lavoro. La segmentazione non avviene per durata fissa, ma per cambio di tema o per episodio/pratica, così da far coincidere la segmentazione con unità informative effettive. I segmenti vengono annotati con campi minimi: trascrizione, sintesi del segmento, pratica, spazio, attori sociali, struttura/manufatto, temporalità, lessico locale, note interpretative, collegamenti a foto, GIS e modelli 3D. Il lavoro di annotazione è pensato per essere svolto in ELAN, che consente di sincronizzare testo e audio e di costruire *tiers* distinti ma collegati (Sloetjes et al., 2007; Boyd, 2013).
Il collegamento tra segmento orale e dato materiale avviene quando il brano contiene almeno un riferimento esplicito a una pratica, un riferimento localizzabile a un luogo o a una struttura, e un riferimento identificabile a uno o più attori sociali. In questo modo non tutte le porzioni d'intervista vengono trattate allo stesso modo: i segmenti più utili sono quelli che consentono di mettere in relazione una pratica concreta, uno spazio riconoscibile e un soggetto o gruppo sociale. Un esempio utile è offerto dall'intervista a F.E.R. (58 minuti), centrata su vita rurale, domesticità, *montes* (terre collettive) e architettura vernacolare. In essa il segmento dedicato ai *hornos de carbón* e alla *yesera* collega una pratica produttiva, luoghi specifici del versante — Cerro Corona e Peñón de Cuatro Picos — attori familiari identificabili, quali il padre e lo zio dell'intervistata, e strutture materiali come carbonaie e cave di gesso. Il segmento diventa così un'unità informativa che potrà essere messa in rapporto con future UT, fotografie e dati topografici, mostrando in forma concreta come la memoria orale possa essere resa confrontabile con la documentazione archeologica. La scheda d'intervista e la trascrizione confermano inoltre che si tratta della seconda intervista fatta alla testimone, ma della prima registrata integralmente in audio e testo, con informazioni di prima mano e alcuni tagli dovuti a interruzioni, già dichiarati nella documentazione tecnica.
Nei prossimi mesi il protocollo delle fonti orali sarà formalizzato anche per quanto riguarda gli aspetti dell'anonimizzazione selettiva e dei livelli di accesso e riuso dei materiali audio e testuali. Le interviste già raccolte costituiscono per ora un corpus di lavoro interno; la forma definitiva di pubblicazione e consultazione dei materiali resta quindi in fase di definizione.
Nel caso di Sinergy, il digitale non coincide con il solo impiego di software specialistici per il rilievo o la trascrizione, ma con la costruzione di un ambiente di relazione tra unità topografiche, segmenti di intervista, immagini, modelli 3D e, in seguito, fonti cartografiche e archivistiche. In altri termini, il punto non è solo produrre più dati, ma produrli in modo che risultino confrontabili, interrogabili e riusabili. L'unità minima dell'informazione spaziale è la UT; l'unità minima dell'informazione orale è il segmento annotato; l'unità minima della documentazione tridimensionale è il modello riferito a una struttura o a

una sua parte; per cartografia e fonti documentarie il corrispettivo sarà costituito da *feature*, estratti o record singoli, non da blocchi indistinti di informazione. In questo modo il progetto non accumula soltanto documentazione, ma costruisce un sistema di relazione tra oggetti, luoghi, pratiche e comunità.
Il rapporto tra patrimonio materiale e immateriale viene così trattato in termini operativi: le strutture documentate sul terreno possono essere interrogate insieme ai segmenti di intervista che ne descrivono costruzione, uso, manutenzione, conflitti, lessico e temporalità; allo stesso tempo, un segmento orale può essere collegato non solo a un luogo, ma a una foto, a un modello 3D, a una UT e, in prospettiva, a una fonte archivistica o cartografica. È precisamente in questo passaggio che il digitale smette di essere un semplice strumento di supporto e diventa parte del dispositivo conoscitivo e interpretativo del progetto (Boyd, 2013; Vrachliotou & Papatheodorou, 2024).
A partire da giugno 2026 è previsto l'avvio della componente di *deep mapping* in ambiente QGIS per La Cebadilla. In questo caso il GIS non dovrà limitarsi a visualizzare punti su mappa, ma organizzare una stratigrafia di dati relativi a edifici, funzioni, memorie, immagini, cartografie e trasformazioni. La finalità è coagulare le informazioni attorno a quattro poli: singoli oggetti, singoli luoghi, singole pratiche e singole comunità, per poi ricostruire temi, fenomeni e processi più ampi. In questa prospettiva l'obiettivo del progetto coincide con la costruzione di un'infrastruttura di conoscenza riusabile e interrogabile, capace di rendere confrontabili fonti nate in contesti e formati diversi (Roberts, 2016).

4. CONCLUSIONI

Nel corso del 2025 il progetto ha visto svolgersi due campagne di ricognizione estensiva nell'area di Bubión, la mappatura di strutture e UT relative alle pratiche di gestione delle risorse ambientali, i primi rilievi fotogrammetrici 3D di strutture selezionate, le prime interviste orali e l'osservazione e partecipazione ad attività comunitarie di pulizia delle *acequias*. Le attività previste per il 2026 comprendono nuove interviste orali e ulteriori rilievi fotogrammetrici a partire da maggio, l'avvio del *deep mapping* in ambiente QGIS a partire da giugno, la prosecuzione delle ricognizioni, la ricerca d'archivio relativa a La Cebadilla.
I primi risultati non consistono ancora in una sintesi conclusiva del paesaggio storico dell'Alpujarra, ma in qualcosa di più elementare e, per ora, forse altrettanto utile: la verifica operativa che dati topografici, fonti orali e documentazione 3D possono essere prodotti secondo unità minime compatibili tra loro. Questo consente già ora di individuare linee di lavoro precise, anche funzionali all'obiettivo del progetto di realizzare un *abaco* delle strutture utilizzate nella gestione delle risorse ambientali: individuazione di pratiche e lessici tecnici, relazione tra memoria orale e trasformazioni osservabili del paesaggio, definizione di un protocollo replicabile per l'integrazione fra dati materiali e immateriali.
Nel breve periodo il progetto mira a ordinare e rendere confrontabili dati topografici, modelli 3D e fonti orali. Nel medio periodo alla costruzione di una mappa profonda de La Cebadilla e di un corpus annotato di interviste. Nel lungo periodo si intende infine produrre una base conoscitiva riusabile, utile sia alla ricerca sia alla restituzione pubblica e alla gestione del patrimonio rurale (le forme di disseminazione pubblica e di coinvolgimento partecipativo, non ancora pienamente sviluppati nel progetto, fanno parte della logica stessa di riuso dei dati).
Il rapporto con le comunità locali non assume per ora la forma di una co-produzione pienamente formalizzata del sapere, ma si fonda su collaborazione informale, accesso ai luoghi, disponibilità alle interviste e partecipazione a pratiche di manutenzione collettiva, come la pulizia delle *acequias*.
Il progetto Sinergy mostra che, nel contesto del Barranco del Poqueira, la documentazione archeologica delle strutture non può rimanere fine a sé stessa, come pure le memorie orali. Punto nodale è la messa in relazione strutture, pratiche, lessici, attori sociali e cronologie differenti, pur non riducendoli a un unico formato. Da questo punto di vista, il contributo non intende applicare semplicemente strumenti digitali a un caso archeologico, ma bensì tentare di costruire una forma di conoscenza multimodale e interrogabile, in cui il digitale serve a dare struttura, confrontabilità e riuso a dati materiali e immateriali.
L'aspetto più promettente del progetto non sta quindi soltanto nell'oggetto di studio, ma nella possibilità di definire un protocollo operativo per integrare survey archeologico, fotogrammetria 3D, fonti orali annotate e deep mapping all'interno di una stessa infrastruttura documentaria, e di offrire non solo un caso di studio locale, ma un modello trasferibile anche ad altri contesti di paesaggio rurale storico.

BIBLIOGRAFIA

Abellán Santisteban, J., Stagno, A. M., Paciotti, Y., Tacca, M., Brenu, N., Panetta, A., Bizzarri, G., González Montes, B., Menéndez-Blanco, A., & Santeramo, R. (2025). Jurisdictional conflicts as a form of sharing practices: water and canals between Cerdagne and Sierra Nevada (18th–21st centuries). In M. Di Tullio, A. M. Stagno, & M. L. Fagnani (Eds.), *The History of Environmental Resource Management in Europe: Sustainable Practices Through Time* (pp. 98–143). Routledge. https://doi.org/10.4324/9781003469070-5

Borrero, M., & Stroth, L. R. (2020). A proposal for the standardized reporting of error and paradata regarding Structure from Motion (SfM) 3D models used in recording and consolidating archaeological architecture. *Advances in Archaeological Practice, 8*(2), 171–187.

Boyd, D. (2013). OHMS: Enhancing access to oral history for free. *Oral History Review, 40*(1), 95–106. https://doi.org/10.1093/ohr/oht031

García-Pulido, L. J. (2025). Water Management in a UNESCO Biosphere Reserve: The Village of Pampaneira at the Barranco de Poqueira Heritage Site. *Blue Papers*, 4(2). https://doi.org/10.58981/bluepapers.2025.2.10

Green, S., Bevan, A., & Shapland, M. (2014). A comparative assessment of Structure from Motion methods for archaeological research. *Journal of Archaeological Science, 46*, 173–181. https://doi.org/10.1016/j.jas.2014.02.030

López Osorio, J. M., Lara García, J., & Peñalver Martín, S. (2020). Water and architecture in the Alpujarra region of Granada (Spain): An approach to the hydraulic systems of fountains and washing places of the region. *Journal of Traditional Building, Architecture and Urbanism*, 1, 395–410. https://doi.org/10.51303/jtbau.vi1.366

Magnani, M., Douglass, M., Schroder, W., Reeves, J., & Braun, D. R. (2020). The digital revolution to come: Photogrammetry in archaeological practice. *American Antiquity, 85*(4), 737–760. https://doi.org/10.1017/aaq.2020.59

Martín Civantos, J. M. (2014). Mountainous landscape domestication. Management of non-cultivated productive areas in Sierra Nevada (Granada-Almería, Spain). *European Journal of Post-Classical Archaeologies*, 4, 99–130.

Martín Civantos, J. M., et al. (2022). Singular Cultural Landscapes of the Sierra Nevada. In R. Zamora & M. Oliva (Eds.), *The Landscape of the Sierra Nevada: A Unique Laboratory of Global Processes in Spain* (pp. 31–46). Springer.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). *Robust speech recognition via large-scale weak supervision*. arXiv. https://doi.org/10.48550/arXiv.2212.04356

Remondino, F., & El-Hakim, S. (2006). Image-based 3D modelling: A review. *The Photogrammetric Record, 21*(115), 269–291. https://doi.org/10.1111/j.1477-9730.2006.00383.x

Roberts, L. (2016). Deep mapping and spatial anthropology. *Humanities, 5*(1), 5. https://doi.org/10.3390/h5010005

Ruiz-Ruiz, F. J., & Martín Civantos, J. M. (2017). La gestión comunitaria del agua en la cara norte de Sierra Nevada: Acción colectiva y saberes etnoecológicos en los sistemas de riego de origen andalusí. *e-rph*, 20, 76–103.

Sloetjes, H., Russel, A., & Klassmann, A. (2007). ELAN: A free and open-source multimedia annotation tool. In *Proceedings of Interspeech 2007*.

Vrachliotou, M., & Papatheodorou, C. (2024). Interoperability of oral history metadata: An ontological model. *Oral History Review, 51*(2), 446–474. https://doi.org/10.1080/00940798.2024.2379367

Reveduti, Correcti, Aprobati et Confirmati: The Digital Edition of the Statutes of Ascoli

MICHELA PARMA[1], FABIO MARIANI[2]

[1]Johannes Gutenberg-Universität Mainz, Germany – michela.parma@uni-mainz.de
[2]University of Augsburg, Germany – fabio.mariani@uni-a.de

ABSTRACT (ENGLISH)

This paper presents the digital edition of the statutes of Ascoli, a normative corpus promulgated in 1377 and preserved in a printed edition of 1496. The project aims to make a significant source for the legal and institutional history of the city accessible through a digital edition, based on a full transcription encoded in XML according to the TEI guidelines. The edition highlights the structural complexity of the statutory text by making explicit references to people, places, events, and thematic domains, and provides differentiated modes of access for both specialist and non-specialist audiences. Particular attention is devoted to the sustainability and long-term preservation of the project by adopting a static architecture. Finally, an experiment in automated named entity recognition shows that large language models locate entity-bearing regions with reasonable accuracy, while span delimitation remains challenging, reflecting the complexity of the source's vernacular register.
Keywords: digital scholarly edition; communal statutes; named entity recognition

ABSTRACT (ITALIANO)

Reveduti, Correcti, Aprobati et Confirmati: l'Edizione Digitale degli Statuti di Ascoli
Il contributo presenta l'edizione digitale degli statuti di Ascoli, un corpus normativo promulgato nel 1377 e conservato in un'edizione a stampa del 1496. Il progetto mira a rendere accessibile una fonte significativa per la storia giuridica e istituzionale della città attraverso un'edizione digitale, basata su una trascrizione codificata in XML secondo le linee guida TEI. L'edizione valorizza la complessità strutturale del testo statutario rendendo espliciti i riferimenti a persone, luoghi, eventi e ambiti tematici, e offre percorsi di consultazione differenziati per pubblici specialistici e non specialistici. Particolare attenzione è dedicata alla sostenibilità e alla durata nel tempo del progetto, attraverso l'adozione di un'architettura statica.
Infine, un esperimento di riconoscimento automatico di *named entities* dimostra che i *large language models* individuano le regioni contenenti entità con una discreta accuratezza, mentre la delimitazione precisa degli intervalli rimane complessa, riflettendo la complessità del registro vernacolare della fonte.
Parole chiave: edizione digitale; statuti comunali; named entity recognition

1. INTRODUCTION

Communal statutes represented one of the main forms of normative production at the local level and constituted the foundation of the urban legal order in communal Italy during the medieval period. The statutes should be understood as dynamic instruments, capable of translating into legal form political balances, institutional structures, and social practices in continuous transformation (Meccarelli, 1999). These codifications regulate a wide range of domains: from civil and criminal law to the organisation of the local government, from urban space management to the regulation of economic activities and social practices, thus offering an integrated vision of the city.
The statutes of Ascoli (modern-day Ascoli Piceno) emerge as a significant example in central Italy. As emphasised by Zdekauer and Sella, these statutes complete the series of statutory codifications of the major communes located along the main Italian trade routes (Zdekauer & Sella, 1910: 10). This chronological sequence includes Siena in 1310, Florence in 1325, Perugia in 1342, and Ascoli in 1377. *Statuti del Comune e del Popolo di Ascoli* were compiled and promulgated in 1377 by a group of local jurists, judges, and legal experts, tasked with reviewing and updating the existing normative corpus (De Santis, 1988: 199–205). This legal code remained the main regulatory reference for the city throughout

the 15th century, and, in 1496, was translated from Latin into the vernacular and published in print; this edition now forms the textual basis for the digital edition.[1]
The Ascoli corpus allows us to analyse the functioning of the communal statute both as a normative instrument and as a historical source, as well as to experiment, through a digital edition, with new analysis methods and to enhance a significant statutory collection of the communal age. The project for a digital edition of the statutes of Ascoli responds to the need to improve access to a text of major relevance to the city's legal and institutional history, whose specialised language and articulated structure have traditionally limited its readership to specialists. As occurred in 1496 with the translation into the vernacular and the printing of the statutes, this digital edition aims to enhance the statutory text through contemporary technologies. From this perspective, the project was guided by a set of key research questions:

- How can the hierarchical and cross-referential structure of a statutory corpus be represented in TEI to support non-linear navigation?
- How can semantic modelling through named entities and thematic categories transform a legal text into a network of meaningful, explorable resources?
- How can the structural and semantic modelling of the statutes be translated into an interface that supports both specialist research and public-oriented exploration?

These questions guided both the editorial choices as well as the design of the interface and consultation tools. The digital edition that emerged from this process is based on a transcription encoded according to the XML-TEI standard, on the semantic modelling of people, places, events, and thematic domains, and the development of mediation tools alternative to linear reading. From this perspective, the digital edition is not conceived merely as a transcription of the source, but as a scholarly infrastructure that supports multiple levels of access and interpretation for different audiences.

2. STATE OF THE ART

Long before the emergence of digital scholarly editions, medieval communal statutes had attracted the attention of historians, expressed through practices of collection, cataloguing, and preservation. A significant example is *Catalogo della raccolta di statuti, consuetudini, leggi, decreti, ordini e privilegi dei comuni, delle associazioni e degli enti locali italiani dal Medioevo alla fine del secolo XVIII* preserved at *Biblioteca del Senato della Repubblica "Giovanni Spadolini"*, which documents a wide corpus of statutory sources. Access to the catalogue has been extended through the digital facsimile of 150 statutes, including the statutes of Ascoli.[2] While not constituting digital editions in a strict sense, these resources provided an essential foundation for later developments in the field.
Building on this long-standing scholarly and institutional interest, since the early 2000s, Digital Humanities have devoted increasing attention to enhancing medieval normative sources through new editorial models, especially digital editions. In the Italian context, where communal laws constitute a core body of medieval normative production, projects such as *Statuto comunale di Bologna del 1376*, *Codice Diplomatico della Lombardia Medievale*, and the digital edition of *Statuto di Vicenza del 1264* represented some of the first research experiences in this historical field to employ XML-TEI encoding to make normative documents available in digital form (Zorzi, 2007; Siciliano & Salardi, 2009). These initiatives laid the foundations for the development of shared methodologies for the encoding and publication of statutory texts. They thereby contributed to the establishment of XML-TEI as the reference standard for the structured representation of texts and their semantic entities.
Within this evolving landscape, the field of digital editions of medieval sources was further enriched by projects that combined philological and editorial rigour with an increasing attention to issues of accessibility and audience. The diffusion of Edition Visualization Technology (EVT, Rosselli Del Turco, 2019) played a crucial role by providing a stable and widely adopted environment for the online

[1] The digital edition can be consulted at: https://statutiascoli.it/. The project material is available in the GitHub repository: https://github.com/statutiascoli/statutiascoli.github.io.
[2] The catalogue currently comprises 790 manuscripts dating from the XIII to the XIX century, 43 incunabula, and more than 4,000 printed editions produced between the XVI and the XX century. The catalogue is available at: https://www.senato.it/w3/Biblioteca/catalogoDegliStatutiMedievali.nsf/home?OpenPage. The digital reproduction of the 150 statutes is available at: https://www.senato.it/w3/Biblioteca/StatutiMedievali.nsf/home?OpenPage.

publication of digital scholarly editions, facilitating their circulation beyond a strictly specialist domain.[3] Within this framework, projects such as *Codice Pelavicino* and *Statuti di Monterosso* represented significant attempts to address a broader public, particularly at the local level. *Codice Pelavicino* employed the EVT to facilitate access for users interested in local history, offering multiple points of access to the text through images, indexes, full-text search, and structured lists of named entities (including places, people, roles, coins, and notable terms). The digital edition of *Statuti di Monterosso* followed the same model, further extending it through the integration of an Italian translation displayed alongside the transcription and the facsimile (Salvatori et al., 2017; Salvatori, 2022).

While these projects demonstrate how EVT can support diversified modes of access and improve the usability of scholarly editions for non-specialist audiences, the overall interaction model remains largely text-centred. In particular, tools for extratextual and spatial navigation, such as maps or timelines, are not natively integrated into the EVT environment. This limitation is especially relevant for editions of sources related to urban history, where spatial visualisations can play a crucial role in connecting documentary content to the historical geography of the city, fostering a more immediate and intuitive form of engagement.

The need for integrated extratextual tools can be understood as part of a shift in the theory of digital editing, where increasing attention has been paid to the audience. Recent scholarship has challenged the idea that critical editions are intended exclusively for specialist users, emphasising how digital publishing radically transforms access and potentially broadens the audience to include students, local history experts, and citizens interested in cultural heritage (Salvatori, 2022; Hajo, 2025). In this perspective, tools such as maps, indexes, and exploratory paths assume a central role, supporting diverse users and use cases through forms of guided consultation, while also fostering dynamics of discovery and serendipity (Barabucci et al., 2017; Chapman et al., 2025). In this sense, guided paths foster self-learning by encouraging users to explore the collection on their own (Salvatori, 2022: 235).

Alongside the redefinition of audience, a further critical dimension of digital editions concerns the sustainability and longevity of digital projects. Analyses on the life cycle of DH initiatives have shown that a significant percentage of projects become inaccessible within a few years, often due to technological obsolescence or the cessation of infrastructural support (Andreose et al., 2025). These critical issues have contributed to focusing attention on design strategies oriented towards persistence and infrastructural simplicity.

3. THE DIGITAL EDITION

The digital edition of the statutes of Ascoli was conceived as a tool for broad and differentiated use. It aims to transform the statutory text into digital form while providing a system that supports multiple modes of exploration and consultation, tailored to both specialist audiences and non-expert users.

The transcription is based on the critical edition of *Statuti di Ascoli Piceno dell'anno MCCCLXXVII* (Zdekauer & Sella, 1910).[4] The text of the 1910 edition was initially processed through OCR techniques and subsequently revised and corrected manually, with systematic reference to the 1496 incunabulum preserved at the Biblioteca del Senato della Repubblica, which constitutes the textual basis of the digital edition.[5] The final transcription was then encoded in XML in accordance with the guidelines of the Text Encoding Initiative (TEI).[6] The structure of the edition reflects the organisation of the 1496 printed text and is divided into two main volumes, *Statuti del Comune* and *Statuti del Popolo*, each subdivided into books and rubrics. Figure 1 offers an overview of the internal organisation of the statutes, illustrating both the hierarchical structure of the corpus and the thematic orientation of each book.

In total, the corpus comprises 513 rubrics, which constitute the fundamental textual units of the edition. Each rubric is encoded as an autonomous unit, identified by a sequential number, a title, and the full normative text. Alongside the textual structure, the XML-TEI file integrates structured lists of named

[3] The EVT software is open source. It is available at: http://evt.labcd.unipi.it/.

[4] Although the statutes were originally promulgated in 1377 in Latin, the original version is no longer extant. The only complete version of the statutes transmitted to us is the vernacular edition printed in 1496. The critical edition edited by Zdekauer and Sella, used as the reference for this digital edition, reproduces the 1496 incunabulum held at the Biblioteca del Senato della Repubblica "Giovanni Spadolini".

[5] OCR processing was performed using Tesseract OCR: https://github.com/tesseract-ocr/tesseract.

[6] The TEI P5 Guidelines are available at: https://tei-c.org/guidelines/p5/.

entities for people, places, and events, defined in the *teiHeader* and annotated throughout the document. These lists comprise 28 individuals, 126 places, and 68 festivities or events, forming the semantic backbone of the edition and providing the basis for alternative modes of access to the statutory text.
The hierarchical structure encoded in the XML file is reflected in the online representation of the text, which allows navigation by volume, book, and rubric. Each rubric is accompanied by digital images of the 1496 incunabulum alongside its transcription.[7] This choice maintains a clear link to the original source and ensures editorial transparency by enabling direct comparison between the transcription and the printed text.
A central aspect of the methodology involves creating alternative access paths to the text, as a complement to linear reading, based on spatial and temporal dimensions. Starting from the lists of places and events encoded in the XML-TEI file, the edition offers interactive maps of the city and its territory, as well as a calendar of festivities, which allow users to explore the statutes starting from urban and territorial geography and the festive cycle.[8] These visualisations directly connect space and time to the statutory rubrics, transforming these dimensions into keys for accessing the normative text.

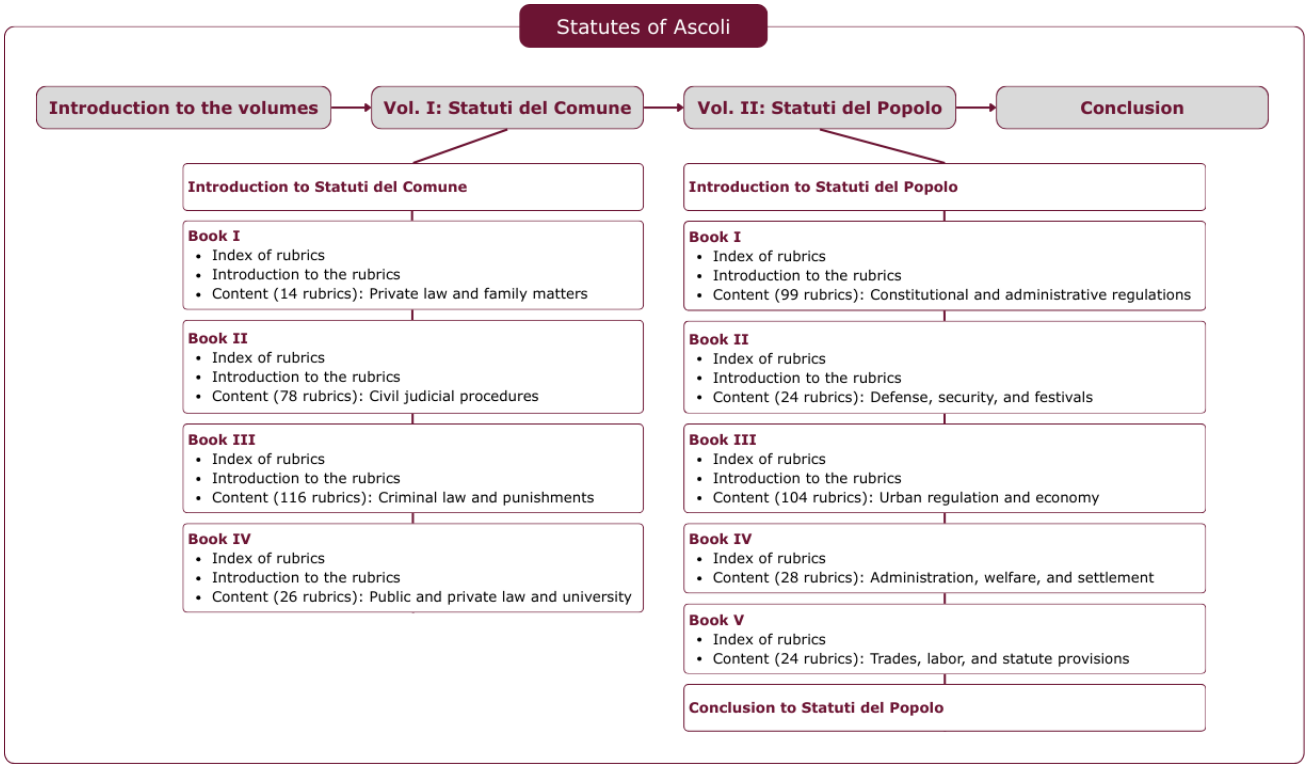


Figure 1. Structural overview of the statutes of Ascoli.

In addition to spatial and temporal paths, the edition includes a list of people mentioned in the statutes, compiled from the codification of names occurring in the text, and a thematic taxonomy composed of 20 categories, applied to a selection of rubrics. The thematic categories organise the corpus according to recurring conceptual domains of the institutional, social, and economic life of the Ascoli community (see Figure 2). Since not all statutory provisions lend themselves to an unambiguous classification, thematic categorisation was applied only to a selection of rubrics (203, approximately 40% of the 513 rubrics). For each categorised rubric, a summary of the normative content was also included as a mediation tool for

[7] The digital images of the 1496 incunabulum are reproduced courtesy of the Biblioteca del Senato della Repubblica "Giovanni Spadolini".

[8] The construction of the maps is based on a process of historical and topographical reconstruction derived from the indications contained in the statutes, integrated with later cadastral and cartographic sources (see the project *Catasto 1381 Ascoli*, available at: https://catasto1381ascoli.archiviodistatoap.it/). The festive calendar was reconstructed with reference to Cappelli (1998: 60, 137–189).

non-specialist audiences. These summaries accompany the original text and facilitate comprehension where the legal language is particularly complex.
The digital encoding also accounts for the internal cross-referential nature of the statutory text, as certain rubrics explicitly refer to other rubrics within the corpus as normative precedents or points of clarification. A total of 27 internal references were identified and encoded as hyperlinks, enabling direct navigation between related rubrics and supporting a non-linear mode of consultation that reflects the interconnected structure of the statutes.
In line with an approach oriented towards public history and user participation, the digital edition also integrates mechanisms for audience engagement. In fact, several sections of the website feature a *call-to-action* inviting users to report any transcription errors or to contribute information regarding people and places mentioned in the statutes, particularly in cases where toponyms or other terms may reflect vernacular forms that still coexist in local dialect usage. This feedback channel is conceived as a tool for dialogue with the user community and as a possible support for the progressive improvement and enrichment of the edition.
Alongside its scholarly and user-oriented aims, the project placed emphasis on the technical implementation of the digital edition, guided by a commitment to long-term sustainability. The digital edition of the statutes of Ascoli was developed using a fully static architecture. The website is published via GitHub Pages and is based exclusively on static files (HTML, CSS, JavaScript, and XML), ensuring autonomous operation and minimal maintenance over time. This implementation choice was guided by concerns for sustainability and persistence, aiming to reduce risks and costs related to infrastructure maintenance and to ensure the durability and continued accessibility of the edition.



Figure 2. Visualisation of the Categories access path. The selected category Family displays the list of rubrics associated with this domain, allowing users to explore the corpus through a thematic perspective.

Together, these methodological choices enable a shift from linear reading to multi-path exploration, in which space, time, people, and themes act as orienting devices. XML-TEI encoding thus constitutes the infrastructure that connects the normative text to its multiple dimensions of meaning, supporting a model of digital edition that is attentive to the complexity of the source and to the needs of accessibility and comprehension of different audiences.

4. NAMED ENTITY RECOGNITION

The named entity annotation produced during the encoding process represents one of the most philologically demanding components of the edition. The 222 unique entities manually identified and encoded across the corpus (28 individuals, 126 places, and 68 festivities or events) were annotated according to a set of principles grounded in the specific documentary conventions of the source. For places, the annotated span consistently encompasses the full nominal group as it appears in the statutory text, including the place-denoting noun where present: *castello de Appognano*, *ecclesia de Sancta Maria majore*, *palazzo del Popolo* are tagged in their entirety rather than reduced to the bare toponym. For individuals, spans include the full name as given, with patronymic or place of origin where the text provides them: *Muctio de Johanni de Bernardo*, *Joanni da Theramo*. For festivities, the full liturgical

formula is retained: *festa de sancto Emidio*, *venardì sancto*. These conventions reflect the structure of the source itself, in which names rarely appear in isolation but are embedded in the syntactic and formulaic patterns of communal statutory language.

To assess the extent to which such annotation could be replicated automatically, we conducted an experiment using two large language models: LLaMA 3.3 70B and Qwen 2.5 72B.[9] Both models were prompted with a structured definition of the three entity categories (PER for person, LOC for location, and FES for festivity) and instructed to return entity spans. The manual annotations served as the gold standard. Evaluation was performed under two criteria: exact match, requiring both the correct category and the precise character span, and overlap match, defined here as intersection over union (IoU) ≥ 0.5, which accepted predictions with the correct category whose span overlapped at least half of the gold span (Table 1).

Under exact matching, both models achieved modest overall F1 scores. LLaMA reached 0.277 and Qwen reached 0.348. PER yielded the strongest results, with F1 values of 0.444 and 0.526 respectively. LOC and FES proved more challenging, with F1 values ranging from 0.258 to 0.370. The discrepancy between exact and overlap scores highlights a systematic pattern. Under overlap matching, overall F1 rises dramatically to 0.635 for LLaMA and 0.655 for Qwen. The largest gains are in LOC (+0.33 and +0.30) and FES (+0.49 and +0.36). This pattern reveals that the predominant error mode is not a failure to identify the correct referent, but a difference in span delimitation. Models identify the correct text region but struggle to align with editorial span boundaries, tending to omit the place-denoting noun (e.g., reducing *castello de Appognano* to *Appognano*) or the liturgical designation (e.g., reducing *festa de sancto Emidio* to *sancto Emidio*). PER shows a smaller gap overall, as personal name spans tend to be syntactically more compact in this corpus.

	LLaMA 3.3 70B			**Qwen 2.5 72B**		
Exact match	P	R	F1	P	R	F1
PER	0.341	0.636	0.444	0.429	0.682	0.526
LOC	0.293	0.230	0.258	0.315	0.334	0.324
FES	0.323	0.244	0.278	0.352	0.389	0.370
Overall	**0.306**	**0.253**	**0.277**	**0.333**	**0.365**	**0.348**
Overlap (IoU≥0.5)	P	R	F1	P	R	F1
PER	0.439	0.818	0.571	0.514	0.818	0.632
LOC	0.672	0.528	0.591	0.607	0.644	0.625
FES	0.889	0.672	0.765	0.697	0.771	0.732
Overall	**0.702**	**0.580**	**0.635**	**0.625**	**0.687**	**0.655**

Table 1. Precision (P), recall (R), and F1 for LLaMA 3.3 70B and Qwen 2.5 72B on named entity recognition, evaluated with exact and overlap (IoU ≥ 0.5) matching.

These results speak to a broader characteristic of the source. The vernacular of the 1496 printed statutes of Ascoli is a hybrid register that combines medieval legal formulae with local dialectal features. This linguistic specificity poses a particular challenge for models trained predominantly on modern text (Esposito, 2025: 15-16). The vernacular character of the statutes thus lends itself to further linguistic investigation, and the experiment reported represents a first step in that direction, establishing a baseline against which future approaches can be meaningfully compared.

5. CONCLUSIONS

The digital edition of the statutes of Ascoli aims to contribute to the debate on scholarly digital editions of medieval normative sources, addressing in an integrated manner the issues of scholarly rigour, accessibility, and long-term sustainability. Starting from a transcription encoded in XML-TEI, the project

[9] Model resources: LLaMA 3.3 70B Instruct (https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct) and Qwen 2.5 72B Instruct (https://huggingface.co/Qwen/Qwen2.5-72B-Instruct). The codebase, prompts, and experimental results are available at https://github.com/statutiascoli/ner. Computations were performed using HPC resources provided by the Erlangen National High Performance Computing Center (NHR@FAU) at Friedrich-Alexander Universität Erlangen-Nürnberg (FAU), within the framework of the BayernKI initiative, funded by the Bavarian state government.

highlights the structural complexity of the text and enables differentiated modes of exploration, grounded in spatial, temporal, and thematic dimensions, designed for specialist and non-specialist audiences. The adoption of a static architecture responds to the critical issues related to the durability of digital projects, proposing the edition as a resource for the study of the institutional history of the city and as a sustainable and replicable model for digital editions of medieval legal sources.
Finally, the named entity annotation that underpins the access paths also served as the basis for an experiment in automated recognition. The results confirm that large language models can identify entity-bearing regions in the text with reasonable accuracy, even in the absence of training data specific to this domain. At the same time, the gap between exact and overlap performance reveals that span delimitation is shaped by editorial conventions that are not easily recovered from a prompt description alone, and that the vernacular register of the source adds a further layer of difficulty. These findings point to productive directions for future work, combining NLP-based approaches with the crowdsourcing of local and dialectological knowledge from the community of Ascoli, with the aim of deepening the linguistic study of the corpus and supporting the development of a dedicated glossary.

REFERENCES

Andreose, E., Crosilla, G., Grillo, R., & Spinaci, G. (2025). Life and death of DH projects: A preliminary investigation of their lifecycles in Italy. In S. Rebora, M. Rospocher, & S. Bazzaco (Eds.), Diversità, equità e inclusione: Sfide e opportunità per l'informatica umanistica nell'era dell'intelligenza artificiale (Atti del XIV Convegno Annuale AIUCD, Verona, 11–13 giugno 2025).

Barabucci, G., Spadini, E., & Turska, M. (2017). Data vs. presentation. What is the core of a scholarly digital edition? In P. Boot, A. Cappellotto, W. Dillen, F. Fischer, A. Kelly, A. Mertgens, A.-M. Sichani, E. Spadini, & D. Van Hulle (Eds.), Advances in Digital Scholarly Editing: Papers presented at the DiXiT conferences in The Hague, Cologne, and Antwerp (pp. 37-46). Sidestone Press.

Cappelli, A. (1998). Cronologia, cronografia e calendario perpetuo: Dal principio dell'era cristiana ai nostri giorni (7th ed.). Ulrico Hoepli.

Chapman, A., Holmes, M., Fralick, K., Fukushima, K., Bakhtvar, N., & Pinto, S. (2025). Browse, search and serendipity: Building approachable digital editions. In J. O'Sullivan, M. Pidd, S. Whittle, B. Wessels, M. Kurzmeier, & Ó. Murphy (Eds.), Digital editing and publishing in the twenty-first century. Scottish Universities Press. https://doi.org/10.62637/sup.GHST9020.6

De Santis, A. (1988). Ascoli nel Trecento (Vol. II, 1350–1400).

Esposito, S. (2025). Named entity recognition for historical Italian texts: Overcoming data limitations through strategic annotation and model adaptation (Master's thesis, Alma Mater Studiorum – Università di Bologna, Department of Computer Science and Engineering).

Hajo, C. (2025). Who are we editing for? How digital publication changes the role of the scholarly edition. In J. O'Sullivan, M. Pidd, S. Whittle, B. Wessels, M. Kurzmeier, & Ó. Murphy (Eds.), Digital editing and publishing in the twenty-first century. Scottish Universities Press. https://doi.org/10.62637/sup.GHST9020.2

Meccarelli, M. (1999). Statuti, potestas statuendi e arbitrium: La tipicità cittadina nel sistema giuridico medievale. In E. Menestò (Ed.), Gli Statuti delle città: l'esempio di Ascoli nel secolo XIV: atti del convegno di studio svoltosi in occasione della dodicesima edizione del "Premio internazionale Ascoli Piceno", Ascoli Piceno, 8-9 maggio 1998. (pp. 87–124).

Rosselli Del Turco, R. (2019). Designing an advanced software tool for Digital Scholarly Editions: The inception and development of EVT (Edition Visualization Technology). Textual Cultures, 12(2), 91–111.

Salvatori, E. (2022). Edizioni digitali di fonti medievali tra ricerca e pubblico. In E. Salvatori (Ed.), Il medievista come public historian (pp. 227–250). ISIME.

Salvatori, E., Rosselli Del Turco, R., Alzetta, C., Di Pietro, C., Mannari, C., & Miaschi, A. (2017). Il Codice Pelavicino tra edizione digitale e Public History. Umanistica Digitale, 1(1). https://doi.org/10.6092/issn.2532-8816/7232

Siciliano, L., & Salardi, V. (2009). The digital edition of the Statuta communis Vicentie of 1264. Literary and Linguistic Computing, 24(3), 329–338. https://doi.org/10.1093/llc/fqp021

Zdekauer, L., & Sella, P. (Eds.). (1910). Statuti di Ascoli Piceno dell'anno MCCCLXXVII. Istituto Storico Italiano.

Zorzi, A. (2007). L'edizione digitale degli statuti dei comuni italiani. Qualche riflessione a modo di premessa. Mélanges de l'École française de Rome. Moyen Âge, 119(2), 323–326

Trascrivere automaticamente i registri catastali sardi del primo Novecento con Transkribus: risultati di un caso di studio

Andrea Pergola[1], Cecilia Tasca[2]

[1] Università degli Studi di Cagliari, Italia - andrea.pergola@unica.it
[2] Università degli Studi di Cagliari, Italia - tasca.cecilia@tiscali.it

ABSTRACT (ITALIANO)

L'articolo presenta i risultati del processo di digitalizzazione e trascrizione automatica di una serie di registri catastali sardi dei primi del Novecento attraverso l'utilizzo dell'applicativo Transkribus. Tale processo ha permesso la creazione di tre modelli personalizzati (tabella, layout, trascrizione) sviluppati attraverso l'etichettatura e la trascrizione manuale di 21 registri, che hanno portato a un significativo miglioramento dei modelli base proposti dalla piattaforma, consentendo di ottenere dataset strutturati interrogabili con un margine di errore contenuto.
Parole chiave: Transkribus; HTR; Sardegna; Archivistica.

ABSTRACT (ENGLISH)

Automatic Transcription of Early Twentieth-Century Sardinian Cadastral Registers Using Transkribus: A Methodological Case Study
The article presents the results of the digitization and automatic transcription process applied to a series of early twentieth-century Sardinian cadastral registers using the Transkribus platform. This process enabled the creation of three customized models (table, layout, and transcription), developed through the manual labeling and transcription of 21 registers. These models significantly improved the baseline configurations provided by the platform, allowing the generation of structured and searchable datasets with a limited margin of error.
Keywords: Transkribus; HTR; Sardinia; Archives.

1. INTRODUZIONE

Negli ultimi anni, i sistemi di Handwritten Text Recognition (HTR) sono stati utilizzati con sempre maggior frequenza in progetti di analisi, descrizione e valorizzazione dei patrimoni documentari. In questo contesto, Transkribus, sviluppata nell'ambito dei progetti Transcriptorium (Sánchez et al., 2013) e READ, oggi READ-COOP (Terras et al., 2025), si è affermata come piattaforma di riferimento per la trascrizione automatica di documentazione storica manoscritta (Muehlberger et al., 2019; Nockels et al., 2022)[1], grazie alla sua capacità di trasformare immagini digitalizzate in risorse interrogabili (Romein et al., 2025). Seppur ancora frammentati, diversi casi di studio (Carboni & Molino Lova, 2025; Iorio, 2019; Loss et al., 2025; Schwarz-Ricci, 2022) dimostrano un uso sempre più sistematico di Transkribus in contesti archivistici che può facilitare enormemente l'interrogabilità di fondi documentari consistenti attraverso la loro trasformazione in basi dati ricercabili. Questa possibilità non ha solo impatti diretti sulla ricerca storica e, più in generale, sulle pratiche di accesso alle informazioni in ambito digitale, ma può agevolare concretamente azioni di descrizione archivistica di fondi di grande entità. A questo proposito, il presente contributo fornisce informazioni sull'esperienza portata avanti da un gruppo di ricerca dell'Università degli Studi di Cagliari[2] su un *corpus* di registri catastali sardi dei primi del Novecento, con lo scopo di permettere una valutazione dei risultati raggiunti, ossia la creazione di tre modelli personalizzati (tabella, layout, trascrizione) che ha permesso di elaborare una base di dati con un margine di errore contenuto.

[1] Sebbene Transkribus si sia attestata nel tempo come piattaforma di riferimento, va ricordata anche la piattaforma web Escriptorium, piattaforma open-source per il riconoscimento di testo manoscritto (HTR) nei documenti storici, integrata con il motore Kraken (Kiessling et al., 2019; Sichera et al., 2025).
[2] Il gruppo di ricerca, sotto il coordinamento di Giampaolo Salice e Cecilia Tasca, è composto da Bianca Fadda, Mariangela Rapetti, Eleonora Todde, Andrea Pergola e Beatrice Schivo. Hanno collaborato alle attività di ricerca anche Stella Barbarossa e, nelle ultime fasi, Giommaria Carboni e Davide Pisanu.

2. CORPUS E OBIETTIVI

L'utilizzo dell'applicativo è stato il fulcro di un'indagine archivistica volta allo studio di documenti utili all'individuazione di territori soggetti ad uso civico in Sardegna[3], realizzata in seno al Centro di Umanistica Digitale dell'Università degli Studi di Cagliari (DH.UniCA)[4]. L'attività di ricerca è stata condotta sui registri catastali del fondo Ufficio Tecnico Erariale (UTE) dell'Archivio di Stato di Cagliari[5]: unità manoscritte, prodotte da diversi Comuni della Sardegna e dunque caratterizzate da una forte eterogeneità in termini di formato, qualità di scrittura e stato di conservazione (presenza di pieghe, bruciature, disallineamenti, differenze di contrasto). Il contenuto informativo di questi registri è organizzato in strutture tabellari più o meno regolari, con colonne comprensive di dati numerici e testuali relativi ad appezzamenti di terreni e ai loro possessori, alla loro localizzazione geografica, alla qualità delle colture, alla stima economica e, in ultimo, a eventuali osservazioni. L'attività è stata realizzata attraverso diverse fasi, condotte in collaborazione con la Società Linkalab S.r.l. di Cagliari, che ha curato gli aspetti prettamente informatici della sperimentazione e supportato con un ausilio formativo di base[6]. Questa modalità di lavoro ha permesso, non solo, di prendere dimestichezza con il software e i suoi processi, ma ha anche dato modo al gruppo di ricerca diverificare in maniera graduale il livello di risposta dei modelli forniti dal software, specialmente su questo specifico tipo di documentazione che, pur prodotta dalla medesima tipologia di soggetto produttore, presenta spesso difformità grafiche e organizzative. La sperimentazione ha portato al raggiungimento di un triplice obiettivo: migliorare il margine di errore dei modelli base di Transkribus attraverso la creazione di modelli personalizzati, portando a ridurre quanto più possibile la correzione manuale dei risultati ottenuti; riuscire a produrre un file strutturato unico ed esportabile – necessario per procedere a interrogazioni puntuali della documentazione analizzata–; creare i presupposti per procedere con la trascrizione automatica di ulteriore documentazione di tipologia simile a quella di partenza con tempi maggiormente rapidi rispetto alla trascrizione manuale.

3. METODOLOGIA E WORKFLOW

L'attività di trascrizione automatica è stata impostata in un flusso di lavoro in più fasi, pensato per rispondere alle caratteristiche specifiche del corpus: registri catastali manoscritti, fortemente eterogenei per qualità materiale, organizzazione grafica e fisionomia scrittoria. La scelta di utilizzare Transkribus come strumento di lavoro principale risponde non solo alla sua consolidata affidabilità scientifica, ma anche alla sua recente adozione come standard nazionale per i 102 Archivi di Stato italiani[7]. Questa piattaforma offre un ecosistema integrato che permette di gestire in un unico workflow la complessa segmentazione delle strutture tabulari sarde e la successiva trascrizione dei campi testuali, facilitando la creazione di database pronti per la ricerca avanzata: analisi del layout, riconoscimento di strutture tabellari, trascrizione automatica del testo manoscritto, correzione manuale ed esportazione dei risultati

[3] Il progetto "Indagine archivistica nei fondi medievali e moderni, acquisizione, digitalizzazione, trascrizione documenti reperiti attraverso uno studio paleografico e diplomatistico, creazione mappa geolocalizzata, individuazione territori soggetti ad uso civico in Sardegna", è stato finanziato dall'Assessorato della Pubblica Istruzione, Beni Culturali, Spettacolo, Sport e Informazione della Regione Autonoma della Sardegna, L.R. 21 febbraio 2023, n. 1 (Legge di stabilità 2023), art. 1, comma 3, Tabella D); responsabile scientifico Prof. Giampaolo Salice, Dipartimento di Lettere, Linge e Beni culturali. Si tratta della prosecuzione del progetto "Dai Feudi agli Usi civici: la mappa archivistica per la storia della proprietà in Sardegna", anch'esso finanziato dalla Regione Autonoma della Sardegna L.R. 7/07 annualità 2017) a valere sul Fondo di sviluppo e coesione 2014-2020 (Patto per lo sviluppo della Regione Sardegna - Area tematica 3 Linea d'azione 3.1 "Interventi di sostegno alla ricerca"), responsabile scientifico Prof.ssa Cecilia Tasca, Dipartimento di Lettere, Lingue e Beni culturali, conclusosi nel dicembre 2022. Hanno collaborato alle attività di ricerca Andrea Pergola e Myriam Mereu.

[4] (DH.UniCa, s.d.).

[5] Costituito da 6.616 unità (cc. 3917, regg. 2458, b. 1) relative ai secoli XIX e XX, il fondo contiene la documentazione prodotta in seguito alla misurazione dei terreni realizzata per la formazione del primo catasto terreni, e comprende, oltre al materiale cartografico, i registri delle matricole dei possessori, i registri delle partite e i sommarioni (*Archivio di Stato di Cagliari*, s.d.).

[6] Si ringrazia inoltre Linkalab per la predisposizione di un report dettagliato sugli aspetti tecnici delle operazioni condotte, parzialmente utilizzato per la stesura di questo contributo (*Homepage*, 2024).

[7] A questo proposito, l'ICAR nel luglio 2025 ha dato il via al progetto Transkribus, che ha l'obiettivo di «fornire ai 102 Archivi di Stato italiani e ad ogni ricercatore la possibilità di accedere in modo più semplice e rapido alla lettura di documenti trascritti automaticamente in formato digitale». (Archivi | Via al progetto «Transkribus», 2025).

in formati riutilizzabili. Questa integrazione è risultata particolarmente adatta per la documentazione oggetto del presente studio, che non presenta un problema esclusivamente grafico, ma anche strutturale e tabellare.
Il lavoro è stato preceduto da un'operazione di acquisizione digitale dei registri in formato TIFF[8]. Poiché Transkribus non consente il caricamento diretto di questo formato, è stato necessario realizzare, sempre con il supporto tecnico della Società Linkalab, uno script Python per la conversione massiva dei file in PDF e PNG.
Una volta predisposte le immagini, i documenti sono stati organizzati in collezioni distinte sulla base del Comune di provenienza. Tale scelta ha risposto a un'esigenza sia tecnica sia metodologica: da un lato, ha consentito una gestione ordinata del corpus, dall'altro, ha permesso di strutturare con maggior precisione la variabilità interna dei documenti, evitando che il training si concentrasse su un unico tipo di mano, di impaginazione o di stato di conservazione. Includere registri diversi per provenienza, struttura e condizioni materiali, ha permesso di aumentare la capacità di generalizzazione dei modelli pur a discapito di una loro maggiore precisione.
Il flusso di lavoro è stato articolato realizzando gradualmente tre modelli distinti, ciascuno dei quali rispondeva a una funzione specifica. Il primo, il modello tabellare, era destinato al riconoscimento della struttura delle tabelle presenti nei registri. Nei documenti analizzati, infatti, le informazioni sono organizzate in griglie regolari, composte da righe e colonne contenenti dati testuali e numerici. La segmentazione tabellare è stata trattata come un passaggio autonomo, preliminare rispetto alla trascrizione vera e propria, perché in questo tipo di documentazione un errore nella rilevazione della struttura può rendere instabile l'associazione tra cella, contenuto e posizione e compromettere irrimediabilmente l'interrogabilità del dato.
Il secondo modello, dedicato al layout, è stato impiegato per l'individuazione delle linee testuali e numeriche e per la definizione delle *baseline* su cui poi si fonda il riconoscimento dei caratteri. Questa fase non va confusa con la trascrizione, poiché riguarda l'individuazione della struttura grafica del testo e non la lettura del contenuto. Il modello layout ha quindi avuto il compito di stabilire con maggiore precisione l'assetto delle linee, riducendo gli errori causati da righe spezzate, campi disallineati o elementi grafici poco leggibili. In presenza di registri con cifre, apici, simboli di ripetizione e altre convenzioni amministrative ricorrenti, questa fase si è rivelata decisiva per limitare i falsi positivi e migliorare la qualità dell'input destinato al modello di trascrizione (Schwarz-Ricci, 2022).
Il terzo modello, infine, è stato dedicato alla trascrizione automatica del testo manoscritto. L'avvio del training è stato effettuato a partire da modelli base disponibili nella piattaforma, individuati in base alla loro rassomiglianza, tipologica e cronologica, con le grafie del corpus. Una scelta di partenza, dunque, non arbitraria, ma guidata dalla volontà di ridurre la distanza tra base model e documentazione di riferimento, limitando il rischio di trasferire al nuovo training configurazioni troppo lontane dalle caratteristiche reali dei registri. Questo modello è stato l'ultimo applicato alla lavorazione, perché riconosce i caratteri a partire dalle linee identificate dal modello layout e li colloca all'interno delle sezioni individuate dal modello tabella. La trascrizione automatica si basa su sistemi di Handwritten Text Recognition (HTR), che non coincidono con l'OCR. A differenza dei sistemi OCR, che scompongono il testo in singoli glifi indipendenti, la tecnologia HTR interpreta l'andamento orizzontale della scrittura come un flusso unitario (Bazzaco, 2024). Questo approccio permette al software di non limitarsi al riconoscimento delle lettere isolate, ma di analizzare l'intera riga di testo, sfruttando modelli probabilistici per decodificare le grafie talvolta compresse o irregolari tipiche dei compilatori dei registri catastali
L'addestramento dei tre modelli è avvenuto in modo progressivo, con un incremento graduale del numero di immagini annotate e trascritte manualmente. Questa strategia è stata adottata per due ragioni principali. In primo luogo, perché consentiva al gruppo di ricerca un maggior controllo sulla qualità delle annotazioni e di correggere gli errori nelle fasi iniziali, prima che si propagassero nelle versioni successive dei modelli. In secondo luogo, perché permetteva di misurare con maggiore chiarezza il contributo di ciascun ciclo di addestramento al miglioramento delle prestazioni. Le immagini impiegate per l'addestramento sono state selezionate in modo da rappresentare la maggiore varietà possibile di

[8] Il lavoro di reperimento, schedatura, acquisizione digitale e metadatazione dell'imponente documentazione reperita negli Archivi di Stato di Cagliari e di Sassari, nell'Archivio centrale dello Stato di Roma e nell'Archivio dell'Assessorato dell'Agricoltura di Cagliari, che ammonta allo stato attuale a oltre 20.000 file digitalizzati, è stato realizzato da Stella Barbarossa (Borsa di ricerca n. 09/2023, Dipartimento di Lettere, Lingue e Beni culturali) a partire dal 1° dicembre 2023.

condizioni: registri di Comuni diversi, mani differenti, tabelle regolari e irregolari, pagine in buono stato e pagine danneggiate da strappi, pieghe, bruciature, macchie o disallineamenti dovuti alla digitalizzazione.

Data l'estrema variabilità delle mani e dello stato di conservazione dei registri dell'UTE, la creazione di un ground truth[9] accurato è stata la fase più delicata. Per i modelli tabellare e layout, le annotazioni sono state inizialmente prodotte con il supporto dei modelli base della piattaforma e poi corrette manualmente; per il modello di trascrizione, a causa dell'unicità della scrittura manoscritta, nonostante la loro rassomiglianza con le grafie del corpus, i modelli di trascrizione di base sviluppati dalla community di Transkibus, hanno spesso dato risultati non accettabili. Per questo motivo, dunque, nelle prime fasi di utilizzo dell'applicativo è spesso necessario procedere con una profonda revisione manuale dei risultati ottenuti con i modelli di base o direttamente con una trascrizione manuale del materiale. Questa scelta rispondeva alla necessità di disporre di dati di addestramento il più possibile accurati, poiché la qualità del *ground truth* incide in modo determinante sull'efficacia del modello finale. I dati così prodotti hanno costituito la base per successive iterazioni di training e validazione.
Il processo si è svolto in due fasi principali. Nella prima, il training è stato condotto su un campione iniziale di documenti, con il duplice obiettivo di verificare la risposta dei modelli base e misurare la distanza tra questi e la documentazione di riferimento. Nella seconda fase, i modelli sono stati riaddestrati su un corpus più ampio, comprendente un numero consistente di immagini provenienti da più registri e più Comuni. Questa procedura ha consentito una progressiva riduzione del *Character Error Rate* (CER) [10] per il modello di trascrizione e un miglioramento della qualità di segmentazione per i modelli tabella e layout. Nonostante ciò, il modello tabellare ha mantenuto un margine di errore più elevato rispetto agli altri due, soprattutto nelle pagine con condizioni conservative peggiori e nelle tabelle più irregolari, a conferma del fatto che la struttura documentaria rappresenta, in questo caso, la componente più complessa da automatizzare.
Completato l'addestramento, i documenti sono stati esportati da Transkribus in formato Excel e sottoposti a una fase di post-processing. Questa fase non è stata considerata un semplice intervento correttivo finale, ma una parte integrante del workflow, indispensabile per trasformare i risultati dell'HTR in un dataset realmente interrogabile. I file esportati sono stati aggregati in un unico foglio di calcolo e normalizzati, soprattutto per quanto riguarda toponimi, abbreviazioni e alcune forme ricorrenti di valore numerico. Il lavoro di normalizzazione ha permesso di ridurre ulteriormente le ambiguità residue e di rendere il dataset adatto a successive interrogazioni e analisi quantitative.

4. RISULTATI RAGGIUNTI

4.1 Modello tabella

Il modello tabella personalizzato per i registi catastali sardi (Table Model v2) è stato creato utilizzando come base di partenza il modello "Amsterdam Tables" fornito da Transkribus.
Durante la fase uno, il team di Linkalab ha addestrato un primo modello, tracciando le tabelle su 50 immagini estratte da registri catastali di 10 Comuni del Marchesato di Quirra[11]. È stata successivamente addestrata una seconda versione del modello, con l'applicazione dei risultati a 100 immagini di ulteriori registri. Nella fase due, il modello è stato applicato su 21 unità archivistiche di registri catastali (una media fra 40 e 170 carte per un totale di 1.120 file immagine); i risultati sono stati corretti manualmente dal team di UniCa[12].

[9] Nel contesto del machine learning e, in particolare, con questo termine si indicano i testi utilizzati dal software per confrontare e affinare le proprie trascrizioni dei documenti digitalizzati. Il ground truth può essere realizzato completamente a mano dall'utente oppure generato in modo semi-automatico, partendo da un modello già esistente (ad esempio un public model sviluppato da altri gruppi di ricerca) e correggendone gli errori. Questo materiale testuale costituisce la base su cui il sistema apprende, per cui le trascrizioni devono essere quanto più accurate possibile, altrimenti il modello apprenderà erroneamente. (*Preparazione dei dati*, s.d.)

[10](*Character Error Rate e curva di apprendimento*, s.d.).

[11] I Comuni di Arbus, Assemini, Genuri, Guspini, Iglesias, Las Plassas, Lunamatrona, Muravera, Pabillonis e Pauli Arbarei.

[12] Nel rispetto dell'ordine dei 20 Comuni selezionati, la correzione è stata realizzata sui primi 35.965 record (su un totale di 90.387) del file excel in cui il team di Linkalab ha riversato i contenuti dei files immagine. Sempre a cura del team di Unica sono stati realizzati alcuni apparati esemplificativi: anzitutto l'elenco normalizzato delle abbreviazioni riscontrate nei diversi registri e l'elenco dei numerosissimi toponimi, per il quale ci si è avvalsi delle informazioni

Al termine dell'addestramento, il modello tabella presentava ancora un errore relativamente alto (CER del 40–50%), causato principalmente dalla forte varietà di condizioni sia materiali che digitali (contrasto irregolare dell'immagine, pieghe, bruciature e strappi di pagina, disallineamenti tra righe e colonne relativi alle operazioni di digitalizzazione). Nonostante ciò, il confronto con il modello base "Amsterdam Tables" ha evidenziato un miglioramento netto: il modello di base spesso non rilevava righe, aggiungeva colonne non presenti ai margini della pagina e mostrava disallineamenti marcati rispetto all'immagine, differentemente dal modello personalizzato in cui, ad eccezione di casi critici, la struttura tabellare risulta coerente e utilizzabile come base per i passaggi successivi
Nel complesso, il modello tabella, pur non raggiungendo ancora livelli di accuratezza pienamente soddisfacenti, ha risposto positivamente - soprattutto nei casi in cui la qualità delle immagini era migliore - nell'identificare la struttura delle tabelle presenti nei registri, segmentando correttamente righe e colonne anche in presenza di layout irregolari, delimitazioni grafiche non sempre visibili e contenuti multi-riga all'interno della stessa cella, riducendo sensibilmente il lavoro di correzione richiesto per ottenere tabelle leggibili e interrogabili.

4.2 Modello layout

Il modello layout personalizzato (Baseline model v1) è stato implementato utilizzando quello base "Universal Lines", progettato come modello generico per il rilevamento di linee su una vasta gamma di tipologie documentarie, ma non specificamente ottimizzato per tabelle miste testo/numero. Nonostante questi presupposti, al termine dell'addestramento il modello ha risposto positivamente alla documentazione sarda, raggiungendo un CER intorno al 9,5% nel riconoscimento dei campi testuali e numerici. Si tratta di un valore che può essere considerato complessivamente buono, considerato che nei registri catastali sono spesso presenti diffusamente cifre, apici e simboli di difficile distinzione. In alcune colonne, specialmente quelle in cui erano presenti esclusivamente dati numerici, il modello di base "Universal Lines" produceva dei falsi positivi individuando linee inesistenti.
Il modello personalizzato ha invece eliminato i falsi positivi e migliorato significativamente la copertura delle colonne numeriche, con errori residui, limitati al mancato riconoscimento di doppie virgolette in celle consecutive (utilizzate per indicare la ripetizione su colonna di una stessa cifra) e di alcuni valori in colonne di stima. Questo risultato ha ridotto considerevolmente il numero di correzioni manuali necessarie per ottenere linee coerenti su cui applicare il modello di trascrizione.

4.3 Modello di trascrizione

Il modello di trascrizione personalizzato (Transcription model v5) è stato sviluppato con un training sui modelli base "Italian - 20th Century Minutes of Mediobanca's Board of Directors & Exec. Committee"[13] e "Transkribus Italian Handwriting M1"[14] che la piattaforma presenta con un CER teorico di partenza rispettivamente del 2.38% e del 6,7%. Alla loro prima applicazione, i modelli hanno presentato un CER molto elevato (50% e 60%) con errori rilevanti su quasi tutte le tipologie di campo, con trascrizioni poco affidabili dei nomi di persona, dei toponimi e dei valori numerici, tanto da risultare inadatti per un uso archivistico diretto su questo specifico tipo di registri. Per questo motivo, l'addestramento è avvenuto talvolta partendo da una revisione manuale dei risultati ottenuti con il modello di base, talaltra trascrivendo in maniera totalmente manuale i registri, in modo da fornire al sistema indicazioni quanto più corrette per il raggiungimento del *ground truth* necessario e adattarlo alle specificità grafiche e linguistiche dei registri catastali sardi. Anche in questo caso, il lavoro è stato graduale nella fase uno, con un addestramento del modello conseguito trascrivendo manualmente un numero sempre maggiore di immagini (10 di tre Comuni; 25 di 3 Comuni; 50 di 5 Comuni; 80 di 10 Comuni) e applicando i risultati a 100 immagini relative a registri di 20 Comuni differenti[15].

normalizzate presenti nel Dizionario dei toponimi della Sardegna disponibile on line attraverso il Sardegna Geoportale e il Servizio ricerca toponimi (Portale Toponimi 1.0.5, s.d.)

[13] Modello creato da Silvia Carboni per trascrivere i verbali manoscritti del Consiglio di Amministrazione e del Comitato Esecutivo di Mediobanca. Il Training Set comprende 13 diverse mani e 90.729 parole (circa 50 pagine per ciascuna mano). È stato utilizzato come modello base uno preesistente e non pubblico, addestrato su 25 pagine di ground truth per ciascuna delle 6 mani più frequenti nei verbali.

[14] Modello generico, sviluppato su diverse tipologie di scritture manoscritte in lingua italiana del XVI-XIX secolo. Si tratta di un modello generico addestrato su un dataset diversificato.

Successivamente a questi cicli di training, il modello personalizzato, basato sul modello base Transkribus "Italian Handwriting M1" ha raggiunto un CER tra l'8 e il 10%. Sebbene possa sembrare un valore ancora relativamente alto, va considerato che si tratta di un margine di errore di riconoscimento che si riferisce a documenti stilati da diverse persone, quindi con grafie differenti. In generale, pur essendo ancora perfettibile, il modello personalizzato individua e trascrive correttamente le parole nelle intestazioni e nei titoli di colonna e gran parte dei contenuti testuali e numerici, con errori concentrati specialmente su numeri, apici e alcune grafie particolarmente irregolari. Dal punto di vista operativo, ciò permette di utilizzare le trascrizioni come base per ricerche per parola chiave e analisi quantitative, limitando l'intervento umano alla revisione di casi problematici e alla normalizzazione dei dati.

5. CONCLUSIONI

L'esperimento ha mostrato che anche con un numero limitato di unità è possibile ottenere modelli personalizzati in grado di migliorare sensibilmente le prestazioni rispetto ai modelli base di Transkribus, in particolare per il riconoscimento delle linee e per la trascrizione. Analizzando i risultati raggiunti, non emerge tanto la possibilità di automatizzare integralmente il trattamento dei registri catastali, quanto piuttosto la possibilità di costruire un flusso di lavoro semi-automatico in cui l'HTR riduce in modo consistente il carico di lavoro umano, pur non eliminando la necessità di controllo, correzione e normalizzazione.
In questo senso, il valore principale dei modelli personalizzati non consiste nella sostituzione di chi trascrive la documentazione, ma nella possibilità di rendere possibile la trasformazione dei registri in un dataset interrogabile e riusabile. Relativamente ai base model, inoltre, è utile evidenziare che seppur la loro scelta può essere stata metodologicamente decretata in base alla prossimità tipologica, cronologica e grafica rispetto ai documenti da trattare non sempre può garantire buone prestazioni sul nuovo corpus.
In conclusione, la sperimentazione condotta sui registri catastali sardi del primo Novecento ha mostrato che l'impiego di Transkribus può produrre risultati utili anche in presenza di documentazione archivistica complessa.
I risultati ottenuti consentono di distinguere con maggiore chiarezza il comportamento dei tre moduli impiegati. Il modello layout e il modello di trascrizione, una volta personalizzati, hanno raggiunto livelli di accuratezza sufficienti per sostenere attività di interrogazione, ricerca per parola chiave e analisi quantitativa su larga scala, pur mantenendo un margine di errore non trascurabile nei campi numerici e nei segni abbreviativi. Il modello tabella, invece, si è confermato il segmento più problematico dell'intero flusso di lavoro, poiché la varietà dei layout, la presenza di celle irregolari e le condizioni materiali dei registri hanno inciso in modo significativo sulla stabilità della segmentazione.
Dal punto di vista archivistico, il guadagno principale risiede nella diminuzione del margine di errore che porta all'ottenimento di trascrizioni sufficientemente accurate che permettono di svolgere interrogazioni mirate e analisi su larga scala.
Tale esperienza, dunque, suggerisce come sia ancora necessario combinare i modelli HTR con strategie avanzate di normalizzazione. Al netto di queste limitazioni, l'uso di Transkribus e dei modelli personalizzati permette di trasformare semplici digitalizzazioni in un *dataset* strutturato interrogabile e integrabile con altre basi dati e si rivela, dunque, una soluzione praticabile ed efficace per l'interrogazione dei dati derivanti da registri catastali storici ad alta complessità. Questa stessa metodologia, a partire dalla definizione del *ground truth* all'addestramento dei modelli, fino alla normalizzazione dei dati esportati, può essere riapplicata, adattandola, a tipologie di registri amministrativi di tipologia simile o ad altri contesti archivistici.

BIBLIOGRAFIA

Archivio di Stato di Cagliari. (s.d.). SIAS. Recuperato 7 febbraio 2026, da https://sias-archivi.cultura.gov.it/cgi-bin/pagina.pl?TipoPag=strumcorr&Chiave=%2033254&RicProgetto=as%2dcagliari

Bazzaco, S. (2024). La trascrizione automatica di documenti a stampa antichi. Appunti per un modello di riconoscimento della tipografia in corsivo. *DigItalia*, *19*(1), 63–86. https://doi.org/10.36181/digitalia-00094

[15] I Comuni di Assemini, Barrali, Decimoputzu, Elmas, Goni, Gonnesa, Guamaggiore, Lunamatrona, Monastir, Pirri, Portoscuso, Pula, S. Pietro Pula, Salto di Quirra, Sarroch, Segaiu, Setzu, Seurgus, Sicci e Villasimius.

Carboni, S., & Molino Lova, T. (2025). L'utilizzo dell'Handwritten Text Recognition nei sistemi di consultazione degli archivi: Il caso dell'Archivio Storico Mediobanca "Vincenzo Maranghi" e Transkribus. *I quaderni de Il mondo degli archivi*, *1*. https://www.ilmondodegliarchivi.org/i-quaderni-de-il-mondo-degli-archivi-1-2025-lutilizzo-dellhandwritten-text-recognition-nei-sistemi-di-consultazione-degli-archivi-il-caso-dellarchivio-storico-me/

DH.UniCa. (s.d.). DH UNICA. Recuperato 7 febbraio 2026, da https://dh.unica.it/

Homepage. (2024, settembre 27). Linkalab. https://linkalab.it/

Iorio, S. (2019). L'utilizzo della piattaforma Transkribus nell'Archivio Storico del Banco di Napoli: Il "Progetto Pandetta". *Quaderni dell'Archivio Storico - nuova serie online*, *1*(1), 195–208.

Kiessling, B., Tissot, R., Stokes, P., & Stökl Ben Ezra, D. (2019). eScriptorium: An Open Source Platform for Historical Document Analysis. *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, *2*, 19–19. https://doi.org/10.1109/ICDARW.2019.10032

Loss, E., Guernaccini, F., & Carassai, M. (2025). From Manuscript to Metadata: Experiments on Handwritten Text Recognition, Tagging and Importation for the Memoriali series (1265-1452). *JLIS.It*, *16*(2), 59–85. https://doi.org/10.36253/jlis.it-641

Muehlberger, G., Seaward, L., Terras, M., Ares Oliveira, S., Bosch, V., Bryan, M., Colutto, S., Déjean, H., Diem, M., Fiel, S., Gatos, B., Greinoecker, A., Grüning, T., Hackl, G., Haukkovaara, V., Heyer, G., Hirvonen, L., Hodel, T., Jokinen, M., … Zagoris, K. (2019). Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study. *Journal of Documentation*, *75*(5), 954–976. https://doi.org/10.1108/JD-07-2018-0114

Nockels, J., Gooding, P., Ames, S., & Terras, M. (2022). Understanding the application of handwritten text recognition technology in heritage contexts: A systematic review of Transkribus in published research. *Archival Science*, *22*(3), 367–392. https://doi.org/10.1007/s10502-022-09397-0

Romein, C. A., Rabus, A., Leifert, G., & Ströbel, P. B. (2025). Assessing advanced handwritten text recognition engines for digitizing historical documents. *International Journal of Digital Humanities*, *7*(1), 115–134. https://doi.org/10.1007/s42803-025-00100-0

Sánchez, J. A., Mühlberger, G., Gatos, B., Schofield, P., Depuydt, K., Davis, R. M., Vidal, E., & de Does, J. (2013). tranScriptorium: A european project on handwritten text recognition. *Proceedings of the 2013 ACM symposium on Document engineering, DocEng '13*, 227–228. https://doi.org/10.1145/2494266.2494294

Schwarz-Ricci, V. I. (2022). Handwritten Text Recognition per registri notarili (secc. XV-XVI): Una sperimentazione. *Umanistica Digitale*, (13), 171–181. https://doi.org/10.6092/issn.2532-8816/14926

Sichera, P., Grosso, A. M. D., Mazzagufo, L., & Spampinato, D. (2025). ZoneRW: verso un'integrazione con Kraken ed eScriptorium per il riconoscimento e la gestione avanzata delle regioni di interesse. In S. Rebora, M. Rospocher, & S. Bazzaco (A c. Di), *Diversità, Equità e Inclusione: Sfide e Opportunità per l'Informatica Umanistica nell'Era dell'Intelligenza Artificiale, Proceedings del XIV Convegno Annuale AIUCD2025*. AIUCD.

Terras, M., Anzinger, B., Gooding, P., Mühlberger, G., Prien, M., Nockels, J., Romein, C. A., Stauder, A., & Stauder, F. (2025). The artificial intelligence cooperative: READ-COOP, Transkribus, and the benefits of shared community infrastructure for automated text recognition. *Open Research Europe*, *5*, 16. https://doi.org/10.12688/openreseurope.18747.2

Via al progetto «Transkribus»: Trascrizione automatica documenti tramite intelligenza artificiale. (2025, luglio 1). Archivi. https://archivi.cultura.gov.it/archivio-notizie/notizia?tx_news_pi1%5Baction%5D=detail&tx_news_pi1%5Bcontroller%5D=News&tx_news_pi1%5Bnews%5D=2064&cHash=a028de822f75028c8b6fcb2c167b600c

Ricostruzione virtuale di un patrimonio disperso: le collezioni di disegni a Genova nel XIX secolo

GIULIA PILOSU
UNIVERSITÀ DI GENOVA, ITALIA – GIULIA.PILOSU@EDU.UNIGE.IT

ABSTRACT (ITALIANO)

Il progetto presentato si concentra sulla ricostruzione digitale delle collezioni di disegni presenti a Genova nel XIX secolo, un ambito ancora in parte inesplorato rispetto alle raccolte nobiliari cittadine di dipinti e statuaria che si sviluppano nei secoli precedenti, a partire dal Cinquecento. Sebbene in questo periodo il collezionismo grafico assuma caratteri più specifici, coinvolgendo tanto l'aristocrazia quanto il ceto artistico, la natura fortemente privata e personale delle raccolte ha portato in diversi casi alla massiccia dispersione del patrimonio alla morte del collezionista. L'indagine si sofferma sulle vicende collezionistiche dello scultore e professore accademico Santo Varni e del marchese Marcello Durazzo: due percorsi speculari che permettono di testare diverse tipologie di virtual reunification su nuclei con genesi e destini diversi. L'obiettivo finale è la ricomposizione dell'assetto originale di questi complessi grafici, oggi frammentati tra istituzioni pubbliche, mercato antiquario e raccolte private. Metodologicamente, la ricerca si fonda sulla definizione di una struttura di dati capace di connettere le diverse fonti disponibili, tra cui cataloghi d'asta, archivi personali e carteggi inediti, per poter mappare lo spostamento delle opere. L'esito ultimo è lo sviluppo di un database, concepito come catalogo virtuale, capace di ricomporre l'unità delle collezioni e di restituirne l'accessibilità, al fine di comprendere le scelte collezionistiche e la visione complessiva delle raccolte.
Parole chiave: Virtual reunification; Collezionismo grafico; Patrimonio disperso; Mercato artistico genovese.

ABSTRACT (ENGLISH)

Virtual reconstruction of a lost heritage: the collections of drawings in Genoa in the 19th century.
The project focuses on the digital reconstruction of the collections of drawings present in Genoa in the 19th century, an area that is still largely unexplored compared to the city's noble collections of paintings and statuary that developed in previous centuries, starting in the 16th century. Although graphic collecting took on more specific characteristics during this period, involving both the aristocracy and the artistic class, the highly private and personal nature of the collections led in certain cases to the massive dispersion of the heritage upon the death of the collector. The study focuses on collecting activities of the sculptor and academic professor Santo Varni and the marquis Marcello Durazzo: two mirror-image paths that allow us to evaluate several types of virtual reunification on nuclei with different origins and destinies. The goal is to recompose the original structure of these graphic complexes, now fragmented between public institutions, the antiques market, and private collections. Methodologically, the research is based on the definition of a data structure capable of connecting the various available sources, including auction catalogues, personal archives, and unpublished correspondence, to map the movement of the works. The outcome is the development of a database, conceived as a virtual catalogue, capable of recomposing the unity of the collections and restoring their accessibility, to understand the collecting choices and the overall vision of the collections.
Keywords: Virtual reunification; Graphic collections; Dispersed heritage; Genoese art market.

1. STATO DELL'ARTE

Il fenomeno del collezionismo di disegni a Genova nel XIX secolo rappresenta un elemento essenziale per comprendere lo sviluppo culturale della città in un'epoca di profondi mutamenti sociali e, di conseguenza, artistici. Già nel corso del Seicento, epoca d'oro per l'arte e l'ambiente culturale genovese, si hanno notizie di raccolte grafiche di notevole interesse, spesso legate alle diverse botteghe dei principali pittori o a mecenati aristocratici. Tuttavia, tali nuclei presentano evidenti criticità strutturali per il loro studio: la consistenza delle collezioni appare, almeno dagli inventari a noi pervenuti, contenuta e la documentazione d'archivio, lacunosa e poco precisa, riflette una percezione del disegno ancora

subordinata alle "arti maggiori". Questa natura rende perciò ad oggi estremamente complessa la ricostruzione sistematica di tali raccolte.
È invece tra il Settecento e l'Ottocento che il collezionismo di opere su carta a Genova, come in tutta Europa, vive una metamorfosi sostanziale, assumendo una specificità e una capillarità senza precedenti. In questo periodo, le raccolte smettono di essere semplici appendici delle quadrerie per diventare realtà autonome e specializzate, crescendo esponenzialmente nel numero di pezzi. Figure come Marcello Durazzo e Santo Varni arrivano a radunare diverse migliaia di fogli, creando corpora che per ricchezza e vastità superano quelli visti dei secoli precedenti. Tuttavia, la loro natura intrinsecamente privata e personale ne ha favorito in molti casi lo smembramento dopo la scomparsa dei proprietari. Il caso di Santo Varni, la cui immensa collezione fu dispersa quasi interamente all'asta nel 1887, è in tal senso emblematico della fragilità di questo patrimonio.
La storiografia sul collezionismo genovese ha tradizionalmente privilegiato il cosiddetto "Secolo dei Genovesi", concentrandosi sulle grandi quadrerie di età moderna e sugli assetti patrimoniali delle famiglie nobiliari (Boccardo, 2004). Sebbene contributi fondamentali (Raggio, 2000; Montaldo Spigno, 1996) abbiano tracciato un quadro rigoroso delle dinamiche collezionistiche tra Ancien Régime e Restaurazione, l'attenzione dedicata specificamente alla produzione grafica resta ancora oggi episodica. Anche le ricerche più recenti sull'ambiente artistico ligure del XIX secolo (Vazzoler, 2013), hanno privilegiato il contesto politico ed economico o la produzione pittorica, lasciando in ombra la fisionomia di raccolte che proprio in questo momento acquisivano una nuova autonomia.
Il panorama degli studi attuali riflette dunque la frammentarietà di questo patrimonio: gran parte della letteratura si è focalizzata su singoli aspetti biografici (Montaldo Spigno, 2001) o su nuclei legati a specifiche occasioni espositive (Cavelli & Sborgi, 1985; Cavelli Traverso & Olcese Spingardi, 1996), ma manca ancora un'analisi sistematica che metta in relazione la circolazione dei disegni con le reti del mercato antiquario. Tale lacuna è accentuata dalla natura stessa delle fonti primarie: i cataloghi d'asta ottocenteschi[1] e gli inventari d'archivio offrono spesso descrizioni identificative sommarie, rendendo complesso il compito di ricondurre i fogli dispersi alle loro serie originarie. Persino nei casi di collezioni rimaste in gran parte integre o donate a istituzioni cittadine, la critica (Cabrini et al., 2018; Pastorino, 2021) ha spesso privilegiato l'indagine su altre tipologie di opere d'arte presenti nelle collezioni, come le raccolte archeologiche o le stampe, relegando il disegno ai margini dell'analisi.
In questo scenario, la sfida non è più soltanto documentaria ma metodologica: l'opzione di una virtual reunification (Punzalan, 2014; Assonitis, 2020) si impone come l'unica strategia possibile per recuperare una visione unitaria di queste raccolte. Risulta evidente la necessità di una ricerca che adotti una prospettiva capace di ricomporle virtualmente, trasformando inventari lacunosi in un ecosistema digitale integrato. Attraverso la strutturazione digitale dei dati è possibile, perciò, restituire coerenza a un patrimonio frammentato e difficilmente leggibile nella sua interezza, rendendolo nuovamente accessibile.
Nel panorama internazionale, la scelta di un approccio digitale per la ricostituzione di grandi corpora grafici è visibile in alcuni esempi di grande rilievo, primo fra tutti quello della Fondation Custodia a Parigi. L'operazione di digitalizzazione condotta sull'interezza della collezione di Frits Lugt, collezionista e studioso vissuto tra Ottocento e Novecento, e la successiva pubblicazione online, rappresenta un punto di riferimento metodologico imprescindibile: la piattaforma non si limita alla semplice catalogazione, ma garantendo una visione d'insieme di una raccolta personale, permette di comprendere l'intento intellettuale e le scelte del collezionista.

2. CASI STUDIO E DIFFERENTI METODOLOGIE

La validità del modello di ricerca proposto trova riscontro nell'applicazione a casi studio cardine del panorama genovese, ciascuno dei quali richiede un approccio metodologico specifico, dettato dallo stato di conservazione attuale e dalla sorte subita dalle opere nel tempo. Questa diversificazione della metodologia permette di testare come il modello possa adattarsi a contesti differenti.
Il caso dello scultore Santo Varni introduce un elemento di sfida metodologica legato alla gestione di una collezione oggi profondamente frammentata. Parliamo di un corpus vasto, comprendente oltre 2000

[1] Vedasi il catalogo della vendita della collezione di Santo Varni: Catalogo della collezione Santo Varni di Genova: marmi, gessi e terre cotte, porcellane e maioliche, armi, ferri, bronzi, monete e medaglie, oggetti diversi e da vetrina, antichità classiche, quadri, miniature, stampe, disegni: di cui la vendita al pubblico incanto avrà luogo in Genova [...] lunedì 14 novembre 1887 [...]. Milano: Tip. Luigi di Giacomo Pirola, 1887.

opere su carta: se alcuni piccoli nuclei sono approdati, dopo la vendita all'asta, nelle collezioni pubbliche genovesi in diversi momenti e con differenti donazioni, lasciti e acquisti, una vasta parte della raccolta è confluita nel mercato antiquario. Il principale ostacolo alla ricomposizione è rappresentato dalla natura sommaria degli inventari dell'asta ottocentesca, le cui descrizioni generiche rendono complesso ricondurre i singoli fogli alle cartelle in cui originariamente erano conservati. Tuttavia, l'abitudine sistematica di Santo Varni di apporre la propria firma sul recto delle opere permette oggi di intercettare questi disegni nelle aste contemporanee e nelle istituzioni pubbliche. Proprio per questa sua fisionomia dispersa, la ricerca non ambisce a presentare un database statico o definitivo, ma si configura come un sistema digitale in continua espansione. Il lavoro mira a creare una struttura aperta, capace di accogliere nuove scoperte a mano a mano che i fogli riemergono sul mercato o in collezioni private ancora ignote. Integrando le opere già individuate con il fondo di lettere e appunti privati conservato presso l'Accademia Ligustica, il database si pone l'obiettivo di mappare, accanto alla raccolta in sé, la fitta rete di scambi che legava lo scultore al mercato artistico. Attraverso l'analisi dei carteggi, è possibile risalire non solo alla provenienza dei disegni, ma anche alle motivazioni critiche che spingevano Varni verso determinati acquisti. In quest'ottica, la struttura dei dati potrà sfruttare i linked open data, utilizzando l'ontologia CIDOC-CRM per formalizzare le relazioni tra fogli, eventi d'asta e passaggi di proprietà (Hyvönen, 2012). L'uso di identificativi univoci (URIs) permetterà di connettere i frammenti della collezione a dataset internazionali, trasformando il database in un grafo dinamico capace di integrare automaticamente nuove occorrenze dei disegni disperse nel mercato globale. Si tratta, dunque, di un cantiere di ricerca permanente: uno strumento che, pur consapevole delle lacune ancora esistenti, fornisce la base per un censimento che continuerà ad aggiornarsi nel tempo, trasformando ogni nuovo ritrovamento in un tassello essenziale per la ricostruzione di una situazione profondamente complessa.
Complementare è la ricerca dedicata al Marchese Marcello Durazzo: in questa situazione il patrimonio, per volere del collezionista, è rimasto unito (viene donato infatti nel 1848 alla biblioteca cittadina e confluisce in seguito alla loro fondazione nella collezione dei Musei di Strada Nuova a Genova), ma ha perso la sua forma originaria a causa delle scelte catalografiche di inizio Novecento che hanno modificato sostanzialmente l'ordine delle cartelle voluto dal collezionista per imporre un rigido criterio di catalogazione alfabetica. Questa operazione ha di fatto modificato e reso complessa la comprensione dei legami logici e le affinità che Durazzo aveva stabilito tra le opere della sua raccolta. In questo contesto, il progetto agisce come un'operazione di archeologia digitale volta a decostruire l'attuale assetto museale per ritrovare e comprendere la visione originale del collezionista. A tal fine, è stata già completata la fase di analisi della struttura originaria delle cartelle Durazzo, propedeutica alla creazione dell'interfaccia di navigazione del portale. Il focus dell'intervento sarà la creazione di una sezione dedicata all'interno del portale dei Musei di Strada Nuova, progettata per rispecchiare fedelmente la divisione in cartelle originaria. Questa interfaccia permetterà di visualizzare le opere non come singoli oggetti isolati, ma all'interno della loro cornice storica, esaltando le attribuzioni originali ottocentesche, in alcuni casi sostanzialmente diverse da quelle attuali e possono essere un dato interessante per comprendere come si è sviluppato lo studio e l'evoluzione critica in quasi due secoli. Tale operazione è fondamentale per comprendere il gusto dell'epoca, restituendo dignità a una logica intellettuale che il passaggio in collezioni più complesse ha messo in secondo piano. Questa pagina specifica sarà messa in costante dialogo direttamente con le schede tecniche dei vari disegni, aggiornate e implementate nel catalogo online dei Musei di Genova[2]. In questo modo, si potrà trasformare il portale in un luogo interconnesso, dove la ricerca scientifica sulla collezione storica si traduce in una nuova esperienza di navigazione, capace di far dialogare l'occhio ottocentesco di Durazzo con la nostra sensibilità moderna. L'integrazione tra i sistemi permetterà infine di generare nuove connessioni, rendendo il patrimonio grafico genovese un modello di riferimento per la valorizzazione digitale delle collezioni grafiche.

3. CONCLUSIONI E PROSPETTIVE FUTURE

L'integrazione dei casi studio analizzati dimostra come la strutturazione digitale delle relazioni tra opere documenti costituisca oggi un metodo critico imprescindibile per il recupero della memoria storica di patrimoni complessi. Il progetto ha evidenziato la necessità di adottare strategie differenziate per rispondere a dinamiche di dispersione eterogenee: dal monitoraggio dei flussi del mercato antiquario per i

[2] https://catalogo.museidigenova.it/

fogli autografi di Santo Varni, fino alla ricostituzione dell'assetto originario per il nucleo di Marcello Durazzo, rimasto in gran parte integro ma soggetto nel tempo a riclassificazioni.
Le prospettive future della ricerca si focalizzeranno sulla finalizzazione di un sistema capace di superare la staticità del cataloghi tradizionali e di ricomporre la logica intellettuale che legava le opere nelle collezioni storiche. L'implementazione in particolare della struttura della raccolta di Marcello Durazzo all'interno del portale ufficiale dei Musei di Genova garantirà una ricaduta concreta alla ricerca, rendendo accessibile un patrimonio frammentario e poco comprensibile nella sua attuale forma. In definitiva, la creazione di questo ecosistema mira a ristabilire il legame tra la produzione grafica e il contesto culturale genovese del XIX secolo, offrendo una diversa chiave di lettura per la comprensione del collezionismo in un'epoca di profonde trasformazioni. L'uso dei linked data diventa così un metodo tangibile per ristabilire i legami originali tra l'opera e il suo contesto, garantendo l'interoperabilità e rendendo la virtual reunification un modello riproponibile per altri patrimoni grafici dispersi.

BIBLIOGRAFIA

Assonitis, A. (2020). The Expanding Role of Digitized Collections: The Medici Archive. Brown, K. (Eds.), The Routledge companion to digital humanities and art history. New York: Routledge, 266-274.

Boccardo, P. (2004). L'età di Rubens: dimore, committenti e collezionisti genovesi. Milano: Skira.

Cabrini, L., Extermann, G., Fontanarossa, R. (2018). Santo Varni: conoscitore, erudito e artista tra Genova e l'Europa. Chiavari: Società Economica di Chiavari.

Cavelli, C., Sborgi, F. (1985), Santo Varni scultore (1807-1885): Genova, Museo dell'Accademia ligustica di belle arti, 9 maggio-30 giugno 1985. Genova: Sagep.

Hyvönen, E. (2012). Publishing and Using Cultural Heritage Linked Data on the Semantic Web. Synthesis Lectures on the Semantic Web: Theory and Technology. 2 (1). 1-159. DOI:10.2200/S00452ED1V01Y201210WBE003

Montaldo Spigno, M.G. (1996). Corrispondenze d'arte : Genova nella prima metà dell'Ottocento. Genova: Brigati.

Montaldo Spigno, M.G. (2001). Le collezioni d'arte del marchese Marcello Durazzo d'Ippolito. Genova: Brigati.

Olcese Spingardi, C., Cavelli Traverso, C. (1996). Studi dall'Archivio dell'Accademia: Santo Varni. Quaderni del Museo dell'Accademia, 20.

Pastorino, A.M. (2021). La Collezione di Sculture Antiche di Santo Varni. Genova: Sagep.

Punzalan, R.L. (2014). Understanding Virtual Reunification. The Library Quarterly: Information, Community, Policy, 84, N. 3, 294-323. DOI:10.1086/676489

Raggio, O. (2000). Storia di una passione: cultura aristocratica e collezionismo alla fine dell'Ancien Régime. Venezia: Marsilio.

Vazzoler, M. (2013). Genova tra Rivoluzione e Impero : patrimonio artistico, mercato dell'arte, progetti museografici. Firenze: Edifir.

Catalogo della collezione Santo Varni di Genova: marmi, gessi e terre cotte, porcellane e maioliche, armi, ferri, bronzi, monete e medaglie, oggetti diversi e da vetrina, antichità classiche, quadri, miniature, stampe, disegni: di cui la vendita al pubblico incanto avrà luogo in Genova [...] lunedì 14 novembre 1887 [...]. Milano: Tip. Luigi di Giacomo Pirola, 1887.

Dai workflow descrittivi ai workflow applicativi tramite AEON e l'IA

Francesco Pinna[1], Emiliano Degl'Innocenti[2]

[1]CNR Istituto Opera del Vocabolario Italiano, Consiglio Nazionale delle Ricerche (OVI-CNR), Italia – francesco.pinna@cnr.it

[2]CNR Istituto Opera del Vocabolario Italiano, (OVI-CNR), Italia – emiliano.deglinnocenti@cnr.it

ABSTRACT (ITALIANO)

AEON (dAriah sErvice Oriented iNfrastructure) è la piattaforma web sviluppata da DARIAH-IT in H2IOSC per catalogare servizi e comporre, validare ed eseguire workflow applicativi nelle Scienze Umane e Sociali, in coerenza con i principi FAIR e l'ecosistema EOSC. Il paper analizza il divario tra workflow descrittivi – ricchi di contesto e utili alla trasparenza metodologica e alla discovery (come quelli dello SSHOC Marketplace), e pipeline eseguibili, che richiedono contratti di dato espliciti, specifiche API e controlli formali. Proponiamo una metodologia in cui un chatbot/LLM agent legge un workflow narrativo, segmenta gli step, distingue componenti computabili e concettuali, estrae azioni, entità e vincoli, quindi mappa ogni azione sui servizi AEON vincolandosi ai JSON manifest (input/output, formati, schemi, parametri e requisiti d'accesso). L'agente genera una proto-pipeline (Pipeline.json e bozza XML per WSO2 Micro Integrator/CAR) che viene sottoposta ai controlli del Connector e al ciclo di validazione, generazione e deploy di AEON, preservando tracciabilità, riproducibilità e governance. L'approccio è esemplificato sul workflow "Automatic Text Recognition Roadmap", mostrando come la narrazione possa essere tradotta in una catena minima di servizi (ingestion → HTR → post-processing → TEI export). Discutiamo infine implicazioni e limiti (mediazione semantica, granularità degli step, lacune nei metadati) e delineiamo sviluppi futuri verso un "workflow translator agent" integrato nativamente in AEON.
Parole chiave: AEON; workflow descrittivi; agente LLM, DARIAH-IT, H2IOSC, infrastrutture critiche.

ABSTRACT (ENGLISH)

From descriptive workflows to application workflows using AEON and AI

AEON (dAriah sErvice Oriented iNfrastructure) is the web platform developed by DARIAH-IT in H2IOSC to catalogue services and compose, validate and execute application workflows in the Humanities and Social Sciences, in line with FAIR principles and the EOSC ecosystem. The paper analyses the gap between descriptive workflows – rich in context and useful for methodological transparency and discovery (such as those of the SSHOC Marketplace) – and executable pipelines, which require explicit data contracts, specific APIs and formal controls. We propose a methodology in which a chatbot/LLM agent reads a narrative workflow, segments the steps, distinguishes computable and conceptual components, extracts actions, entities and constraints, and then maps each action to AEON services by binding to JSON manifests (input/output, formats, schemas, parameters and access requirements). The agent generates a proto-pipeline (Pipeline.json and XML draft for WSO2 Micro Integrator/CAR) which is subjected to Connector checks and the AEON validation, generation and deployment cycle, preserving traceability, reproducibility and governance. The approach is exemplified in the "Automatic Text Recognition Roadmap" workflow, showing how the narrative can be translated into a minimal chain of services (ingestion → HTR → post-processing → TEI export). Finally, we discuss implications and limitations (semantic mediation, granularity of steps, gaps in metadata) and outline future developments towards a "workflow translator agent" natively integrated into AEON.
Keywords: AEON; descriptive workflow; LLM agent, DARIAH-IT, H2IOSC, critical infrastructures.

1. INTRODUZIONE

DARIAH-IT è il nodo italiano di DARIAH-EU (Digital Research Infrastructure for the Arts and Humanities, 2024), infrastruttura europea che supporta le comunità di ricerca nelle arti e scienze umane con servizi digitali federati e pratiche di interoperabilità. Insieme ai nodi italiani delle infrastrutture di ricerca CLARIN, E-RIHS e OPERAS, ha realizzato la federazione di quattro infrastrutture di ricerca afferenti al dominio

delle scienze umane e sociali in Italia attraverso il progetto PNRR H2IOSC (Humanities and cultural Heritage Italian Open Science Cloud). Nel contesto del progetto H2IOSC, DARIAH-IT ha progettato e implementato AEON, dAriah sErvice Oriented iNfrastrucure, la piattaforma che consente di comporre visivamente, eseguire e condividere workflow applicativi nelle Scienze Umane e Sociali (SSH), integrando servizi digitali eterogenei nel rispetto dei principi FAIR (Findable, Accessible, Interoperable and Reusable, Wilkinson et al., 2016) e in coerenza con l'ecosistema EOSC (Degl'Innocenti & Pinna, 2025). Grazie alle funzionalità del Composer e del
Connector, si verifica la compatibilità tra input e output dei servizi inanellati nei workflow, e se ne generano gli script di orchestrazione; l'esecuzione è mediata da componenti d'integrazione basati su WSO2 (API Manager e Micro Integrator)[1] che instradano e trasformano i dati, garantendo tracciabilità e riproducibilità. Sicurezza e privacy sono assicurate da autenticazione e autorizzazione federate (OpenID Connect/OAuth2)[2], parallelamente a gestione granulare di ruoli e permessi.

1.1 Obiettivi della ricerca

Partendo dalla presentazione di AEON, viene proposto un potenziamento che consenta di derivare workflow applicativi da descrizioni testuali, valorizzando i workflow descrittivi come blueprint.

1.2 Contesto e stato dell'arte

Lo SSHOC Marketplace[3] espone workflow descrittivi contestualizzati (insieme a tool, dataset e training) utili alla trasparenza metodologica e alla discovery, ma non nativamente eseguibili; resta quindi critico il divario tra descrizione e automazione.

1.3 Contributo del paper

Descriviamo un chatbot/LLM agent che estragga azioni, entità computabili e vincoli dai testi, li mappi sui servizi del catalogo AEON vincolandosi ai manifest, e generi una proto-pipeline (Pipeline.json → XML WSO2/CAR) da sottoporre ai controlli del Connector e a deploy controllato in piattaforma (Rubies, 2026, sezione "3. Crea il tuo chatbot IA").

2. AEON COME WORKFLOW MANAGER PER LE DIGITAL HUMANITIES

AEON, sviluppata da DARIAH-IT nell'ambito di H2IOSC, è un ambiente web che consente di pubblicare servizi in un catalogo interno e di comporre, validare ed eseguire workflow applicativi, ossia pipeline eseguibili che orchestrano servizi via API (Sichera et al., 2024), nelle SSH. L'obiettivo è ridurre la barriera tecnica, offrendo un'esperienza in cui l'utente progetta la sequenza operativa a partire da servizi già descritti e governati, senza intervenire direttamente su script o configurazioni di integrazione (Degl'Innocenti & Pinna, 2025). Dal punto di vista architetturale, AEON adotta un modello modulare basato su: (i) un Flow Designer UI (user interface) per la composizione visuale a nodi (drag&drop) di servizi REST e mediatori; (ii) un backend (Spring Boot) che gestisce il ciclo di vita delle pipeline; (iii) un layer di API management e integrazione basato su WSO2 (API Manager e Micro Integrator) per l'orchestrazione; (iv) un'infrastruttura di autenticazione e autorizzazione[4] federata (OIDC/OAuth2, tramite Keycloak[5]) con ACL[6] per la governance degli accessi a servizi e workflow. Nel Flow Designer l'utente costruisce il flusso su un canvas, configura parametri e connessioni e avvia procedure di salvataggio, validazione e test. La pipeline è serializzata in un file denominato Pipeline.json e inviata al backend, il quale: interpreta e ordina i nodi secondo le dipendenze, verifica requisiti minimi e coerenza dei parametri, quindi compila la rappresentazione in XML eseguibile per WSO2 Micro Integrator

[1] Per approfondire: https://mi.docs.wso2.com/en/4.2.0/develop/api-led-integration (cons. 10/01/2026).

[2] OpenID Connect / OAuth 2.0: standard di protocolli per autenticazione (OIDC) e autorizzazione delegata (OAuth 2.0).

[3] Social Sciences & Humanities Open Cloud Marketplace, catalogo online di risorse (tool, servizi, dataset, training e workflow descrittivi) per la comunità SSH; https://marketplace.sshopencloud.eu (cons. 10/01/2026).

[4] AAI (Authentication and Authorization Infrastructure): sistema che gestisce l'identità digitale degli utenti e i permessi di accesso alle risorse.

[5] Piattaforma open source di Identity and Access Management (IAM) per Single Sign-On e gestione di identità, ruoli e token (supporta OIDC/OAuth 2.0).

[6] Access Control List (in italiano: "lista di controllo degli accessi"): elenco di permessi associati a una risorsa (ad es. lettura/scrittura/esecuzione).

(eventualmente impacchettabile in file CAR[7]), predisponendo il deploy sul runtime. In parallelo, il backend genera gli artefatti necessari alla pubblicazione e al riuso (es. documentazione OpenAPI e manifest), e gestisce lo stato della pipeline (Draft/Ready/In Catalog) e le operazioni CRUD[8] nel rispetto delle ACL. L'esecuzione si fonda sul paradigma a mediatori di WSO2 Micro Integrator (in AEON sono disponibili call, property, enrich, payload factory, filter, respond[9]), che consente routing, trasformazione dei messaggi, logging e gestione degli errori in modo deterministico e tracciabile, favorendo riproducibilità e controllo. In questo quadro, AEON si concentra sui workflow applicativi (eseguibili), mentre i workflow descrittivi, non gestiti in AEON, rimangono strumenti metodologici utili a documentare e contestualizzare processi di ricerca: la distinzione abilita una divisione del lavoro in cui la narrazione guida la progettazione, ma l'automazione avviene solo dove esistono servizi invocabili e contratti di dato espliciti. In AEON, nell'ambiente di creazione dei workflow la funzionalità del Connector supporta la composizione guidata tramite controlli di compatibilità (I/O già operativa; conformità di schema e coerenza semantica come estensioni progettate), riducendo errori in fase di design e prevenendo failure a runtime.

3. I WORKFLOW NELLO SSHOC E NELL'H2IOSC MARKETPLACE

Questa sezione chiarisce la differenza tra workflow descrittivi (scenari metodologici) e workflow applicativi (pipeline eseguibili), mettendo a confronto SSHOC Marketplace e H2IOSC Marketplace.[10]

3.1 Orchestrazione nell'H2IOSC Marketplace

Nella versione di staging l'orchestrazione end-to-end di pipeline applicative non risulta ancora disponibile. L'impostazione corrente è un flusso di instradamento basato su WSO2 API Manager con integrazione SSO/IdP[11], che inoltra richieste verso servizi remoti: non è previsto un motore univoco per pipeline né una UI low-code per comporre flussi. La definizione dei "flussi" richiede la produzione e il caricamento di configurazioni XML di mediazione (scripting ESB[12]) e l'import di OpenAPI dai manifest. In questo quadro, restano distintive di AEON la composizione visuale a nodi, i controlli guidati di compatibilità e la fruizione dei log in interfaccia (Sichera et al., 2024).

3.2 Tipologia dei workflow nello SSHOC Marketplace

Nel Marketplace SSHOC i workflow sono roadmaps di ricerca a step, orientate al ciclo di vita dei dati (Barbot et al., 2024): descrizioni testuali e collegamenti a tool, servizi, dataset, pubblicazioni e materiali formativi. La riproducibilità è soprattutto metodologica: convivono passi concettuali (scelte di standard, valutazioni qualitative, pianificazione) e passi tecnici, senza vincolare l'utente a uno specifico stack. Una parte significativa del corpus deriva dal progetto PARTHENOS (Standardization Survival Kit[13]), dove i workflow erano modellati come oggetti TEI, in una struttura a passaggi (Romary et al., 2016).

3.3 Esempio selezionato da SSHOC Marketplace: Automatic Text Recognition Roadmap

La roadmap per ATR (Automatic Text Recognition)[14] disponibile su SSH Marketplace articola nove fasi dalla definizione degli obiettivi alla produzione dell'output e al riuso. In ottica di automazione, risultano

---

[7] Carbon Applications Archives, formato privilegiato per esportare e salvare i workflow con WSO2 Micro Integrator: https://apim.docs.wso2.com/en/4.2.0/integrate/develop/integration-development-kickstart/ (cons. 10/01/2026).

[8] Create, Read, Update, Delete, ossia le quattro operazioni fondamentali su una risorsa (creazione, lettura, modifica, cancellazione).

[9] Documentazione ufficiale sui mediators: https://mi.docs.wso2.com/en/4.4.0/reference/mediators/about-mediators/ (cons. 10/01/2026).

[10] Piattaforma del progetto H2IOSC pensata come punto d'ingresso unico per scoprire e accedere a servizi, strumenti, dataset, materiali formativi e workflow delle infrastrutture federate (CLARIN-IT, DARIAH-IT, E-RIHS.it, OPERAS-IT), con interoperabilità verso l'ecosistema EOSC. In prospettiva, il Marketplace prevede, oltre al catalogo descrittivo, un livello di orchestrazione di servizi per esecuzioni remote; https://www.h2iosc.cnr.it (cons. 10/01/2026).

[11] SSO sta per "Single Sign-On": autenticazione unica che consente di accedere a più servizi con un solo login. IdP sta per "Identity Provider": fornitore di identità che autentica l'utente e rilascia le credenziali (token) per i servizi.

[12] Enterprise Service Bus: script di orchestrazione che instrada chiamate e dati tra componenti applicativi.

[13] Lo Standardization Survival Kit (SSK) è un ambiente online concepito all'interno del progetto europeo PARTHENOS per fornire a ricercatori in discipline umanistiche e del patrimonio digitale informazioni, documentazione e risorse sugli standard applicabili alle attività di ricerca digitale; https://hal-lara.archives-ouvertes.fr/hal-01513531 (cons. 10/01/2026).

[14] Consultabile qui: https://marketplace.sshopencloud.eu/workflow/o8ilLU?step=N6Lt4I (cons. 10/01/2026).

tipicamente computabili: acquisizione immagini, pre-processing, layout analysis, riconoscimento, valutazione della qualità e produzione/normalizzazione dell'output; i primi passi restano in gran parte organizzativi.

3.4 Sfida principale: dal testo descrittivo alla sequenza di operazioni computabili e servizi invocabili

La sfida è tradurre la narrazione in una catena invocabile via API con contratti di dato espliciti. Fra i passaggi cruciali per l'implementazione di questo processo, si segnalano: (i) distinguere step computabili vs concettuali, (ii) esplicitare formati/MIME, schemi e vincoli, (iii) mappare ogni azione su servizi con manifest e specifiche API verificabili.
AEON può fungere da ponte perché vincola la traduzione ai manifest e la sottopone a controlli di compatibilità prima dell'esecuzione.

4. METODOLOGIA: DAL WORKFLOW DESCRITTIVO ALLA PIPELINE AEON MEDIATA DA UN CHATBOT

La metodologia proposta mira a trasformare workflow descrittivi (narrativi e contestualizzati) in workflow applicativi eseguibili in AEON, mantenendo AEON come "catena di garanzia" per interoperabilità, tracciabilità e governo del ciclo di vita. Assumiamo, in modo non rigoroso ma operativo, che un chatbot LLM con funzionalità di agente sia un sistema conversazionale basato su un Large Language Model che non si limita a generare risposte linguistiche, ma è in grado di percepire obiettivi, pianificare azioni, interagire con strumenti esterni (API, database, web, software) e operare nel mondo digitale come soggetto operativo, non solo discorsivo.

4.1 Lettura analitica del workflow descrittivo.

L'agente segmenta il testo in step, ricostruendone la struttura argomentativa, e separa ciò che è prescrittivo e organizzativo da ciò che è computabile. Per ogni step computabile estrae: (a) azioni (ingest, transform, analyse, export), (b) entità computabili (strumenti, formati, standard, parametri: es. OCR/HTR; TIFF/PNG; TEI/XML; JSON), (c) vincoli (input/output attesi, prerequisiti, dipendenze). Questo processo di disambiguazione semantica riduce sinonimie e ambiguità (es. tokenization vs lemmatization; OCR vs HTR) e produce mapping espliciti verso gli I/O attesi dai servizi.

4.2 Allineamento ai servizi AEON tramite manifest

L'efficacia del chatbot dipende dall'adesione rigorosa alle specifiche dichiarative del catalogo AEON: la bozza prodotta[15] viene ancorata ai manifest dei servizi, trattandone campi e vincoli, come i tipi e i formati di input/output, gli schemi, i parametri, e i requisiti d'accesso, che rappresentano le condizioni di progetto non negoziabili. L'agente non si sostituisce alla catena di garanzia: le proposte sono consegnate ai controlli formali del Connector e alla pipeline di validazione, generazione e deploy che converte la rappresentazione in risorse pubblicabili (manifest e OpenAPI) ed eseguibili sul motore d'integrazione. L'aderenza ai manifest consente inoltre di rispettare i vincoli semantici espressi da vocabolari o codifiche controllate, mantenendo tracciabilità delle scelte e verificabilità ex post. Così il chatbot svolge un ruolo assistivo e trasparente, mentre AEON assicura coerenza, riproducibilità e governo del ciclo di vita dei workflow.

4.3 Generazione della proto-pipeline e compilazione

L'output è una pipeline in doppia rappresentazione: Pipeline.json come grafo logico (nodi, archi, parametri, riferimenti a schemi) per UI e revisione umana, e XML interpretabile da WSO2 Micro Integrator (eventualmente impacchettabile in CAR) che materializza la catena di mediatori necessaria a invocare servizi REST, trasformare payload e costruire l'output.

4.4 Esempio concreto

Si prenda come riferimento il già citato workflow ATR Roadmap del SSHOC Marketplace, che inquadra un percorso metodologico dall'acquisizione delle immagini alla scelta del formato finale, includendo pre-processing, layout analysis e riconoscimento OCR/HTR. Su questa base, l'agente proposto estrae una catena minima, mirata alla computabilità: ingestion → HTR → post-processing → TEI export. Dalla scheda

[15] Un file in formato XML, di fatto la parte essenziale per l'orchestrazione del contenuto dei file CAR.

narrativa il chatbot individua i segmenti operativi e li caratterizza: *ingestion* (recupero/ingresso immagini), *HTR* (riconoscimento testuale), *post-processing* (pulizia/normalizzazione testuale), *TEI export* (serializzazione strutturata). I passaggi concettuali restano fuori dalla pipeline eseguibile, in coerenza con la distinzione computabile/organizzativo. La sequenza viene allineata al catalogo dei servizi di AEON generando quattro nodi: ingestionAPI, HTR-API, NLP-service (post-processing) e TEI-normalizer. Per ciascun nodo si definiscono e si vincolano input, output, MIME type e schema ai rispettivi JSON manifest. Le relazioni tra i nodi sono rese esplicite come contratti di scambio dei dati (dall'immagine al testo, dal testo "grezzo" al testo normalizzato, quindi a TEI/XML). In questa fase AEON conserva la rappresentazione di riferimento nel file Pipeline.json. Il backend di AEON compila la proto-pipeline in XML per WSO2 Micro Integrator, inserendo i mediatori necessari a costruire le richieste, raccogliere le risposte e applicare le trasformazioni intermedie. Qui, un esempio, fpinnahub (2026). La logica utilizza log, payloadFactory, call, enrich e respond, cioè i mediatori standard per orchestrare chiamate REST, trasformare il payload e costruire l'output finale; la trasformazione dei dati e l'instradamento sono così resi deterministici e tracciabili.

4.5 Validazione, deploy controllato, architettura e test preliminari

La pipeline proposta è sottoposta ai controlli del Connector (compatibilità I/O; estendibile a conformità di schema e coerenza semantica) prima del passaggio allo stato Ready. Gli esiti (mismatch, errori, suggerimenti) costituiscono un prezioso feedback capace di attivare correzioni automatiche o intervento umano. Solo pipeline validate possono avanzare nell'ambiente di esecuzione della piattaforma. Riassumendo, l'agente è dunque organizzato come catena estrazione → retrieval → mapping → generazione → validazione (vd. Fig.1), dove estrazione comprende l'identificazione e la classificazione degli step, retrieval è la raccolta dal catalogo dei servizi potenzialmente utilizzabili, mapping è la scelta del servizio migliore, generazione è relativa alla proto-pipeline e validazione è il passaggio finale, a valle dei controlli del Connector.

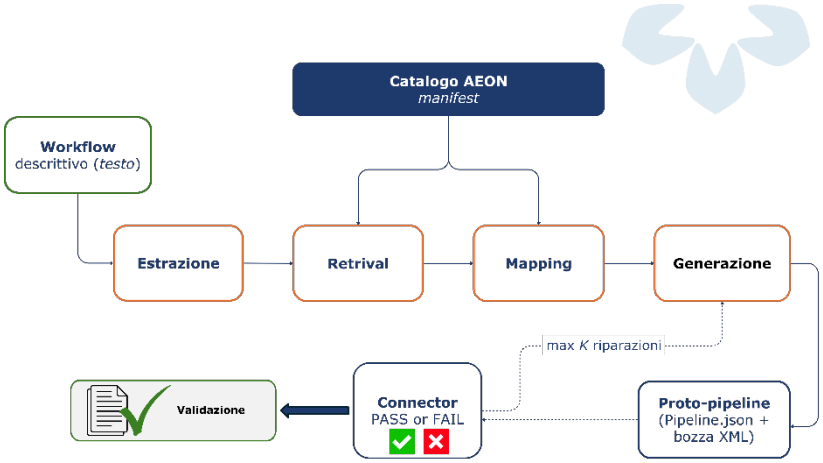


Figura 1. Catena operativa dell'agente in AEON

Nella fase di test preliminare, alla luce di Le Goues et al. (2012) e Bi et al. (2024), la stima dell'efficacia dell'azione agentica segue un ciclo generate-and-validate: l'agente genera una proto-pipeline vincolata ai manifest e la sottopone al Connector come *oracle*; gli errori di compatibilità guidano iterazioni di riparazione fino a un budget massimo (*K*), analogamente ai loop di validazione usati nell'APR (Automated Program Repair). Più specificatamente, come gold standard si selezionano *N* workflow descrittivi eterogenei (per domini e per attività). Per ciascuno, manualmente ne viene identificata: (a) la lista degli step; (b) per ogni step, la classificazione (computabile o concettuale); (c) per ogni step computabile, la relativa azione, gli I/O attesi e il servizio target su AEON (o "nessun match"). Questi *N* workflow vengono poi processati dall'agente, e le risultanti *N* proto-pipeline possono essere valutate con le seguenti metriche:

- **Estrazione step computabili**: F1-Score (oppure Precision, o Recall)[16].
- **Mapping verso AEON**: *coverage* (% step computabili correttamente identificati) e *top-1 accuracy* (frazione dei servizi correttamente identificati).
- **Successo end-to-end**: % proto-pipeline che supera i controlli del Connector (entro *K* riparazioni).
- **Riduzione del lavoro umano**: tempo (min) e numero di modifiche in UI per ottenere una pipeline Ready, confrontando (A) costruzione interamente manuale vs (B) correzione della bozza generata.

Sulla base di queste metriche ne segue la relativa reportistica, che potrà indirizzare future correzioni della piattaforma, l'estensione del catalogo dei servizi e tipi di controlli presenti.

5. RISULTATI ATTESI E IMPLICAZIONI PER INFRASTRUTTURE SSH

L'adozione di manifest strutturati e di controlli sistematici (compatibilità I/O, sintassi, semantica) abilita pattern ripetibili di composizione, rendendo pipeline e output più versionabili, tracciabili e riusabili (nel senso che reusable prende nell'acronimo FAIR). Il Flow Designer, insieme alla validazione del Connector, riduce la barriera tecnica: ricercatori nelle DH e SSH possono configurare, testare e condividere workflow eseguibili senza scrivere XML o codice, disponendo di log consultabili in interfaccia. In prospettiva, un *workflow translator agent* nativo può tradurre workflow descrittivi in proto-pipeline vincolate dai manifest (Pipeline.json + bozza XML), aumentando precisione e copertura tramite addestramento su corpora come lo SSHOC Marketplace.

6. DISCUSSIONE E PANORAMA ATTUALE

Lo SSHOC Marketplace offre workflow descrittivi ricchi di contesto, che risultano utili alla trasparenza metodologica e alla discovery, ma non sono eseguibili. WorkflowHub[17], al contrario, è un registro FAIR di workflow computazionali (CWL, Nextflow, Snakemake, Galaxy ecc.)[18], con versionamento, DOI e metadati: è un repository ma non un orchestratore. VERA[19] supporta co-design e project management in SSH, ma non valida I/O, né automatizza esecuzioni. Nell'H2IOSC Marketplace l'orchestrazione di API è ancora un obiettivo la cui realizzazione non include la riduzione della soglia tecnica. Nel complesso, la criticità maggiore resta la mediazione semantica: domini e pratiche eterogenei richiedono contratti di dato espliciti. Pesano divergenze ontologiche, granularità irregolare degli step e lacune (schemi, MIME, pre/post-condizioni). In questo quadro AEON funge da ponte tra repertori descrittivi e registri computazionali, offrendo controlli formali ed esecuzione riproducibile.

7. CONCLUSIONI E SVILUPPI FUTURI

Abbiamo presentato AEON come ambiente web low-code per catalogare servizi e comporre, validare ed eseguire workflow applicativi nelle SSH e DH, con controlli di compatibilità e deploy su WSO2 Micro Integrator. Abbiamo quindi discusso il divario tra "workflow descrittivi" (SSHOC Marketplace) e pipeline eseguibili, proponendo un approccio in cui un LLM-agent traduce descrizioni narrative in proto-pipeline vincolate dai manifest dei servizi. Si può considerare anche il caso in cui un utente esperto rediga direttamente la descrizione del proprio workflow e la fornisca come input: il chatbot ne restituisce la controparte applicativa in forma di bozza verificabile. Un'evoluzione naturale è l'integrazione diretta del traduttore in AEON come modulo nativo (workflow translator agent), con interfacce leggere per l'editing della proto-pipeline, la resa grafica sulla board e il passaggio controllato di stato (draft → ready → published). In prospettiva, il *workflow translator agent* può uscire dai confini delle DH e lavorare su qualsiasi catalogo di servizi o applicazioni (fino a ecosistemi generalisti come gli app store). In pratica, partendo dalle schede e dalle descrizioni di app e servizi, l'agente può costruire flussi pronti all'uso che combinano funzionalità diverse in scenari reali, riducendo drasticamente il tempo tra idea e adozione.

[16] Consultare il seguente link per le relative definizioni: https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall?hl=it (cons. 10/04/2026).
[17] Raggiungibile qui: https://workflowhub.eu (cons. 10/01/2026).
[18] Si prenda, come spunto, la comparazione: https://vatlab.github.io/blog/post/comparison (cons. 10/01/2026).
[19] Link: https://vera.operas-eu.org/en/home (cons. 10/01/2026).

RINGRAZIAMENTI

Gli autori ringraziano Dedagroup per il contributo tecnico allo sviluppo di AEON, il team DARIAH-IT per il lavoro collaborativo e l'OVI-CNR per il supporto scientifico. Durante la stesura è stato utilizzato ChatGPT 5.2 per controllo ortografico, grammaticale e miglioramento stilistico; gli autori hanno poi revisionato il testo e se ne assumono piena responsabilità.

BIBLIOGRAFIA

Barbot, L., Dolinar, M., Gray, E. J., Grisot, C., Illmayer, K., Kurzmeier, M., & McGillivray, B. (2024). Contextualizing research tools & services through workflows in the SSH Open Marketplace. *Journal of Open Humanities Data, 10*, Article 22. https://doi.org/10.5334/johd.192

Bi, Z., Wan, Y., Wang, Z., Zhang, H., Guan, B., Lu, F., Zhang, Z., Sui, Y., Jin, H., & Shi, X. (2024). *Iterative Refinement of Project-Level Code Context for Precise Code Generation with Compiler Feedback*. In Findings of the Association for Computational Linguistics: ACL 2024 (pp. 2336–2353). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.138

Degl'Innocenti, E., & Pinna, F. (2025). AEON: an innovative platform for workflow orchestration and service integration in Digital Humanities. 2025 IEEE International Conference on Cyber Humanities (IEEE-CH), 1-7.

Digital Research Infrastructure for the Arts and Humanities. (2024). DARIAH-EU Annual Report 2023. Zenodo. https://doi.org/10.5281/zenodo.14007767

fpinnahub. (2026, February 2). *pipeline.xml* (Version 2a68061) [XML file]. GitHub. https://github.com/fpinnahub/DataForPapers/blob/2a68061/pipeline.xml

Le Goues, C., Nguyen, T. V., Forrest, S., & Weimer, W. (2012). *GenProg: A Generic Method for Automatic Software Repair*. IEEE Transactions on Software Engineering, 38 (1), 54–72. https://doi.org/10.1109/TSE.2011.104

Romary, L., Degl'Innocenti, E., Illmayer, K., Joffres, A., Kraikamp, E., Larrousse, N., Ogrodniczuk, M., Puren, M., Riondet, C., & Seillier, D. (2016, October 31). *Standardization survival kit* (Deliverable D4.1). PARTHENOS. https://www.parthenos-project.eu/Download/Deliverables/D4.1_standardization_survival_kit.pdf

Rubies, D. (cons. 10/01/2026). *Come creare un chatbot AI in 7 passaggi [2026]* [How to build an AI chatbot in 7 steps (2026)]. Botpress. https://botpress.com/it/blog/how-to-build-your-own-ai-chatbot

Sichera, P., Marras, C., & Pasini, E. (2024). Orchestrazione API per workflow applicativi nell'ambito delle digital humanities [API orchestration for application workflows in the digital humanities] [Report]. Zenodo. https://doi.org/10.5281/zenodo.14187534

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Bonino da Silva Santos, L., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data, 3*, Article 160018. https://doi.org/10.1038/sdata.2016.18

Gamification per la valorizzazione dei beni musicali. Un'ipotesi di raccolta di dati catalografici in *crowdsourcing*

MARCELLO RANIERI
ALMA MATER STUDIORUM - UNIVERSITÀ DI BOLOGNA, ITALIA – MARCELLO.RANIERI2@UNIBO.IT

ABSTRACT (ITALIANO)

La *gamification* è l'inserimento di dinamiche di gioco nello svolgimento di compiti che normalmente non le prevedono. Uno degli effetti positivi può consistere nell'attrarre pubblici nuovi verso la fruizione di beni culturali. Si ipotizza un esempio di applicazione della *gamification* anche per il coinvolgimento del pubblico nella catalogazione, raccogliendo dati in *crowdsourcing*. Per i brani estratti da opere di arti performative, ad esempio incisi su disco, le attuali norme di descrizione catalografica non prevedono l'indicazione della posizione del brano all'interno dell'opera completa. Viene dunque trascurato lo svolgimento nel tempo che è connaturato alla fruizione delle arti performative. Per la valorizzazione di questo aspetto, invece, e per una fruizione più rispondente all'ordine originale, risulterebbe preferibile inserire nel catalogo dati descrittivi della posizione dei brani estratti all'interno dell'opera. Tale operazione richiederebbe l'intervento di catalogatori professionali formati appositamente, con impatto economico ed effetti sull'organizzazione del lavoro di catalogazione. La *gamification* del medesimo compito mediante piattaforme di *crowdsourging* non richiederebbe invece ai giocatori competenze iniziali e il risultato potrebbe raggiungere la medesima precisione attraverso correzioni successive durante le sessioni di gioco.
Parole chiave: dischi 78 giri; valorizzazione; catalogazione; *crowdsourcing*; *applied games*

ABSTRACT (ENGLISH)

Gamification for the valorization of musical cultural heritage. Videogames as an instrument for crowdsourced collection of cataloging data.
Gamification is the insertion of typical dynamics of playing in tasks that normally don't imply them. One of the positive side effects can be the attraction of new audiences towards the fruition of cultural heritage. An example of application of gamification is hypothesized to involve people in cataloging, as well, collecting data in crowdsourcing. In cataloging excerpts from performative arts works, for example grabbed on discs, current rules don't include the indication of the position occupied by the piece of music in the whole work. The duration in time that is natural for performative art is then put apart. For the valorization of this characteristic, instead, as for a fruition more corresponding to the original order, it would be preferable to insert in the catalog descriptive data about the position of excerpts inside the work. Such an operation would require the involvement of specifically trained professional catalogers, with an impact on the budget and effects on the organization of cataloging work. The gamification of the same task through crowdsourcing platforms would not require starting competences and the result could amount to the same precision via successive approximations in playing sessions.
Keywords: 78 rpm records; valorization; cataloging; crowdsourcing; applied games

1. INTRODUZIONE

Le arti performative sono accomunate da una fruizione legata a eventi, che devono svolgersi in un flusso di tempo. Dunque, esse risultano necessariamente soggette alla estemporaneità delle esecuzioni o alle testimonianze che ne registrano la memoria. Per quanto riguarda, ad esempio, la *performance* musicale, «la sua documentazione richiede la "tracciabilità" di un dato di per sé immateriale come il suono…», rendendone le testimonianze materiali «… rappresentazioni di una temporalità diveniente» (Boni, 2013: 64-65). Per l'opera lirica, per la quale si è stabilizzato un repertorio fin dall'Ottocento, furono fondamentali la fase pionieristica dell'incisione del suono su cilindro fonografico a partire dalla fine di quel secolo e la fase dell'uso del disco a 78 giri, «…il primo supporto a larga diffusione commerciale, prodotto dai primi anni del Novecento fino agli anni Cinquanta» (Pompilio, 2007: 79). Le incisioni giunte a noi da quel periodo, che costituiscono un patrimonio documentale preziosissimo, nell'attestare l'importanza di

questo linguaggio artistico, ne tramandano l'essenza. «Se ci riferiamo alla musica come "arte dei suoni", ci è facile valutare l'importanza della registrazione nel fissare, per quanto possibile, la manifestazione artistica e poterne quindi usufruire successivamente in contesti e con modalità differenti» (Dandolo, 2007: 47). Le voci degli esecutori più importanti di quelle epoche, decisive nell'avviare una tradizione nella storia dell'interpretazione di ciascun'opera, possono essere ancora oggi ascoltate grazie al fatto che vennero registrate su supporti materiali che ne permettono la riproduzione con il livello di fedeltà raggiunto dalla tecnologia dell'epoca. Sia per l'apprendimento del canto lirico che per studi musicologici e critici, oltre che per ragioni storiche, è essenziale conservare quelle insostituibili testimonianze, catalogarle, valorizzarle.

Una caratteristica tipica delle registrazioni su supporto della prima metà del secolo scorso è la durata limitata, che consente quindi di contenere solo una porzione dell'opera complessiva dalla quale il contenuto è stato estratto: «di solito si trattava di brani piuttosto brevi, poiché ogni facciata durava poco più di quattro minuti» (Sassoon, 2008: 905-906). Questo limite fisico determinava l'obbligo di ridurre l'incisione a estratti selezionati tra i punti salienti dell'opera complessiva. Nell'attribuzione di titoli ai brani estratti e registrati, vi era di solito una pressoché totale arbitrarietà, che perlopiù seguiva logiche commerciali stabilite dall'editore o dal distributore senza particolari connessioni con la realtà musicale dell'opera dalla quale il contenuto veniva estrapolato.

La parzialità delle riproduzioni e l'arbitrarietà dei titoli sono aspetti peculiari di questo genere di materiale, che devono essere gestiti adeguatamente nelle fasi di catalogazione per decine di migliaia di risorse. Le *REgole Italiane di CATalogazione* (REICAT), emanate dall'Istituto Centrale per il Catalogo Unico e la documentazione bibliografica (ICCU), e la norma specifica *Titolo dell'opera musicale*, non prevedono che venga indicata la posizione relativa all'opera completa del singolo brano contenuto in una risorsa di suono registrato: «Non si fa riferimento alla posizione dell'estratto all'interno della composizione».[1] Gli innegabili vantaggi dell'introduzione di un indicatore di posizione nelle descrizioni bibliografiche di tali risorse non risultano al momento prese in considerazione. Ma, qualora questo dovesse avvenire, la corretta collocazione dei brani renderà necessaria la predisposizione di appositi *authority file*, consistenti in indici ordinati controllati dei brani delle più note opere.

La creazione di indici per questa funzione richiederà personale specializzato, appositamente formato, che saprà orientarsi nel repertorio melodrammatico e conoscerà l'opera per la quale di volta in volta svolgerà tale compito. La formazione di queste competenze nei catalogatori rappresenterà una sfida, sia dal punto di vista economico che del reperimento sul mercato del lavoro di operatori che abbiano acquisito le competenze necessarie.

L'inserimento di questo genere di informazione potrebbe però avvenire anche con il contributo di volontari che, senza una specifica formazione di partenza, potrebbero integrare i cataloghi esistenti con il genere di informazione richiesta. Come è stato constatato (San Martin & Weiss, 2017: 175), «...crowdsourced contributions to archives and memory institutions (libraries, museums, etc.) is a growing phenomenon.

Per rendere attrattiva per la platea dei potenziali volontari una simile attività, si propone qui una strategia di *gamification* del compito, attraverso un *applied* (o *serious*) *game*, che non coinvolgerebbe catalogatori di professione, ma cultori della musica operistica o semplici appassionati di videogames, attraverso dinamiche di gioco.

Per chiarire meglio il concetto di *gamification*, esso può essere definito come in letteratura (Papagiannakis & Lee, 2024: 808): «...the use of game design elements in nongame contexts and activities». Con definizione meno sintetica (Deterding et al., 2011: 2): «…an umbrella term for the use of video game elements (rather than full-fledged games) to improve user experience and user engagement in non-game services and applications».

I vantaggi del *game-based crowdsourcing* sono stati teorizzati (Brito *et al.*, 2015: 445), «One strategy to increase user participation is the use of game design techniques, since games have successful strategies to grant enjoyable experience» e l'applicazione in attività complesse di raccolta dati è già stata sperimentata (Yang *et al.*, 2020: 1312), «We devise a two-pronged crowdsourcing scheme ... To effectively utilize the two types of tasks, we introduce a *game-based* crowdsourcing approach CrowdGame».

[1] <https://norme.iccu.sbn.it/index.php?title=Norme_comuni/Authority_file/Titoli_materiale_musicale/Registrazione_di_authority/Dati_specifici/Estratto>

2. CONCEPT DI GIOCO

Presupposto essenziale della realizzazione del progetto sarà la connessione della piattaforma di gioco con un *repository* degli indici posizionali, che costituirà l'*authority file* a disposizione dell'ente responsabile della catalogazione. Se, per mera ipotesi, l'ente responsabile della catalogazione fosse l'Istituto Centrale per i Beni Sonori e Audiovisivi (ICBSA), il *repository* conterrebbe dati controllati sulla posizione, un tipo di entità attualmente non richiesta dagli standard di descrizione bibliografica, che andrebbero a integrare le registrazioni di responsabilità nel catalogo del polo ICBSA visualizzabili attraverso l'OPAC del Servizio Bibliotecario Nazionale (SBN).[2] Il *repository* potrebbe essere inizialmente popolato con voci controllate di posizione inserite da catalogatori relative a un numero ridotto di risorse.
Obiettivo finale del progetto sarà il popolamento globale del *repository* con un indice, che descrive la posizione occupata, nell'opera dalla quale è estratto, da ciascuno dei brani contenuti nelle risorse (dischi) presenti nel catalogo.
L'inserimento dei dati di posizione dei brani, se ancora non presenti, o il controllo e l'eventuale correzione dei dati già inseriti in precedenza, sarà il risultato delle sessioni di gioco.
Ogni giocatore può inserire nuovi dati, che saranno soggetti al giudizio della comunità dei giocatori, o confrontarsi con i dati già inseriti da ciascuno dei giocatori precedenti, anche di altre sessioni di gioco, e può modificarli, migliorando il proprio punteggio o esponendosi a una correzione da parte di altri giocatori, che determinerà invece un peggioramento nel suo punteggio.
I dati da inserire (o da correggere) sono quelli relativi alla posizione da attribuire ai dischi di musica operistica presenti in un dato catalogo.
Non si tratta di un gioco *multiplayer* in senso stretto, ma in maniera asincrona, in un certo senso: ogni giocatore è inserito in una graduatoria e il suo posto in classifica è determinato dalla somma dei punteggi, positivi o negativi, raggiunti nelle successive sessioni di gioco a cui partecipa nel tempo. In pratica, non si può essere vincitori definitivamente, ma il proprio posto in graduatoria dipende da quello degli altri giocatori, che possono mettere in discussione i dati da noi inseriti (che ci hanno permesso di aumentare il nostro punteggio) e apportare modifiche imponendoci cali di punteggio, ma rischiando di perdere punti a loro volta se la comunità dei giocatori non approva le loro modifiche ai nostri dati.

3. SCOPO

L'obiettivo generale dell'operazione è registrare in un *repository*, attraverso il gioco, i dati relativi alla posizione corretta dei brani contenuti nei dischi a 78 giri di un catalogo rispetto al melodramma da cui ciascun brano è stato estratto. L'indicazione della corretta posizione del brano contenuto nel disco, nel contesto dell'opera completa, porterà alla costruzione di un indice ordinato di risorse relative all'opera (cioè dei brani contenuti nei dischi).
L'obiettivo che muove il singolo giocatore (già appassionato di opera o semplice curioso) è il raggiungimento di una reputazione sociale guadagnata sul campo grazie all'esattezza dei dati inseriti o delle modifiche apportate a dati inseriti dagli altri giocatori, nei casi in cui esse si sono rivelate necessarie. L'inserimento di dati corretti corrisponderà a un aumento del proprio punteggio e una ascesa del proprio *nickname* (e del proprio *avatar*) nella classifica generale della piattaforma di gioco.
Si ipotizza che la dinamica ludica permetta di superare la curva di apprendimento richiesta dal compito, secondo modalità note in letteratura (Morschheuser et al., 2017: 1): «…organizations seek to make the crowdsourced work activity more like playing a game in order to provide other motives for working than just monetary compensation». Viene anche detto in altri termini (Prandi et al., 2024: 822): «…Several crowdsourcing systems apply different gamification techniques with different aims and contexts».

4. FASI E MODALITÀ

Entrando nel gioco si seleziona un certo compositore e si sceglie una delle sue opere. Poniamo per ipotesi che si scelga *La Bohème*, una ben nota opera di G. Puccini.
Da quel momento inizia la fase di gioco vera e propria. Durante la sessione si scorre un elenco casuale di dischi tratti dall'opera selezionata, dei quali si visualizzano una per volta le etichette delle due facce (oppure un'etichetta sola se il disco risulta inciso su una sola faccia).

[2] <https://opac.sbn.it/>

In Fig. 1 un *mock-up* del videogioco. Sulla sinistra dello schermo compaiono le etichette dei dischi presenti in catalogo (il disco può essere ancora da catalogare, o solo da approvare), mentre sulla parte destra si può navigare all'interno del libretto dell'opera prescelta.
Quando presenti sull'etichetta, le indicazioni di posizione più generali (ad es. "atto I", "scena III", ecc.) permettono fin dall'inizio di ridurre le difficoltà nel localizzare meglio il brano.
Una prima strategia di gioco può essere quella di cercare direttamente in modalità testuale il titolo del brano sul libretto. Se il titolo sull'etichetta corrisponde all'incipit di un brano dell'opera, sarà semplice rintracciarlo. Ad es., se il titolo del brano riportato sull'etichetta del disco fosse *Sì, mi chiamano Mimì* (vd. Fig. 1), una ricerca testuale nel libretto aiuterebbe a rintracciare l'incipit della celebre aria e permettendo di posizionare il brano.
Un'altra ricerca preliminare che è possibile compiere è quella dei dischi contenenti un brano con quel medesimo titolo, che se già localizzato da un altro giocatore permette di accelerare il compito, "catturando" l'indice fornito, se considerato plausibile.
In ogni caso il giocatore, se lo ritiene vantaggioso o anche solo per il puro piacere dell'ascolto, ha la possibilità immediata di fruire un breve ascolto iniziale del brano, grazie alla connessione della piattaforma di gioco con la versione digitalizzata presente sull'OPAC SBN dell'ICBSA,[3] disponibile per i primi trenta secondi in deroga alla L. 633/1941 sul diritto d'autore, grazie a uno speciale accordo di concessione stipulato con la Società Italiana Autori ed Editori (SIAE).[4] Grazie a questa possibilità, l'identificazione del brano risulta spesso semplificata, non solo ai melomani che riconosceranno immediatamente il passaggio dalle prime note anche senza il bisogno di ascoltare brani di canto, ma a chiunque se l'ascolto comprende alcuni versi iniziali, che di nuovo saranno da ricercare nel libretto.
Nel caso in cui nessuno di questi metodi abbia permesso l'identificazione del brano e il suo conseguente posizionamento, il giocatore sarà libero di avviare una sua ricerca su ogni catalogo di cui riesca a venire a conoscenza, per reperire quante più informazioni possibili relative a quel disco, a partire dai dati presenti nell'etichetta come farebbe qualsiasi catalogatore. Ogni caso sarà diverso dagli altri; in quello in esempio sarebbe stato d'aiuto il catalogo delle risorse digitalizzate della Bibliothèque Nationale de France.[5]
Una volta individuata una localizzazione per il brano, il giocatore trascina l'icona dell'etichetta sulla posizione del libretto che ritiene corretta (vd. Fig. 1) e a quel punto il compito è stato svolto. L'indice di posizione corrispondente verrà automaticamente associato a quel brano e trasferito nel *repository*. Il giocatore potrà così procedere con l'altra delle due etichette di quel disco, se c'è, o con il disco successivo dell'elenco. La comunità dei giocatori sarà intanto libera da quel momento di valutare la correttezza dell'indice di posizione assegnato dal giocatore.

[3] <https://cloud.sbn.it/opac/DDS/ricercaSemplice>
[4] <http://opac2.icbsa.it/vufind/Record/IT-DDS0000041137000200>
[5] <https://gallica.bnf.fr/ark:/12148/bpt6k128629t.r>

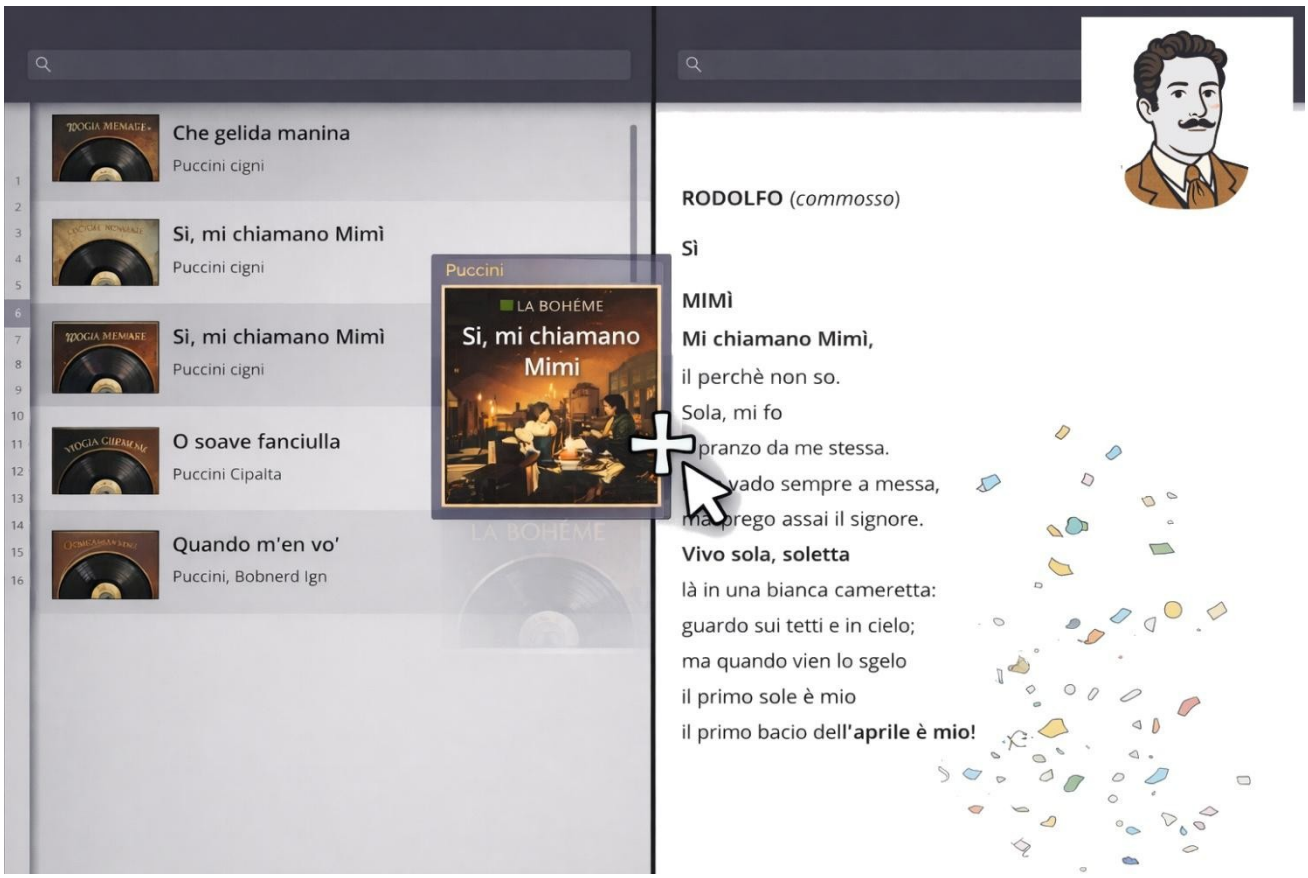


Figura 1

5. CONCLUSIONI

Al presente, la ricerca qui esposta è limitata allo sviluppo di un *proof of concept*. La normativa catalografica vigente non richiede l'indicazione della posizione del brano all'interno dell'opera complessiva, né per le risorse pertinenti alle arti performative, né tantomeno per altri generi di documento. Quindi, non si è ritenuta utile un'implementazione del gioco progettato che non potrebbe essere connessa a un *repository* ufficiale.
Qualora la normativa venisse aggiornata (sarebbe estremamente utile ai fini di una presentazione ordinata delle informazioni sulle risorse) la creazione rapida di indici di posizione con un sufficiente livello di precisione potrebbe avvenire tramite una piattaforma di *crowdsourcing* basata su un modello di gioco come quello qui presentato.
Il gioco può essere strutturato con caratteristiche differenti a seconda delle esigenze specifiche e della creatività di progettisti e sviluppatori. Si ritiene utile presentare di seguito tre possibili esempi diegetici attraverso lineamenti progettuali narrativi da adattare alle reali esigenze.

5.1 Diegesi 1: La principessa Kanouke

Kanouke ha ereditato un regno e con esso un patrimonio sonoro. Il suo compito come futura regina è promuovere la cultura musicale dei suoi sudditi e delle sue suddite, avvicinandoli al canto e alla sua bellezza. Per farlo, le è essenziale esercitarsi ascoltando le voci del passato, incise su misteriose fonti circolari, i dischi solcati. Purtroppo, ha molte difficoltà nell'ordinare i brani che su di essi sono stati registrati e ricostruire le opere dalle quali sono stati originariamente tratti. Se vuole confrontare tra loro le esecuzioni dello stesso brano da parte di diversi artisti, in modo da perfezionarsi e scegliere quella più consona alle sue abilità, deve attendere di ascoltare per caso due dischi che abbiano lo stesso contenuto, dato che i titoli non ne permettono l'identificazione. Una simile confusione non fa che rallentare i suoi progressi. Il ruolo di regina del canto sembra ancora molto lontano, forse irraggiungibile. A meno che qualcuno non cerchi di aiutarla attribuendo la corretta posizione ai brani contenuti nei dischi solcati. Proprio tra i suoi sudditi e le sue suddite, le persone più entusiaste di beneficiare di una simile ricchezza culturale, inizia la corsa a posizionare correttamente quanti più brani sarà possibile.

5.2 Diegesi 2: Cooperative sonore

Si può giocare da soli oppure unendosi ad altri giocatori in una Cooperativa. Nella prima sessione che un giocatore svolge appena entrato in una Cooperativa, il punteggio per l'inserimento di un indicatore di posizione vale il triplo, mentre gli eventuali errori comportano solo una penalità dimezzata.
Dopo la prima sessione in Cooperativa, il punteggio per ogni brano posizionato all'interno di un'opera vale invece doppio (anche gli errori però!). E se l'indicatore di posizione del brano era già stato attribuito da altri giocatori ma si ritiene di poterlo correggere, giocando in una Cooperativa si fa perdere all'avversario (o alla sua Cooperativa) il doppio dei punti.

5.3 Diegesi 3: Meloman

Il mondo sta scivolando verso l'asonia. La mancanza di consapevolezza sull'importanza e l'effettiva consistenza dei beni culturali sta contribuendo al progressivo imbarbarimento della società e in definitiva delle relazioni umane. Solo i paladini della conoscenza possono tentare di opporsi. Eroi dei nostri giorni, conservano quello che resta della consapevolezza culturale e lottano per difenderlo in tutti i campi. Tra di essi, Meloman, che di giorno è uno dei maggiori esperti di musica viventi e di notte sceglie di celare la sua identità dietro una maschera e va con tutti i suoi mezzi alla ricerca dei brani contenuti nei dischi di opera lirica per restituire ad essi il giusto posizionamento in modo da favorirne al massimo la fruizione. Che cosa aspetti a metterti in gioco e unirti a lui per tentare di arginare il disorientamento dilagante? O… aspetta… oppure Meloman sei tu?

RINGRAZIAMENTI

Si ringrazia la Fondazione Scuola Nazionale Patrimonio Attività Culturali per la formazione specifica ricevuta con la frequenza dei corsi Dicolab: "Homo (video)ludens – Il videogioco che genera cultura" tenuto da F. Malgieri, "Game design e patrimonio culturale" tenuto da C. Molinari, "Come stimare e produrre un gioco applicato" tenuto da P. Polsinelli, "Realizzare *applied games* efficaci" tenuto da V. Santalucia. L'immagine è stata realizzata con l'ausilio di OpenAI ChatGPT. Per nessuna parte del testo si è invece ritenuto utile il ricorso a strumenti di intelligenza artificiale.

BIBLIOGRAFIA

Boni, M. (2013). Opera, spartito, esecuzione: i modi di esistere della musica. *Le *arti dello spettacolo e il catalogo*. Bazzocchi, V., & Bignami, P. (eds.), Carocci, 63-68. ISBN 978-88-29-00573-4

Brito, J., Vieira, V., & Duran, A. (2015). Towards a Framework for Gamification Design on Crowdsourcing Systems: The G.A.M.E. Approach. *12th International Conference on Information Technology - New Generations. (ITNG)*, IEEE, Las Vegas, NV, USA, pp. 445–450. https://doi.org/10.1109/ITNG.2015.78

Dandolo, A. (2007). La registrazione sonora una testimonianza. *Il suono riprodotto: storia, tecnica e cultura di una rivoluzione del Novecento atti del convegno annuale del «Laboratorio per la Divulgazione Musicale» Parma 10 e 11 novembre 2006*. Rigolli, A., Russo, P., & Casa della musica Parma (eds.). EDT, 47-49. ISBN 978-88-6040-176-2

Deterding S., Sicart M., Nacke L., O'Hara K., & Dixon, D. (2011). Gamification using game-design elements in nongaming contexts. *CHI '11 extended abstracts on human factors in computing systems*, Vancouver. https://doi.org/10.1145/1979742.1979575

Morschheuser, B., Hamari, J., Koivisto, J., & Maedche, A. (2017). Gamified crowdsourcing: Conceptualization, literature review, and future agenda. *International Journal of Human-Computer Studies*, *106*, 26–43. https://doi.org/10.1016/j.ijhcs.2017.04.005

Papagiannakis, G., & Lee, N. (2024). Gamification and Serious Games. *Encyclopedia of Computer Graphics and Games*, 808–811. Springer International Publishing. https://doi.org/10.1007/978-3-031-23161-2_90

Pompilio, A. (2007). La qualità del suono come 'oggetto' culturale da conservare. *Il suono riprodotto: storia, tecnica e cultura di una rivoluzione del Novecento atti del convegno annuale del «Laboratorio per la Divulgazione Musicale» Parma 10 e 11 novembre 2006*. Rigolli, A., Russo, P., & Casa della musica Parma (eds.). EDT, 77-80. ISBN 978-88-6040-176-2

Prandi, C., Salomoni, P., Mirri, S., & Lee, N. (2024). Gamification in Crowdsourcing Applications. *Encyclopedia of Computer Graphics and Games*, 819–824. Springer International Publishing. https://doi.org/10.1007/978-3-031-23161-2_46

San Martin, B., & Weiss, S. (2017). People Have the Power: Crowdsourcing Cataloging of Sound Recordings. *ARSC Journal*, *48*(2), 173–183. ISSN: 2151-4402

Yang, J., Fan, J., Wei, Z., Guollang, L., Tongyu, L., Xiaoyong, D. (2020). A game-based framework for crowdsourced data labeling. *The VLDB Journal*, 29, 1311–1336. https://doi.org/10.1007/s00778-020-00613-w

Tra riconoscimento disciplinare e prospettiva pubblica: un modello logico-applicativo per le conoscenze e le competenze nelle DH

LISA REGGIANI

CNR ISTITUTO PER IL LESSICO INTELLETTUALE EUROPEO E LA STORIA DELLE IDEE (CNR-ILIESI), ITALIA - LISA.REGGIANI@CNR.IT

ABSTRACT (ITALIANO)

Il contributo propone un modello logico-applicativo che sistematizza le conoscenze e le competenze nelle DH in uno schema complessivo, evidenziando l'opportunità di trasferirle, quale prezioso corredo contrassegnato da un solido impianto metodologico, sia alle altre comunità di ricerca che a segmenti diversificati della società, per assicurare un'autentica cittadinanza digitale allargata. Tale modello, ispirato a quello per l'informazione-documentazione di Duplessis, presenta tre livelli concentrici: un nucleo metodologico, uno strato intermedio euristico-trasversale e uno più esterno, di natura didattica e formativa. Alcuni tratti distintivi delle DH intersecano tutti e tre i livelli e garantiscono il confronto dialogico con differenti comunità scientifiche e con la società: la multi-, inter- e trans-disciplinarità, una vocazione epistemologica che si traduce nella predilezione per team interfunzionali e nella valorizzazione delle diversità e della complementarità di competenze e professionalità; la co-progettazione, corrispondente al design thinking, in cui la dimensione collaborativa permea tutte le fasi ideative e applicative; la modellizzazione, che investe e innerva l'intera galassia DH; un approccio metodologico, assimilabile al digital thinking, che incorpora, in un'integrazione progressiva, punti di vista, conoscenze, metodi, strategie e strumenti secondo una prospettiva funzionale.
Alcune realizzazioni esemplificano l'applicabilità del modello, fra cui la "Science and Technology Digital Library" e le digital library di ambito DH da essa derivate; il progetto H2IOSC e il PON *Beni e patrimonio culturale: governare la progettazione,* quest'ultimo di carattere gestionale e organizzativo.
Si sottolinea infine l'utilità del modello nel consolidare il riconoscimento disciplinare delle DH, in quanto giustifica e valorizza la loro specificità e il loro ruolo peculiare nella prospettiva pubblica.
Parole chiave: modello logico-applicativo; co-progettazione; lavoro collaborativo; digital thinking; prospettiva pubblica

ABSTRACT (ENGLISH)

Between disciplinary recognition and public perspective: a logical-applicative model for knowledge and competences in DH.
The contribution sets out a logical-applicative model that systematically organises knowledge and competences in DH within a comprehensive scheme. This highlights the opportunity to transfer them, as a valuable resource marked by a solid methodological framework, both to other research communities and to diverse segments of society, in order to ensure authentic, expanded digital citizenship. Inspired by Duplessis' information-documentation model, the proposed model comprises three concentric levels: a methodological core, an intermediate, heuristic-transversal layer, and an outer, educational and training layer. Some characteristics of DH intersect all three levels, guaranteeing dialogue with different scientific communities and society. These include multi-, inter- and trans-disciplinarity, an epistemological vocation that translates into a preference for cross-functional teams, enhancing diversity and the complementarity of skills and professional profiles; co-design, equivalent to design thinking and characterised by a collaborative dimension permeating all conceptual and application phases; modelling, permeating and underpinning the entire DH galaxy; and a methodological approach equivalent to digital thinking that incorporates points of view, knowledge, methods, strategies and tools through progressive integration, according to a functional perspective.
The model's applicability is illustrated by examples such as the "Science and Technology Digital Library" and the DH digital libraries derived from it, the H2IOSC project, and the managerial and organisational PON project *Beni e patrimonio culturale: governare la progettazione*.

Finally, the model's effectiveness in consolidating the disciplinary recognition of DH is emphasised, as it justifies and enhances their specificity and their unique role in the public perspective.
Keywords: logical-applicative model; co-design; collaborative work; digital thinking; public perspective

1. INTRODUZIONE

Il contributo enuclea e descrive alcuni elementi caratteristici delle *Digital Humanities* (DH) per proporre un modello logico-applicativo per le conoscenze e le competenze in questo settore, in grado di riunirle e sistematizzarle in uno schema visuale sinottico, definendone le interrelazioni. Il modello evidenzia l'opportunità di trasferire ad altri domini, pratiche e comunità di ricerca e a segmenti diversificati della società tali elementi caratteristici, quale prezioso corredo di attitudini, conoscenze e competenze contrassegnate da un robusto impianto metodologico; ciò può avvenire attraverso processi di *cross-fertilization* e attivando strumenti di coinvolgimento/partecipazione in spazi comunicativi intermedi più sfaccettati e profilati – come avviene nelle iniziative di Public Digital Humanities e di Citizen Humanities (Brennan, 2016; Heinisch, 2020) – o attraverso attività formative solidamente strutturate, rivolte al mondo delle professioni, dell'istruzione superiore e della scuola (Dupret et al., 2024).
La messa a punto del modello si situa nel quadro dei più recenti tentativi di tracciare la ricchezza e la vivacità del lavoro alla base delle realizzazioni di DH, che solo parzialmente confluisce in forme documentarie compiute e standardizzate, dando ad esso un risalto e un valore adeguati. Per ricondurre a strutture riconoscibili, verificabili e riutilizzabili attività e modalità di lavoro altrimenti destinate a restare se non proprio invisibili solo parzialmente visibili o comunque relegate nell'ambito del professionalismo esperienziale, gli sforzi si concentrano prevedibilmente sulla modellizzazione. In tal modo si individuano pattern, si valorizzano esperienze, si organizzano dati; si definiscono e si perfezionano metodologie e processi, si disegnano i flussi principali; si offrono, in una parola, rappresentazioni di percorsi progettuali e di ricerca soddisfacenti sia sul versante teorico-metodologico che nella pratica professionale, come accade ad es. nel King's Digital Lab per il ciclo di vita del progetti (Ciula et al., 2023b), nell'Open Science Framework[1] per l'attività di ricerca e le sue fasi canoniche e, limitatamente a singoli aspetti, nella predisposizione di *data management plan* (DMP), ormai capillarmente diffusi nelle realtà di ricerca.
Il modello qui illustrato non vuole essere alternativo ai modelli al centro del dibattito e variamente sviluppati con differenti obiettivi in ambito DH (Ciula & Marras, 2016; Ciula & Eide, 2017; Ciula et al., 2018; Ciula & Marras, 2019; Malazita et al., 2020; Ciula et al., 2023a; Pawlicka-Deger & Thomson, 2023) ma intende piuttosto affiancarli per contribuire ad una più consapevole configurazione identitaria del settore e dunque alla sua riconoscibilità scientifica e al suo posizionamento all'interno della società. Si potrebbe così delineare, auspicabilmente, una fisionomia disciplinare maggiormente caratterizzata e facilmente identificabile nella costruzione dei saperi, dei metodi e delle pratiche di ricerca.

2. IL MODELLO LOGICO-APPLICATIVO: LA STRUTTURA TRIPARTITA

Il modello logico-applicativo che si presenta è tripartito e si rifà al modello disciplinare dell'informazione-documentazione (I&D) messo a punto nell'ambito della didattica accademica da Pascal Duplessis nel 2017; il suo scopo era quello di assicurare al dominio I&D una più marcata "visibilità" scientifico-disciplinare e insieme di definire l'"identità professionale" degli specialisti (Duplessis, 2017).[2]
In forte analogia sia per la struttura che per le finalità, nel modello per le DH qui discusso "le tre matrici disciplinari" si dispongono su tre livelli concentrici, distinti ma stabilmente intercomunicanti (Fig. 1). Al centro il *nucleo metodologico* include modelli, metodi, approcci, tecniche e strumenti cardinali per la teoria e la prassi disciplinare, utili alla ricerca e alla rappresentazione della conoscenza nel settore. Lo strato intermedio corrisponde alla componente *euristica e trasversale*, nella quale trascorrono gli elementi costitutivi dello strato più interno, cui si viene però aggregando una gamma, in continua espansione, di

[1] https://osf.io/ (cons.14/04/2026).
[2] Il modello di Duplessis – https://ltc.apden.org/didactique-information/les-trois-matrices-disciplinaires-de-l-information-documentation.html (cons. 14/04/2026) – organizza in un quadro unitario e coerente le molteplici componenti del settore I&D, quale si è venuto configurando in Francia attraverso un lungo dibattito, grazie alla combinazione della prestigiosa tradizione documentaria con l'universo delle scienze e delle pratiche dell'informazione. Tale ambito scientifico-disciplinare ha trovato da tempo in Francia uno spazio riconosciuto, in particolare nella formazione scolastica e universitaria mediante la creazione di portfolio strutturati di competenze I&D, *référentiels* e curricula diversificati (Reggiani, 2008).

strategie, tecnologie, linguaggi e buone pratiche convergenti in alcune fondamentali *soft skill*. Da ultimo, sulla superficie esterna si colloca la dimensione *didattica e formativa* che è fortemente interconnessa con le altre due grazie a un'osmosi ininterrotta; in essa sono in evidenza fattori chiave – quali la cognizione e la metacognizione, la riflessione, l'argomentazione, il *problem solving* e il pensiero critico – che non costituiscono una prerogativa delle DH, ma che nel dominio vengono a connotarsi peculiarmente.

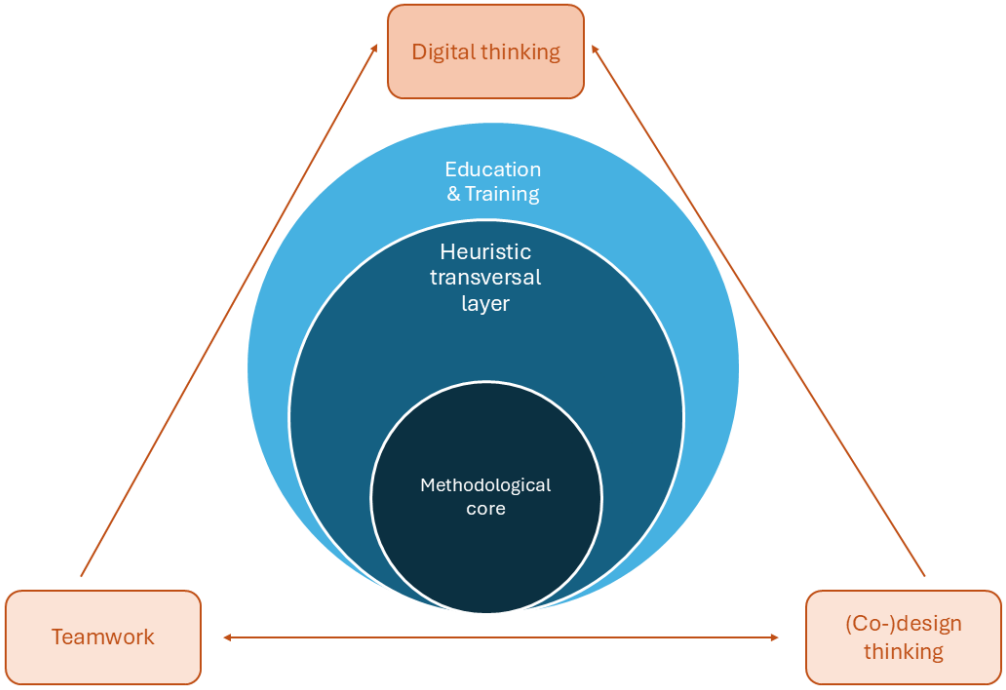


Figura 1. Modello logico-applicativo per le conoscenze e le competenze nelle DH

3. IL MODELLO LOGICO-APPLICATIVO OLTRE LE MATRICI DISCIPLINARI: LE CARATTERISTICHE TRASVERSALI

Attraversano tutti e tre i livelli alcuni tratti e tendenze trasversali, caratterizzanti l'intera costellazione disciplinare, che possono garantire una saldatura sistematica e un confronto dialogico sia con le altre comunità di ricerca che con differenti componenti della società, tramite la condivisione e la diffusione di saperi, pratiche e modalità di lavoro, l'utilizzo di strategie comunicative di coinvolgimento/partecipazione che sfruttano la crescente segmentazione delle conoscenze e dei pubblici e l'offerta strutturata di proposte didattiche e formative.

In primo piano vi è la *multi-, inter- e trans-disciplinarità* connaturata alle DH, che fa del settore uno spazio multiplo singolarmente dinamico e poroso, intrinsecamente pionieristico, incline all'apertura – a campi di indagine inediti, a nuovi ambiti e comunità di ricerca – e impegnato a superare confini, a intersecare, ibridare e innestare territori limitrofi o tangenti e traiettorie di ricerca, nel solco della visionaria *intertwingularity,*[3] concetto/termine creato da Ted Nelson nel 1974 (Nelson, 1987; Burdick et al., 2012; Castellucci, 2009). Questa vocazione epistemologica si rispecchia, in ambito lavorativo, nella metodica predilezione per *team interfunzionali*, polivalenti, duttili e reciprocamente permeabili, in cui si valorizzano le diversità e la complementarità di competenze e professionalità nel segno della partecipazione multifunzionale (Khalil et al., 2013; Lundene & Mohagheghi, 2018; Bastiaansen & Van Dun, 2025).

Un secondo elemento, che incrocia tutte e tre le matrici disciplinari, individua nella *co-progettazione* il suo perno. Nelle DH il *design thinking* trova infatti un'applicazione paradigmatica su larga scala, in quanto

[3] resa con "interconnessabilità" nella traduzione italiana (*Umanistica_Digitale*, 2014: 24) del volume *Digital_Humanities* di Burdick et al. (2012), una pietra miliare negli studi del settore. Gli autori, nel sottolineare che per le DH è cruciale l'integrazione fra la componente testuale e quella visuale, si riferiscono esplicitamente al "neologismo ludico di Theodor Holm Nelson" per rilevare più in generale che "la questione chiave riguarda l'individuazione degli strumenti più efficaci per la creazione di *interconnessazioni significative*".

offre una chiave interpretativa orientata all'innovazione, in grado di ricomporre in una sintesi unitaria processi cognitivi e strategici, organizzativi e operativi, scandendoli in fasi, spesso non lineari ma ricorsive, di efficacia sperimentata: ci si muove dalla iniziale condivisione preliminare del lavoro di ricerca alla messa a punto creativa[4] dell'idea progettuale e alla sua precoce traduzione applicativa attraverso la prototipazione, per giungere nuovamente al confronto con gli utenti, al fine di sperimentare e perfezionare il prototipo. Qui le fasi ideative si coniugano e si misurano tempestivamente con quelle applicative e la dimensione collaborativa permea e sostanzia entrambi i momenti (Balagtas, 2024).
Prossima alla progettazione è la *modellizzazione*, che investe l'intera galassia DH innervandone tutte le componenti e tutti i livelli (politico-strategico, organizzativo, scientifico e metodologico, semantico, tecnico e tecnologico) e rappresenta quasi una specificità disciplinare (Svensson, 2010; Ciula & Eide, 2017) con particolare riferimento alle trasformazioni mediali, alle metodologie di visualizzazione computazionale, all'invenzione e all'utilizzo su larga scala di metafore visive, spaziali e locali (Wouter & Uyttenhove, 2012; Elleström, 2014; Eide, 2014-2015; Averbukh, 2015; Flanders & Jannidis, 2015; Flanders & Jannidis, 2019; Elkins & Fiorentini, 2021; Ciotti, 2023; Ciula et al., 2023a): a titolo esemplificativo, dalle infrastrutture digitali alla rappresentazione della conoscenza, dall'architettura dell'informazione alla codifica dei testi, dalla descrizione, organizzazione, gestione e fruizione delle risorse ai modelli e sistemi di ricerca e navigazione, agli ambienti di lavoro virtuali, per gran parte collaborativi e interattivi.
Un'applicazione emblematica di co-design e modellizzazione nelle DH, che coniuga esemplarmente discipline umanistiche e metodi computazionali e si rivolge "tanto alla comunità scientifica quanto a un pubblico più ampio", è costituita dalle edizioni critiche digitali: grazie all'abbandono della forma tipografica tradizionale in favore di un "approccio centrato sui dati", la trasformazione mediale segna un vero e proprio "cambio di paradigma"; e la dimensione progettuale, attraverso l'approccio iterativo, si dilata a "processo continuo, aperto, esteso" destinato ad un arricchimento progressivo, cui sono chiamati a collaborare esperti di differenti domini e utenti esterni, così da favorire "una più generale partecipazione al processo della conoscenza" (Ciotti et al., 2022).
Un ulteriore tratto distintivo trasversale consiste in un *approccio metodologico* di derivazione informatica il quale ingloba via via, assimilandoli in un'*integrazione progressiva*, una molteplicità di angolazioni, metodi e aree di ricerca, singole conoscenze, strategie, strumenti etc. secondo una prospettiva funzionale che utilizza o trascura ogni singola componente modulare in un'ottica *problem solving*. Ciò consente alle DH di alimentare un patrimonio pressoché illimitato di conoscenza – dalla riflessione epistemologica e dallo statuto disciplinare alle metodologie fino alle specificità più granulari di applicazioni, linguaggi, standard e formati – cui poter attingere per impieghi ben consolidati o imprevedibili finalizzati alla risoluzione di problemi e ipotesi di ricerca; e permette loro di aggregare instancabilmente nuove professionalità e specializzazioni nei gruppi di lavoro, ridisegnando e aggiornando i team, quanto a profili, ruoli, configurazioni complessive, attraverso la metodica valorizzazione delle differenze. Quest'ultima caratteristica, corrispondente al *digital thinking*, incorpora le altre (Radin & Danylchenko-Cherniak, 2025): comprende la visione sistemica e il pensiero computazionale, la capacità di riutilizzare in modo nuovo il già noto e di esplorare in maniera non sequenziale, grazie a libere associazioni e a contaminazioni inaspettate, nuovi territori e nuove idee; accentua il peso di competenze informative sofisticate e del pensiero critico; prevede altresì, all'interno dei contesti digitali avanzati tipici dell'universo DH, il governo di una *dimensione comunitaria*, relazionale e comunicativa, che supera i codici più formali e proceduralizzati e investe la sfera *socio-emotiva,* in cui riaffiora la forte caratterizzazione cooperativa delle DH (Eshet-Alkalai, 2009).
Le DH mostrano quindi uno spettro di elementi strutturali, profondamente interrelati, che possono essere messi sistematicamente a frutto per lo sviluppo di competenze e abilità sofisticate, di importanza decisiva nella società contemporanea, sia per le comunità di ricerca che per garantire, *lato sensu*, un'autentica cittadinanza digitale: l'apertura, l'integrazione e l'innesto di punti di vista, conoscenze e comunità disciplinari; la co-progettazione e la modellizzazione; soprattutto, ben oltre il quadro pur ricco e organico del *DigComp 3.0* (Cosgrove & Cachia, 2025), il *digital thinking*, inteso come cultura e forma mentis digitale, da declinarsi in termini di visione strategica e ideazione creativa, attitudine e consapevole adesione all'innovazione e alla sperimentazione, adattabilità, condivisione e diffusione di saperi teorici e pratiche collaborative (Leonardi & Neeley, 2022).

[4] comprendente anche la tecnica del *reverse thinking*, che riesce a superare, grazie a un approccio dialettico, le insufficienze e i limiti strutturali insiti nei meccanismi del *problem solving*.

4. ALCUNE REALIZZAZIONI

Il modello proposto ha preso forma attraverso la riflessione critica riguardo a una serie di esperienze e iniziative di ambito DH e di campi limitrofi.
Un primo caso di studio degno di nota, direttamente afferente al dominio DH, lo ha fornito l'evoluzione delle infrastrutture digitali dell'Istituto per il Lessico intellettuale Europeo e Storia delle Idee del Consiglio Nazionale delle Ricerche (ILIESI-CNR) in ambienti digitali integrati sempre più interconnessi e aperti per la ricerca collaborativa, la condivisione della conoscenza su piattaforme interoperabili e l'apertura alla società; evoluzione avvenuta in concomitanza prima con i progetti Discovery - Digital Semantic Corpora for Virtual Research in Philosophy (2006-2009) e Agora - Scholarly Open Access Research in European Philosophy (2011-2013) e poi con il progetto PARTHENOS - Pooling Activities, Resources and Tools for Heritage E-research Networking, Optimization and Synergies (2015-2019).[5] Questo sforzo di modellizzazione e co-progettazione multi- e inter-disciplinare partendo da solide basi metodologiche ha sviluppato, sul piano euristico, teorie e pratiche di ricerca e soluzioni innovative per la pubblicazione, la gestione, l'accesso e il riuso di testi filosofici e scientifici, dati e informazioni nella cornice dell'Open Access e dell'Open Science; si è concentrato su questioni di natura linguistico-terminologica decisive rispetto alla condivisione degli standard e alla definizione di linee guida per la gestione delle collezioni digitali. A tale scopo ha applicato la piramide degli standard NISO adattandola alla specifica realtà dell'Istituto mediante un processo dinamico di modellizzazione teorico-pratica, enucleandone alcune componenti di interesse dimostratesi più efficaci e produttive (Liburdi et al., 2019).
Una seconda realizzazione, di natura generalista e tuttavia rappresentativa del modello proposto e delle sue maggiori caratteristiche, è ravvisabile nella "Science and Technology Digital Library" (S&TDL), un modello certamente un po' datato, ma ancora efficace. Si tratta di un progetto triennale (2012-2015) nato nell'ambito di un protocollo tra l'allora Ministero dell'Istruzione, dell'Università e della Ricerca (MIUR) e il Consiglio Nazionale delle Ricerche (CNR) con l'obiettivo di sviluppare un sistema integrato per l'accesso all'informazione scientifica e al patrimonio storico-culturale, in grado di rivolgersi sia alla comunità scientifica nazionale, sia a un'utenza più ampia e variegata, garantendo una gestione e un utilizzo efficienti (Lancia & Reggiani, 2013). Al crocevia fra ricerca scientifica, storia e cultura, la sua realizzazione ha dovuto affrontare le sfide di un'estrema ricchezza e varietà – di contenuti, contesti, formati, sistemi… – e ha quindi comportato l'integrazione di una pluralità di prospettive, saperi e discipline, in un'ottica multi-, inter- e trans-disciplinare. Si è pertanto fondata sull'impegno di gruppi di lavoro versatili e polivalenti, formati di esperti dei differenti domini presenti, in particolare *digital humanist* – specialisti dell'informazione, di comunicazione e di didattica, filologi e archeologi; su una sistematica co-progettazione e su un approccio multidimensionale ed evolutivo ispirato al *digital thinking*, che ha abbracciato lo sviluppo e la sperimentazione non soltanto di un'infrastruttura complessa, ma anche di sistemi, metodologie e strumenti organizzativi e gestionali avanzati e di applicazioni e servizi digitali e informativi innovativi, concepiti per pubblici diversificati.
Gli anni successivi hanno visto numerosi esempi di riutilizzo funzionale, spesso almeno in parte innovativo, di componenti, soluzioni, metodologie, tecnologie e servizi messi a punto nel progetto citato. In particolare, si è assistito alla gemmazione dalla sezione DH della S&TDL – significativa ma non predominante – di non poche *digital library* finalizzate alla gestione, valorizzazione e conservazione a lungo termine di patrimoni documentari e archivistici, come ad es. quella dedicata alle risorse di interesse storico-culturale e giuridico-amministrativo di ARSIAL, relative agli usi civici e ai domini collettivi del Lazio.
Nell'evoluzione della S&TDL da progetto a infrastruttura collaborativa (DOMUS - Infrastructure for Digital Objects)[6] e negli impieghi di quest'ultima prevalentemente nel perimetro delle DH, si sono evidenziati attributi e parole chiave ancora attuali: l'interoperabilità, la modularità, la logica di federazione; una modellizzazione e una progettazione dinamiche ed evolutive, che si sono calibrate via via sulle esigenze di utenze reali; un'architettura distribuita e *multi-layered*, la quale ha garantito strategie stratificate per la gestione dell'intero ciclo di vita delle risorse, superandone l'incoerenza ma salvaguardandone

[5] finalizzato alla promozione della collaborazione fra le discipline umanistiche e le rispettive comunità disciplinari grazie alla definizione di standard condivisi e all'armonizzazione delle policy. Cfr. https://www.parthenos-project.eu/ (cons. 14/04/2026).
[6] contraddistinta da elevate affidabilità, robustezza e scalabilità e in grado di gestire grandi volumi di oggetti digitali assai diversificati. Cfr. https://domus.cnr.it/ (cons. 14/04/2026).

l'eterogeneità. Tutti elementi, questi, che ne hanno accentuato, nel passaggio da sistema a servizio, l'impianto partecipato e la vocazione pubblica.
I tratti distintivi sopra rilevati si rinvengono con un grado di sviluppo e perfezionamento assai più avanzato e sofisticato in una della maggiori iniziative CNR di ambito DH, il progetto PNRR H2IOSC - Humanities and cultural Heritage Italian Open Science Cloud (2022-2026).[7] Esso ha previsto la creazione di un cluster collaborativo di infrastrutture di ricerca europee leader nei settori delle scienze umane e sociali e del patrimonio culturale attraverso la federazione dei rispettivi nodi nazionali, così da offrire a ricercatori, imprese e cittadini un ambiente multidisciplinare aperto, che dà accesso a risorse, strumenti e servizi di avanguardia per condurre ricerche innovative ad alta intensità computazionale. Nel Marketplace – il centro nevralgico del sistema e l'ambiente di accesso unico per la condivisione, la gestione e l'utilizzo collaborativi e flessibili di dati, metodologie, strumenti e servizi – il co-design e l'integrazione hanno compiuto un importante balzo in avanti verso la costruzione di un ecosistema della ricerca DH aperto e inclusivo: in esso ricercatori, informatici e utenti esterni possono collaborare ad attività progettuali e di ricerca fortemente interconnesse, nonché creare, condividere e utilizzare soluzioni duttili e altamente personalizzate. Ciò soprattutto grazie al sistema di orchestrazione, principale innovazione del Marketplace, che permette il coordinamento e la gestione armonizzati e scalabili di processi, applicazioni e servizi multipli, assicurandone l'elevata profilazione e ottimizzandone l'integrazione reciproca grazie alla massimizzazione delle prestazioni delle API e all'impiego più efficiente delle risorse computazionali (Sichera et al., 2024; Marras et al., 2025). Ampio spazio è stato dato inoltre alla progettazione e alla diffusione di conoscenze e competenze interdisciplinari e connesse ai domini specifici delle infrastrutture del cluster, per l'insegnamento/apprendimento e in un'ottica di outreach e public engagement: una piattaforma dedicata mette a disposizione delle comunità di ricerca e di gruppi target diversificati una vasta gamma di risorse e servizi per la formazione e la didattica – una pluralità di learning objects, corsi strutturati, moduli ad elevata granularità, eventi formativi… – predisposti con una metodologia ispirata al modello FAIR by design del progetto Skills4EOSC,[8] liberamente accessibili e riusabili (Pedonese et al., 2025). Si sono così enfatizzate e rafforzate la vocazione pubblica dell'iniziativa e le sue profonde sinergie con realtà europee similari e con segmenti crescenti della società nazionale.
Un ultimo esempio paradigmatico è offerto dal progetto *Beni e patrimonio culturale: governare la progettazione* (PON "Governance e capacità istituzionale, 2014-2020"), durato oltre quattro anni (2017-2021) e finalizzato alla definizione di un modello di *governance* inter-istituzionale multilivello (politico-strategico, giuridico-normativo, organizzativo, semantico, tecnico e tecnologico) e multi-stakeholder, in grado di migliorare significativamente il coordinamento e la gestione di progetti di elevata complessità, condotti in regime di partnership inter-istituzionale all'interno di uno scenario, quello dei beni e del patrimonio culturale, vasto ed eterogeneo, contrassegnato da frammentazione e instabilità. La sperimentazione del modello di governance, principale output progettuale, è avvenuta fin dal principio all'interno di un caso di studio, che ha previsto la progettazione di un portale dedicato alla storia e alla cultura politica italiana, in grado di assicurare un accesso unico semplificato ai contenuti archivistici e documentali messi a disposizione dalle istituzioni nazionali e dagli enti culturali presenti nella partnership progettuale.[9] La modellizzazione di alto livello, compiuta in ambito gestionale e organizzativo e ispirata agli approcci informatici e agili, si è dunque precocemente misurata con una realizzazione di natura tipicamente DH, instaurando e alimentando un'interazione dialogica feconda con il caso di studio. La metodologia adottata e la natura stessa del modello, ibride, pluralistiche e adattive, improntate alla multi- e inter-settorialità, ripropongono integralmente tutti i punti nodali precedentemente esaminati e ne forniscono la sintesi compiuta in un prodotto esemplare di *digital thinking*. Di più, il modello di governance, volto allo sviluppo di potenzialità e alla creazione e al rafforzamento di competenze nell'ecosistema della cultura, è stato nativamente configurato per rispondere alle sollecitazioni degli stakeholder e per raggiungere larghi strati della cittadinanza, in un'ottica aperta e inclusiva (Reggiani, 2022).

[7] https://www.h2iosc.cnr.it/ (cons.14/04/2026).
[8] https://www.skills4eosc.eu/ (cons. 14/04/2026).
[9] Si tratta della forma progettuale iniziale del *Portale delle fonti per la storia della Repubblica italiana*, che si è poi evoluta in un'iniziativa autonoma di ampie dimensioni, oggi liberamente accessibile in rete: https://portalefontirepubblicaitaliana.cnr.it (cons.14/04/2026).

5. VERSO LA PROSPETTIVA PUBBLICA

Come si evince anche dalle realizzazioni sopra illustrate, il modello presentato può offrire una soluzione estremamente versatile e nel contempo scalabile a vari livelli di specializzazioni teoriche e di contesti e prassi operative, in relazione sia a necessità strutturali che a punti di vista e obiettivi particolari.
Esso raccoglie una sfida non facile: tenta infatti di chiarire, comporre e governare la tensione fra l'esigenza di un riconoscimento della specificità delle DH e le tendenze centrifughe, intrinseche alla radicale interdisciplinarità che le contraddistingue.
Nell'attuale cornice dell'Open Science e della Citizen Science, la proiezione verso l'esterno, l'altro da sé, e l'apertura nativa alle differenti comunità scientifiche e alla società – qualità costitutive del modello come della costellazione DH – facilitano la trasformazione dei saperi disciplinari in abilità e competenze chiave più generali e trasversali che, pur mantenendo una rigorosa connotazione metodologica, possono diventare in tal modo patrimonio comune delle realtà professionali e dei cittadini. Appare di conseguenza evidente il crescente coinvolgimento del settore nella prospettiva pubblica, coinvolgimento che, enfatizzando la peculiare funzione sociale e civile delle DH, ne giustifica e consolida la consistenza disciplinare e il valore identitario e insieme contribuisce a potenziare la consapevolezza dei cittadini.

BIBLIOGRAFIA

Averbukh, V. (2015). Global Visualization Metaphors. *2nd Conference and Exhibition on Semiotics & Visual Communication. 2-4 October 2015. Lemesos. Cyprus* (2015). https://www.researchgate.net/publication/282655107 (cons. 14/04/2026).

Balagtas, P. (2024). *Making Futures Work: Integrating Futures Thinking for Design, Innovation, and Strategy*. Oreilly & Associates Inc.

Bastiaansen, C. A., & Van Dun, D. H. (2025). Exploring the effective agile team model: a qualitative mixed-methods study among practitioners. *Team Performance Management: An International Journal*, *31*(7-8), 754–778. https://doi.org/10.1108/TPM-09-2024-0103.

Brennan, S. (2016). "Public, First". In M. K. Gold & L. F. Klein (Eds.), *Debates in the Digital Humanities* (pp. 384-90). Minneapolis: University of Minnesota Press.

Burdick, A., Drucker, J., Lunenfeld, P., Presner, T., & Schnapp, J. (2012), *Digital_Humanities*. Massachusetts Institute of Technology. Trad. ital. *Umanistica_Digitale*. Milano: Mondadori, 2014.

Castellucci, P. (2009). *Dall'ipertesto al Web: storia culturale dell'informatica*. Roma-Bari: Laterza.

Ciotti, F., Corradini, E., Cugliana, E., D'Agostino, G., Ferroni, L., Fischer, F., Lana, M., Monella, P., Roeder, T., Rosselli Del Turco, R., & Sahle, P. (2022). Digital Scholarly Editions Manifesto. *Umanistica Digitale*, *6*(12), 103–108. https://doi.org/10.6092/issn.2532-8816/14814. Trad. ital. del Manifest für digitale Editionen. *DHdblog. Digital Humanities im deutschsprachigen Raum.* (2022, 11 March). https://dhd-blog.org/?p=17563 (cons. 14/04/2026).

Ciotti, F. (Ed.). (2023). *Digital Humanities. Metodi, strumenti, saperi*. Roma: Carocci.

Ciula, A., & Eide, Ø. (2017). Modelling in Digital Humanities: Signs in Context. *Digital Scholarship in the Humanities*, *32*(1), i33–i46 https://doi.org/10.1093/llc/fqw045.

Ciula, A., Eide, Ø., Marras, C., & Sahle, P. (2023). *Modelling Between Digital and Humanities. Thinking in Practice.* ISBN Paperback978-1-80511-098-9. https://doi.org/10.11647/OBP.0369.

Ciula, A., Eide, Ø., Marras, C., & Sahle, P. (2018). Models and Modelling between Digital and Humanities - A Multidisciplinary Perspective. *Historical Social Research/Historische Sozialforschung. Supplement*, 31, 46-58. https://www.ssoar.info/ssoar/handle/document/60873.

Ciula, A., & Marras, C. (2016). Circling around Texts and Language: Towards "Pragmatic Modelling". *Digital Humanities Quarterly*, *10*(3). DOI: https://doi.org/10.63744/yv9cxasfbgp4

Ciula, A., & Marras, C. (2019). Exploring a Semiotic Conceptualisation of Modelling in Digital Humanities Practices. In A. Olteanu, A. Stables & D. Borţun (Eds.), *Meanings & Co: The Interdisciplinarity of Communication, Semiotics and Multimodality.* Numanities - Arts and Humanities in Progress (vol. VI, pp. 33-52). Springer International Publishing. https://www.springer.com/us/book/9783319919850.

Ciula, A., Noel, G., Caton, P., Ferraro, G., Smithies, J., & Monteiro Vieira, J. M. (2023). The place of models and modelling in Digital Humanities: some reflections from a Research Software

Engineering perspective. In M. Annisius, E. Arestau, J. Burkhardt, N. Herold & R. Sierig (Eds.), *Vielfalt und Integration - diversitá ed integrazione - diversité et intégration: Sprache(n) in sozialen und digitalen Räumen Eine Festschrift für Elisabeth Burr* (pp. 261-281). Leipzig: Universitätsbibliothek. https://nbn-resolving.org/urn:nbn:de:bsz:15-qucosa2-852367.

Cosgrove, J., & Cachia, R. (2025). *DigComp 3.0: European Digital Competence Framework* (5th ed.). Luxembourg: Publications Office of the European Union. DOI: 10.2760/0001149 (online), 10.2760/7379058 (print).

Duplessis, P. (2017, mars). Les trois matrices disciplinaires de l'Information-Documentation. *Les Trois Couronnes*. *Didactique de l'information-documentation et identité professionnelle de l'enseignant documentaliste. Site de Pascal Duplessis*. https://ltc.apden.org/didactique-information/les-trois-matrices-disciplinaires-de-l-information-documentation.html (cons. 14/04/2026).

Dupret, K., Reyes, E., Ziku, M., Szoniecky, S., & Caridakis, G. (2024). Enhancing cultural and societal engagement through participatory methods: cross-disciplinary learning perspectives. In F. Spigarelli, L. Kempton & L. Compagnucci (Eds.), *Entrepreneurship and Digital Humanities. How Universities Shape the Future of Labour* (pp. 167–183). Edward Elgar Publishing.

Eide, Ø. (2014-2015). Ontologies, Data Modelling, and TEI. *Journal of the Text Encoding Initiative*, 8. https://doi.org/10.4000/jtei.1191.

Elkins, J., & Fiorentini, Erna (2021). *Visual Worlds: Looking, Images, Visual Disciplines.* Oxford: Oxford University Press.

Elleström, L. (2014). *Media Transformation: The Transfer of Media Characteristics Among Media.* London: Palgrave Macmillan.

Eshet-Alkalai, Y. (2009). A Holistic Model of Thinking Skills in the Digital Era. In *Encyclopedia of Distance Learning* (2nd ed., pp. 1088-1093).

Flanders, J., & Jannidis, F. (2015). *Knowledge Organization and Data Modeling in the Humanities*. http://www.wwp.northeastern.edu/outreach/conference/kodm2012/flanders_jannidis_datamodeling.pdf (cons. 14/04/2026).

Flanders, J., & Jannidis, F. (Eds.). (2019). *The Shape of Data in Digital Humanities: Modeling Texts and Text-Based Resources*. Digital Research in the Arts and Humanities. London: Routledge, Taylor & Francis Group.

Heinisch, B. (2020). Citizen Humanities as a Fusion of Digital and Public Humanities? *Magazén*, *1*(2), 143-178. https://edizionicafoscari.it/media/pdf/article/magazen/2020/2/art-10.14277-mag-2724-3923-2020-02-001_dX1nq7p.pdf (cons. 14/04/2026).

Khalil, C., Fernandez, V., & Houy, T. (2013). Can Agile Collaboration Practices Enhance Knowledge Creation between Cross-Functional Teams? In P. J. Benghozi, D. Krob & F. Rowe (Eds.), *Advances in Intelligent Systems and Computing*: *Vol 205. Digital Enterprise Design and Management 2013* (pp. 123–133). Berlin-Heidelberg: Springer. https://doi.org/10.1007/978-3-642-37317-6_11.

Lancia, M. & Reggiani, L. (2013, 15 ottobre). La biblioteca digitale fra Marconi e la scienza aperta. *Scienza in rete*. http://www.scienzainrete.it/contenuto/articolo/maurizio-lancia-lisa-reggiani/biblioteca-digitale-fra-marconi-e-scienza-aperta (cons. 14/04/2026).

Leonardi, P., & Neeley, T. (2022). *The Digital Mindset: What It Really Takes to Thrive in the Age of Data, Algorithms, and AI*. Harvard Business School Press.

Liburdi, A., Marras, C., & Russo, A. (2019). Infrastrutture, terminologie e policy per la ricerca umanistica: note per un confronto interdisciplinare. *Conferenza GARR 2018 - Data (R)evolution - Selected Papers. Cagliari, 3-5 ottobre 2018* (pp. 82-86). ISBN: 978-88-905077-8-6. Associazione Consortium GARR. DOI: 10.26314/GARR-Conf18-proceedings.

Lundene, K., & Mohagheghi, P. (2018). How autonomy emerges as agile cross-functional teams mature. *XP '18: Proceedings of the 19th International Conference on Agile Software Development: Companion.* Article 18, 1-5, Association for Computing Machinery. https://doi.org/10.1145/3234152.3234184.

Malazita, J. W., Teboul, E. J., & Rafeh, H. (2020). Digital Humanities as Epistemic Cultures: How DH Labs Make Knowledge, Objects, and Subjects. *Digital Humanities Quarterly*, *14*(3). DOI: https://doi.org/10.63744/3zv72m4dkb4g.

Marras, C., Fabiani, V., Pasini, E., Reggiani, L., Sichera, P., Ongaro, P., & Rossi, M. (2025). Modelli e tecnologie integrate e innovative per una cittadinanza digitale equa e sostenibile. In S. Rebora, M. Rospocher & S. Bazzaco (a cura di), *Diversità, Equità e Inclusione: Sfide e Opportunità per l'Informatica Umanistica nell'Era dell'Intelligenza Artificiale*, *Proceedings del XIV Convegno Annuale AIUCD2025. Quaderni di Umanistica Digitale* (pp. 538-543). AIUCD. ISBN 978-88-942535-9-7. DOI: 10.6092/unibo/amsacta/8380.

Nelson, T. (1987). *Computer Lib/Dream Machines* [1974]. Redmond (WA): Tempus Books.

Pawlicka-Deger, U., & Thomson, C. (Eds). (2023). *Digital Humanities and Laboratories: Perspectives on Knowledge, Infrastructure and Culture*. London: Routledge. https://doi.org/10.4324/9781003185932.

Pedonese, G., Frontini, F., Ottaviani, R., Boschetti, F., Spadi, A., Francalanci, L., Scognamiglio, A., Restaneo, P., Chaban, A., Striova, J., & Benassi, L. (2025). Learning Materials as FAIR Digital Objects: A Shared Methodology for Training in H2IOSC. *Umanistica Digitale*, *9*(20), 361–380. https://doi.org/10.6092/issn.2532-8816/21190.

Radin, M. A., & Danylchenko-Cherniak, O. (2025). *Creative Learning Styles in the Digital Thinking Era Education 1.0 – 7.0*. Rochester Institute of Technology. https://doi.org/10.1142/14298.

Reggiani, L. (2008). France: the courage to say "Documentation". In C. Basili (Ed.), *Information Literacy at the crossroad of Education and Information Policies in Europe* (pp. 92-164). Roma: Consiglio Nazionale delle Ricerche. ISBN 978-88-8080-096-5.

Reggiani, L. (2022). Un modello di governance progettuale per i beni e il patrimonio culturale. *ICTLABOR - Rapporti tecnici*, n. 5 (RT 005/2022), ISSN 2785-3284.

Sichera, P., Marras C., & Pasini, E. (2024, 19 novembre). *Orchestrazione API per workflow applicativi nell'ambito delle Digital Humanities*. Zenodo. https://doi.org/10.5281/zenodo.14187534.

Svensson, P. (2010). The Landscape of Digital Humanities. *Digital Humanities Quarterly*, *4*(1). https://doi.org/10.63744/5dhf3ymfs2gg.

Wouter, A., & Uyttenhove, P. (2012). Analogous Spaces: An Introduction to Spatial Metaphors for the Organization of Knowledge. *Library Trends*, *61*(2), 259-270. DOI: 10.1353/lib.2012.0037.

Mappe, scale e interpretazione. Per una cartografia digitale semantica dello spazio narrativo

LORENZO SABATINO
ALMA MATER STUDIORUM, UNIVERSITY OF BOLOGNA, ITALY- LORENZO.SABATINO3@UNIBO.IT

ABSTRACT (ITALIANO)

Il contributo propone una riflessione metodologica sul ruolo della cartografia digitale come interfaccia interpretativa, capace di sostenere un paradigma di *scalar reading* per lo studio della letteratura moderna e contemporanea, superando il rischio di riduzione positivista dello spazio narrativo. Attraverso due casi di studio complementari, rappresentati dall'*Atlante illustrato dei Promessi sposi* e il workflow *GaddAtlas* sul Pasticciaccio di Carlo Emilio Gadda, si dimostra come l'integrazione di cartografia esplorativa e modellizzazione ontologica consenta il passaggio progressivo dall'analisi visuale di primo livello a forme di interrogazione semantica specialistiche, formalizzando incertezza, stratificazione e ambivalenza dello spazio del racconto.
Parole chiave: Spatial Humanities; Literary Cartography; WebGIS; Ontologie; Spazio narrativo

ABSTRACT (ENGLISH)

Maps, Scales, and Interpretation. Toward a Semantic Digital Cartography of Narrative Space.
This paper offers a methodological reflection on digital cartography as an interpretative interface supporting a paradigm of scalar reading for the study of modern and contemporary literature, addressing the risk of a positivistic reduction of narrative space. Drawing on two complementary case studies, *The Betrothed Illustrated Atlas* and the *GaddAtlas* workflow applied to Carlo Emilio Gadda's *Quer pasticciaccio brutto de via Merulana* – it demonstrates how the integration of exploratory cartography and ontological modelling allows a progressive shift from first-level visual exploration to specialised forms of semantic querying, making uncertainty, stratification, and ambivalence in narrative space explicit.
Keywords: Spatial Humanities; Literary Cartography; WebGIS; Ontologies; Narrative Space

1. INTRODUZIONE

Negli ultimi decenni, lo *spatial turn* (Lefebvre, 1974; Jameson, 1989; Soja, 1996) ha profondamente inciso sugli studi di geografia letteraria (Moretti, 1997; Jakob, 2005; Collot, 2011) e geocritica (Westphal, 2007; Iacoli, 2008; Sorrentino, 2010), favorendo l'adozione di mappe digitali, GIS e WebGIS come strumenti di analisi esplorativa delle dimensioni spaziali del testo. Tali strumenti hanno dimostrato una notevole efficacia nel far emergere pattern, relazioni topografiche e configurazioni narrative difficilmente osservabili attraverso pratiche di *close reading*. L'adozione di dispositivi cartografici applicati al testo espone tuttavia a un rischio epistemologico noto: la riduzione positivistica dello spazio narrativo a collezione di punti georeferenziati, con un conseguente appiattimento di complessità interpretativa (Harley, 1989; Reuschel & Hurni, 2011; Stockhammer, 2013).
Il presente contributo ripensa la cartografia digitale come ambiente di esplorazione capace di sostenere un lavoro di *scalable reading* (Weitin, 2017; Krautter, 2024), che consenta cioè di muoversi analiticamente tra differenti livelli di astrazione spaziale – dal referente geografico alla funzione narrativa, fino alle costruzioni concettuali e simboliche – all'interno di un'infrastruttura coerente. La tesi portante coincide con la convinzione che questa transizione possa essere mediata efficacemente dall'integrazione tra cartografia esplorativa e modellizzazione ontologica: ontologie e knowledge graph consentono di dichiarare le categorie interpretative che distinguono entità geografiche puntuali, luoghi narrativi vaghi e concetti spaziali non geolocalizzabili, separando formalmente i livelli di referente reale, narrazione legata allo spazio e interpretazione situata (Eide, 2015; Martin-Rodilla & Gonzalez-Perez, 2018; Pratelli, 2025). Tale evoluzione segna il passaggio da un approccio puramente descrittivo a pratiche interpretative in cui la "datafication" dell'informazione spaziale è mediata da un dispositivo che intende sfidare l'egemonia del dato oggettivo, facendo dialogare la formalizzazione semantica con la volontà di preservare l'ambiguità intrinseca dell'ermeneutica testuale (Cooper et al., 2016; Bushell, 2020, Fiorentino et al., 2025).

Tale impostazione è discussa attraverso due casi volutamente eterogenei: i *Promessi Sposi* (opera fuori diritto d'autore caratterizzata da un progetto editoriale-iconografico emblematico) e *Quer pasticciaccio brutto de via Merulana* (opera novecentesca sottoposta a vincoli di copyright), per mettere in evidenza due traiettorie metodologiche complementari, di valorizzazione transmediale del manufatto editoriale e di accesso semantico all'opera tramite dati formalizzati.

2. STATO DELL'ARTE

Lo *spatial turn* ha ridefinito lo spazio narrativo come costruzione culturale e relazionale, influenzando numerose correnti di teoria letteraria, dal cognitivismo (Ryan, 2003; Caracciolo, 2014), alla geografia letteraria, fino alla geocritica, che in particolare ne ha messo in luce la natura polisemica e conflittuale, analizzando lo spazio come nodo di rappresentazioni parziali legate a molteplici regimi discorsivi (Alfano, 2010; Papotti & Tomasi, 2014). In parallelo, le Digital Humanities, a partire dai primi esperimenti di cartografia letteraria e *distant reading* spaziale, hanno sviluppato metodi e progetti legati all'uso di mappe digitali e GIS/WebGIS, mostrando le potenzialità di strumenti cognitivi per l'analisi esplorativa di pattern distributivi, e insieme evidenziandone i limiti legati a processi di riduzione e vincolo espressivo (Reuschel et al., 2013, Drucker, 2014). Una parte crescente della riflessione si è pertanto concentrata non solo sulla possibilità di mappare il dato umanistico, ma sulle condizioni epistemologiche che rendono tale operazione scientificamente fondata, distinguendo fra uso illustrativo ed euristico del dispositivo cartografico (Fischer, 1999; Drecki, 2007; Windhager et al., 2019; El Khatib & Schaeben, 2020).
Negli ultimi anni, dunque, il focus si è progressivamente spostato dallo strumento allo statuto epistemologico della mappa: analogamente a ogni testo, essa non costituisce un contenitore neutro, ma un dispositivo situato e selettivo (Stockhammer, 2013), definito attraverso scelte retoriche e astrazioni simboliche (Piatti, 2008; von Ungern-Sternberg, 2009; Bushell, 2020). Ne deriva una funzione prevalentemente "navigazionale" (Bosco, 2025) ed euristica, orientata alla costruzione di ipotesi più che alla congruenza geometrica con il reale.
Sebbene la riflessione sulla visualizzazione dell'incertezza nei dati umanistici abbia prodotto risultati rilevanti (Thomson et al., 2005; Fisher, 2009; Otten et al., 2018; Windhager et al., 2019), mostrando come vaghezza, incompletezza, ambiguità e conflittualità delle fonti debbano essere trattate alla stregua di proprietà informative multilivello (MacEachren et al., 2005; MacEachren et al., 2012), la loro traduzione in pratiche visive stabili resta ancora fortemente discontinua e scarsamente adattata all'ambito letterario. Di fatto, come osservavano già Senaratne e Gerharz (2011), larga parte dei sistemi di visualizzazione tende ad attenuare il valore dell'incertezza per preservare la leggibilità del dataset e non compromettere l'efficacia della rappresentazione, restituendo *de facto* una rappresentazione più coerente e stabile dei dati di quanto le fonti autorizzino.
Per l'ambito strettamente letterario, iniziative fondanti come *A Literary Atlas of Europe* (2006; Reuschel et al., 2009) hanno mostrato quanto lo strumento cartografico possa dialogare con concetti di "uncertainty" (Fisher, 1999; Piatti et al., 2009; Reuschel & Hurni, 2011) e vaghezza topografica, modellando un più recente parterre di progetti che interpreta la cartografia come spazio critico, supportato da categorie esplicite e dispositivi di commento – come nei recenti sviluppi del progetto PRIN LEDA (n.d.)[1] e del progetto di dottorato *A World of Paper* (Aletta, 2025). Progettualità simili, sebbene teoricamente molto avvertite sul versante critico, tendono a dichiarare il trattamento dello statuto a livello discorsivo, categoriale o metadescrittivo, veicolando l'informazione in convenzioni grafiche non sufficientemente stabili, che rimangono perlopiù idiosincratiche, poco scalabili e difficilmente trasferibili.
Un ulteriore passaggio riguarda il trattamento del dato nel dialogo tra testo e visualizzazione. In questa prospettiva, la modellazione semantica garantisce un processo critico intermediale che rende osservabili gli scarti tra espressione linguistica, articolazione concettuale e rappresentazione cartografica (Eide, 2015; Martin-Rodilla & Gonzalez-Perez, 2018; Car, 2024). Il vantaggio metodologico di uno schema formale esplicito si definisce a questo livello, come ha dimostrato il recente lavoro dottorale (Pratelli, 2025) che propone un framework ontologico per la rappresentazione geospaziale del testo narrativo distinguendo tra luogo fenomenico, esposto all'incertezza delle fonti, e la sua espressione geometrica dichiarativa. Se tale approccio costituisce un avanzamento notevole sul versante dell'organizzazione

[1] For an Atlas of Italian Ecological Literature: From the Great Acceleration to the Pandemic (LEDA) - FINANZIAMENTO PRIN 2022 PNRR, è un progetto realizzato in collaborazione fra l'Università IULM di Milano, l'Università degli Studi di Siena e l'Università di Bologna. URL: https://progettoleda2022prin.github.io/atlante/index.html (cons. 18/04/2026)

dell'informazione, rimane tutt'ora aperta l'applicazione di un linguaggio visuale capace di restituire, in termini di *interaction design*, la stessa complessità epistemica postulata.
In continuità con tali presupposti, il presente contributo si colloca all'intersezione tra modellazione semantica e visualizzazione, proponendo un ambiente interpretativo che mira a integrare la formalizzazione ontologica dell'incertezza con la sua realizzazione grafica, facendo interagire come livelli contigui la semantica del dato e la densità ermeneutica del testo.

3. METODO E CASI DI STUDIO

I due modelli qui presentati riconfigurano la mappa come interfaccia modulare per un'esplorazione scalare dello spazio narrativo, consentendo di muoversi tra referente geografico, funzione narrativa e costruzioni interpretative. Nel caso manzoniano, la scalabilità della mappa valorizza l'assetto transmediale e il progetto editoriale d'autore; nel caso gaddiano, la scalabilità assume una funzione ermeneutica, per cui la mappa diviene medium per visualizzare come il valore del luogo, nel testo, è costruito, trasformato o messo in crisi. In entrambi i casi, la mappa non si pone come risultato del processo interpretativo, ma piuttosto come interfaccia che lo articola in una sequenza di accessi differenziati, in cui il livello esplorativo e quello semantico risultino reciprocamente connessi.
Operativamente, entrambi i workflow combinano: (1) costruzione e normalizzazione dei dati da fonti primarie; (2) pubblicazione WebGIS con filtri e timeline; (3) dispositivi per gestire densità e sovrapposizioni; (4) nel caso manzoniano, visualizzatore IIIF per accesso alle fonti; nel caso gaddiano, modellizzazione ontologica in OWL 2 (Hitzler et al., 2012) e SKOS (Miles & Bechhofer, 2009), e tracciamento delle asserzioni interpretative in PROV-O (Lebo et al., 2013), distinguendo referente, narrazione e interpretazione senza naturalizzare le scelte ermeneutiche. Lo scalar reading è sostenuto dall'architettura modulare: la mappa come layer esplorativo (filtri per tipologia, capitoli e personaggi), l'interazione con card per metadati narratologici, l'accesso a fonti (IIIF) o interrogazioni semantiche per analisi specialistica. Questa impostazione si colloca nell'alveo di una linea di ricerca che attribuisce all'interazione un ruolo di scoperta cognitiva privilegiata, capace di sostenere passaggi controllati fra overview, dettaglio, confronto e ritorno alla fonte, secondo una logica di esplorazione critica più che di sola consultazione (Hinrichs et al., 2015; Morini et al., 2025).

3.1 Atlante Illustrato dei Promessi Sposi

L'Atlante è un WebGIS esplorativo orientato alla valorizzazione della dimensione editoriale e iconografica dell'opera manzoniana, assumendo le illustrazioni inserite nell'edizione in volume della *Quarantana* (Manzoni, 1840) quali unità primarie di accesso allo spazio narrativo e alla volontà autoriale. L'utilizzo della mappa rappresenta dunque un gesto epistemologico situato: l'obiettivo è quello di restituire centralità alla volontà d'autore, che commissiona agli illustratori della bottega di Francesco Gonin una puntuale trasposizione iconografica di passaggi chiave del romanzo. In tal senso, la visualizzazione combinata di testo e immagine, navigabile tramite mappa, consente di rileggere l'opera nella filigrana dei momenti e dei luoghi ritenuti centrali per la comprensione della storia dei personaggi. Più in generale, il caso manzoniano mostra come la cartografia digitale possa funzionare non soltanto come supporto orientativo, ma come dispositivo di ricomposizione transmediale del rapporto tra spazio, illustrazione e progetto editoriale, rendendo osservabili regolarità e densità iconografiche difficilmente ricostruibili attraverso la sola consultazione lineare dell'edizione.
La pipeline adotta formati interoperabili per la riproducibilità: acquisizione metadati dalla Digital Library del FICLIT dell'Università di Bologna tramite API REST Omeka S,[2] integrazione immagini tramite Manifest IIIF, arricchimento di metadati del passo testuale e dell'item illustrato (capitolo, personaggi coinvolti, illustratore individuato), trasformazione in GeoJSON. Il dataset è costituito da 222 occorrenze geolocalizzate, derivanti dalle 505 illustrazioni della Quarantana, associate a coordinate normalizzate tramite geocoding automatico in *Nominatim* (basato su dati *OpenStreetMap*). Dataset, documentazione e codice sono disponibili in una release (v1.0-review) della repository GitHub, per garantire riproducibilità completa.[3]

[2] API REST Endpoint: https://dlrc.ficlit.unibo.it/api/items

[3] GitHub Repository. URL: https://github.com/dh-research-projects/web-atlas/releases/tag/v1.0-review

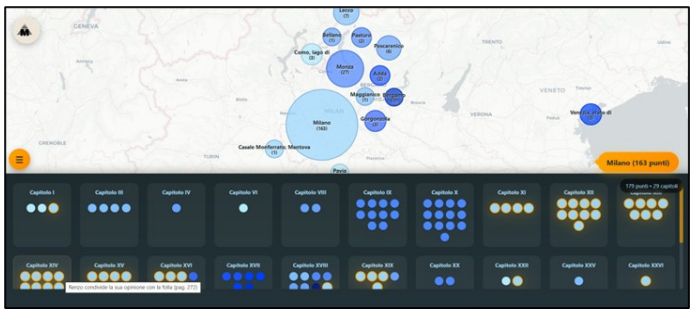


Figura 1: Mappa con timeline aperta e visualizzazione dei cluster tramite *force-directed simulation*

Il front-end (HTML/CSS/JavaScript + Leaflet.js) gestisce la densità dei marker per i singoli item attraverso clustering e strategie di disambiguazione visuale basate su D3.js (vd. Figura 1): una *force-directed simulation*, che redistribuisce dinamicamente marker coincidenti, preservando prossimità semantica e leggibilità; visualizzazioni radiali (*spider graph*) in SVG/D3 per visualizzare le relazioni locali. La navigazione scalare è supportata da timeline interattiva e dall'utilizzo di card, che espongono le informazioni e i metadati dell'item. La timeline, scandita per capitoli, consente di recuperare il singolo item aprendo la card al click del mouse, o di incrociare tutti gli item corrispondenti allo stesso luogo con il passaggio del mouse sull'item, retro-illuminando tutti gli item della timeline corrispondenti allo stesso luogo, con un conteggio dinamico laterale del numero complessivo. Le card, dunque, possono essere attivate tramite interazione diretta con mappa, all'interno della timeline, o essere campionate con l'utilizzo del menù di filtri, che consente di restringere per capitolo, per personaggio, per illustratore, per luogo o per range di pagine.
Un passaggio decisivo verso la fonte è infine garantito dal comando "vedi nell'edizione" della singola card, che apre un Mirador, configurato tramite IIIF Manifest, all'interno della stessa interfaccia (vd. Figura 2), accostando pagina dell'edizione e illustrazione corrispondente. In tal modo, l'atlante non si limita a visualizzare dati derivati dall'opera, ma conserva un nesso metodologicamente forte con il documento editoriale di partenza, evitando che la mediazione cartografica si sostituisca impropriamente alla fonte.

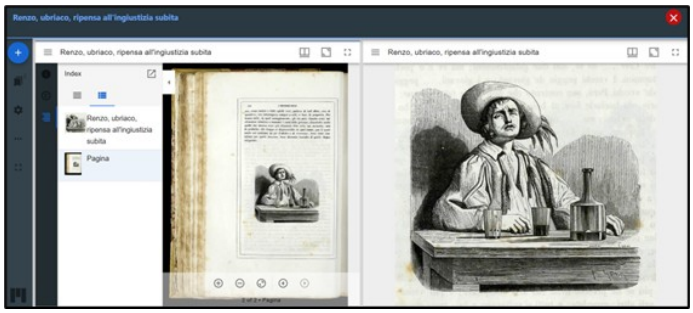


Figura 2: Mappa con timeline aperta e visualizzazione dei cluster tramite *force-directed simulation*

3.2 GaddAtlas

GaddAtlas configura la cartografia come interfaccia ermeneutica per l'analisi del *Pasticciaccio* (Gadda, 2018), un'opera sottoposta a vincoli di copyright, integrando WebGIS esplorativa con un'infrastruttura semantica RDF orientata all'analisi e all'interrogazione strutturata della conoscenza. In questo caso, l'impossibilità di costruire un accesso pienamente fondato sulla riproduzione del testo rende ancora più evidente la necessità di una mediazione formale capace di esplicitare i livelli dell'interpretazione e di trasformare l'interfaccia in luogo di accesso critico ai dati.
La prima fase del workflow prevede la costruzione di un dataset in formato JSON normalizzato, attualmente composto da 355 occorrenze e 25 attributi, derivato dall'*Indice dei Luoghi Garzanti* (Gadda, 1993) e da un *close reading* sistematico del testo. Il dataset, tuttora in fase di ampliamento nell'ambito di un secondo ciclo di analisi ed estrazione, caratterizza ciascuna occorrenza attraverso un insieme strutturato di attributi narrativi e coordinate spaziali. L'interfaccia esplora due assetti visuali e concettuali complementari, che riflettono due diverse modalità di costruzione del rapporto tra testo e mappa. Nel modello *map-centric* (vd. Figura 3), la cartografia costituisce il fulcro dell'esperienza interpretativa: i

contenuti testuali e metatestuali vengono destrutturati e ricombinati all'interno di un percorso di esplorazione affidato prevalentemente alla mappa. Nel modello radiale, o *document-centric* (vd. Figura 4), l'interfaccia è invece organizzata in tre fasce funzionali – mappa-testo al centro, apparato descrittivo a sinistra, commento semantico a destra – che restituiscono una relazione più bilanciata e biunivoca tra testualità e cartografia. Le prestazioni e l'accessibilità dell'interfaccia sono garantite dall'adozione di strategie di *lazy loading*, che consentono un rendering ottimizzato dell'insieme dei 355 marker attualmente presenti. Il confronto tra le due configurazioni non risponde soltanto a un'esigenza di design, ma a una domanda metodologica più precisa: verificare quale forma di articolazione tra visualizzazione, commento e apparato descrittivo risulti più adeguata a sostenere letture specialistiche dello spazio narrativo senza ridurre la complessità del testo a semplice *patterning* topografico.

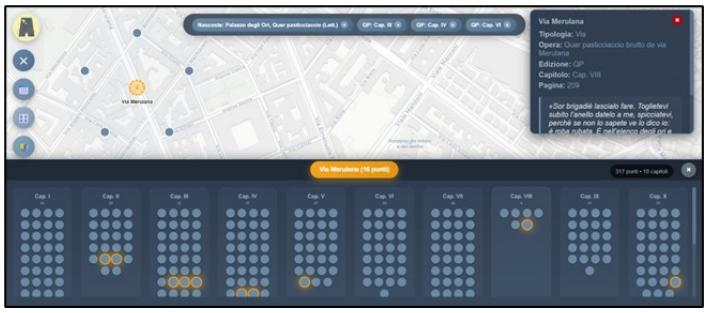


Figura 3: modello *map-centric*: la mappa come fulcro, con timeline/filtri, card e *spider-graph* aperti

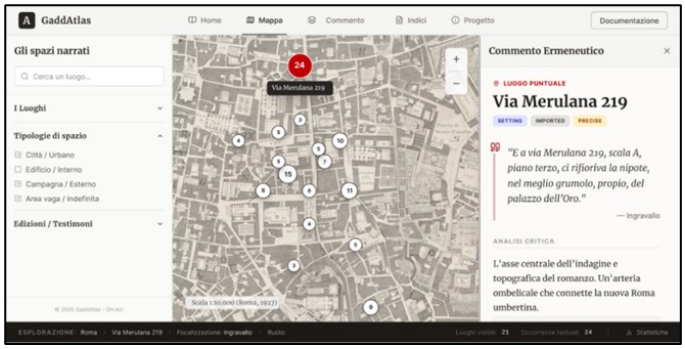


Figura 4: modello *document-centric*: mappa al centro, apparato descrittivo (sx), commento critico (dx)

La seconda fase riguarda la modellizzazione ontologica attraverso un approccio ibrido OWL/SKOS, finalizzato a rendere esplicita la distinzione tra referente spaziale, costruzione narrativa e analisi interpretativa. In questo quadro, OWL definisce la struttura concettuale del dominio (classi e relazioni logiche), mentre SKOS è impiegato per modellare categorie interpretative dal carattere graduale e non rigidamente tassonomico. Il modello consente di articolare tre livelli distinti ma correlati: (1) *ga:NarrativePlace*, relativo al referente spaziale e alla sua possibile controparte reale; (2) un insieme di *skos:ConceptScheme* narratologici che descrivono le modalità attraverso cui lo spazio è costruito, funzionalizzato e percepito nel testo – distinguendo, ad esempio, la funzione diegetica dei luoghi, il loro grado di referenzialità rispetto al reale e il livello di determinazione spaziale; (3) *ga:SpatialInterpretation*, deputato alla formalizzazione delle asserzioni critiche. Una simile articolazione consente di evitare la sovrapposizione fra luogo come oggetto del racconto, luogo come referente geografico e luogo come risultato di un'operazione interpretativa, problema che continua a rappresentare uno dei nodi più delicati nella traduzione del testo narrativo in dati spaziali formalizzati.
Le categorie impiegate nei *Concept Scheme* non sono concepite come strumenti *ad hoc*, ma si fondano su concetti e classificazioni ampiamente attestati nella critica e nelle pratiche di cartografia letteraria, in particolare a partire dai lavori del *Literary Atlas of Europe* e i successivi atlanti letterari più recenti (cfr. stato dell'arte). Tale scelta mira a garantire non solo coerenza interpretativa, ma anche un grado di standardizzazione e interoperabilità nella modellizzazione semantica, estendendo tali principi anche al livello dell'analisi critica. Le asserzioni interpretative sono infine annotate tramite PROV-O, al fine di documentarne provenienza e responsabilità autoriale.

Un esempio concreto di impiego dell'ontologia è presentato in figura (vd. Figura 5), che modella la relazione interpretativa tra il "palazzo dell'Oro" – luogo centrale del romanzo – e "via Merulana, 219", strada nevralgica per il romanzo gaddiano e toponimo in cui il palazzo è situato.

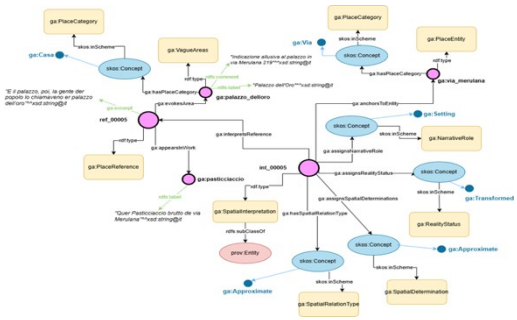


Figura 5: schema di una relazione spaziale complessa tra luogo topografico e spazio narrativo

La trasformazione dei dati in RDF è governata da mapping YAML e da una pipeline ETL in Python (RDFLib). L'architettura ontologica è attualmente sottoposta a una fase di testing e raffinamento, condotta attraverso un insieme di *competency questions* progettate per valutare in modo preliminare la capacità del modello di sostenere interrogazioni critiche specialistiche. In questa fase, le CQs consentono di mostrare come l'integrazione tra ontologia e dati RDF permetta di specializzare l'esplorazione WebGIS, trasformandola progressivamente da strumento di navigazione spaziale a dispositivo di conoscenza orientato all'analisi del valore narrativo dello spazio. È precisamente su questo passaggio che il workflow intende situare il proprio contributo specifico: non nella sola strutturazione di *location data*, ormai ampiamente praticata, ma nella possibilità di passare da una ricognizione visuale di primo livello a interrogazioni che mettano a tema rapporti fra categorie spaziali, personaggi, gradi di precisione, statuti referenziali e modalità di costruzione narrativa del luogo.
Tra le CQs discusse rientrano, ad esempio: "*Qual è il rapporto tra* PlaceEntity *e* VagueArea *in contesti urbani e rurali?*", dove *PlaceEntity* indica luoghi ancorabili a un referente topografico e *VagueArea* aree spaziali indeterminate o percettive; e "*come varia il valore narrativo di uno stesso luogo al variare del personaggio?*". Tali interrogazioni permettono di evidenziare strategie spaziali differenziate – maggiore precisione topografica nello spazio urbano, tensione lirico-percettiva in quello extraurbano – e di formalizzare non solo *dove*, ma *come* e *per chi* lo spazio si costituisce narrativamente. In questo senso, l'interrogazione semantica non aggiunge semplicemente un livello tecnico al WebGIS, ma rende possibile una forma di analisi comparativa e argomentativa che eccede la visualizzazione distributiva dei luoghi e si orienta verso la qualificazione critica delle loro funzioni.
Dataset, ontologia, documentazione e codice ETL saranno rilasciati su GitHub e archiviati su Zenodo con DOI al completamento della validazione del dataset, sotto licenza CC-BY 4.0.

4. CONCLUSIONI E SVILUPPI FUTURI

I due casi di studio mostrano come la cartografia digitale possa operare quale dispositivo modulare di mediazione critica, sostenendo un'esplorazione progressiva dell'opera a diversi livelli di accesso e analisi. Nel caso manzoniano, la mappa assume il ruolo di interfaccia transmediale, fondata sull'integrazione di API pubbliche, IIIF e Mirador, che mette in relazione dati editoriali, volontà autoriali e strumenti interpretativi. Nel caso gaddiano, la cartografia, sviluppata sotto vincoli di copyright, si configura invece come front-end ermeneutico su dati formalizzati, in cui l'ontologia esplicita categorie narratologiche, integra PROV-O per la trasparenza epistemica e abilita interrogazioni semantiche verificabili. Considerati congiuntamente, i due casi consentono di sostenere che il contributo specifico della cartografia digitale agli studi letterari non risieda soltanto nella visualizzazione dello spazio del racconto, ma nella possibilità di organizzare ambienti di ricerca in cui esplorazione, ritorno alla fonte, formalizzazione semantica e interrogazione critica risultino strettamente intrecciati.
Negli sviluppi futuri resta centrale la necessità di rendere espliciti assunti e scelte modellizzanti, evitando la reificazione del modello e governando l'onere manutentivo delle infrastrutture digitali. In questa prospettiva, dataset, documentazione e codice saranno rilasciati in open access (GitHub e Zenodo con DOI). Ulteriori linee di sviluppo riguardano il consolidamento delle tassonomie, la stabilizzazione delle piattaforme e la sperimentazione di visualizzazioni capaci di rendere più leggibile il passaggio

dall'esplorazione visuale all'interrogazione semantica, nonché la conduzione di *user testing* per valutare l'efficacia delle due configurazioni di GaddAtlas (map-centric e document-centric) in termini di usabilità ed esperienza utente.
In una prospettiva più ampia, il progetto mira a sviluppare un'infrastruttura in cui la marcatura semantica esplicita dell'incertezza, resa interoperabile attraverso knowledge graph, sia integrata con dispositivi di visualizzazione capaci di restituirne la funzione narratologica, rendendo osservabile il ruolo che la costruzione dei luoghi assume nella diegesi, nelle strategie cognitive dell'autore e nelle loro relazioni con il contesto biografico e culturale. In tal senso, la traiettoria di ricerca si orienta verso la definizione di ambienti di geo-edizione e di interfacce interpretative avanzate, in cui modellazione semantica e interaction design concorrano a colmare il divario ancora esistente tra formalizzazione dell'incertezza e sua effettiva traduzione in pratiche visuali scalabili e comparabili all'interno delle Spatial Humanities.

BIBLIOGRAFIA

A Literary Atlas of Europe. (2006). https://www.literaturatlas.eu/index.html. (cons. 18/04/2026).

Aletta, A. (2025). A World of Paper: Geography and spatiality in the works of Luigi Pirandello (Tesi di dottorato non pubblicata). University of Toronto.

Alfano, G. (2010). Paesaggi, mappe, tracciati: Cinque studi su letteratura e geografia (Vol. 1). Liguori.

Anderson, J. (2024). Literary Atlas: Plotting a New Literary Geography. Routledge.

Arena, G. (2025). Avant-Texts, Characters and Factoids: Interpreting the Genesis of La luna e i falò Through an Ontology. Humanities, 14(8), 162. https://doi.org/10.3390/h14080162.

Atlas of Literary Landscapes of Mainland Portugal. Atlas Das Paisagens Literárias de Portugal Continental. https://litescape.ielt.fcsh.unl.pt. (cons. 18/04/2026).

Bartalesi, V., & Pratelli, N. (2025). Representing Geospatial Knowledge in Narratives. Journal on Computing and Cultural Heritage, 18(1), 1–15. https://doi.org/10.1145/3703918.

Bodenhamer, D. J. (2010). The Spatial Humanities: GIS and the Future of Humanities Scholarship. Indiana University Press.

Bodenhamer, D. J., Corrigan, J., & Harris, T. M. (Eds.). (2015). Deep Maps and Spatial Narratives. Indiana University Press. https://doi.org/10.2307/j.ctt1zxxzr2.

Bosco, C. L. (2025). Qualche riflessione sul rapporto tra letteratura e cartografia. In Immaginari cartografici: Riconfigurazioni teoriche dello spazio. Meltemi. https://ricerca.uniba.it/handle/11586/539922.

Bushell, S. (2020). Reading and Mapping Fiction: Spatialising the Literary Text (1a ed.). Cambridge University Press. https://doi.org/10.1017/9781108766876.

Car, N. J. (2024). Representing spatial uncertainty and allowing for probabilistic topological functions with SUFF, an extension to GeoSPARQL. GeoLD2024: 6th International Workshop on Geospatial Linked Data at ESWC 2024.

Caracciolo, M. (2014). The Experientiality of Narrative: An Enactivist Approach. DE GRUYTER. https://doi.org/10.1515/9783110365658.

Charles Booth's London. LSE Digital Library. https://booth.lse.ac.uk/map. (cons. 18/04/2026).

Collot, M. (2011). La pensée-paysage: Philosophie, arts, littérature. Actes Sud ENSP.

Cooper, D., & Gregory, I. N. (2011). Mapping the English Lake District: A literary GIS. Transactions of the Institute of British Geographers, 36(1), 89–108. https://doi.org/10.1111/j.1475-5661.2010.00405.x.

Cooper, D., Donaldson, C., & Murrieta-Flores, P. (Eds.). (2016). Literary Mapping in the Digital Age (0 ed.). Routledge. https://doi.org/10.4324/9781315592596.

Digital Literary Atlas of Ireland. http://cehresearch.org/DLAI/. (cons. 18/04/2026).

Drecki, I. (2007). Geographical information uncertainty: The concept and representational challenges. In Proceedings of the 23rd International Cartographic Conference, Moscow, Russia.

Drucker, J. (2014). Graphesis: Visual forms of knowledge production. Harvard university press.

Eide, Øyvind (2015). Media Boundaries and Conceptual Modelling Between Texts and Maps. https://doi.org/10.1057/9781137544582.

El Khatib, R., & Schaeben, M. (2020). Why map literature? Geospatial prototyping for literary studies and digital humanities. Digital Studies/Le champ numérique, 10(1).

Elli, T., Briones Rojas, M., Gobbo, B., Parigini, M., Giustetto, V., Cavalloro, V., & Mauri, M. (2021). Metaphors as Knowledge Activators in Data Visualizations: the case of the Archipelago of Calvino's literary works. In Cumulus Conference Proceedings Series (Vol. 2, pp. 906-924). Aalto University.

Fiorentino, F., Miglio, C., & Valtolina, A. (Eds.). (2025). Immaginari cartografici: Riconfigurazioni teoriche dello spazio. Meltemi.

Fisher, P. F. (1999). Models of uncertainty in spatial data. In P. Longley, M. F. Goodchild, D. J. Maguire, & D. W. Rhind (Eds.), Geographical InformationSystems: Principles, Techniques, Management and Applications: John Wiley & Sons, Vol. 1, 191–205.

Gadda, C. E. (1993). Opere di Carlo Emilio Gadda. 5,2: Bibliografia e indici / a cura di Dante Isella (Ed.). Garzanti.

Gadda, C. E. (2018). Quer pasticciaccio brutto de via Merulana (G. Pinotti, Ed.). Adelphi edizioni.

Giurdanella, L., et al. (2025). *MeMo: Una mappa letteraria digitale per la memoria del Mezzogiorno*. In S. Rebora, M. Rospocher, & S. Bazzaco (Eds.), *Diversità, equità e inclusione: Sfide e opportunità per l'informatica umanistica nell'era dell'intelligenza artificiale. Proceedings del XIV Convegno Annuale AIUCD 2025* (Quaderni di Umanistica Digitale). AIUCD. 10.6092/unibo/amsacta/8380.

Gregory, I. N., & Geddes, A. (Eds.). (2024). Toward spatial humanities: Historical GIS and spatial history. Indiana University Press.

Harley, J. B. (1989). Deconstructing the map. *Cartographica: The international journal for geographic information and geovisualization*, *26*(2), 1-20.

Hinrichs, U., Forlini, S., & Moynihan, B. (2015). Speculative practices: Utilizing infovis to explore untapped literary collections. *IEEE transactions on visualiz*ation and computer graphics, 22(1), 429-438. https://doi.org/10.1109/TVCG.2015.2467452.

Hitzler, P., Krötzsch, M., Parsia, B., Patel-Schneider, P. F., & Rudolph, S. (2012). OWL 2 Web Ontology Language Primer (2nd ed.). W3C Recommendation. World Wide Web Consortium. https://www.w3.org/TR/owl2-primer/.

Iacoli, G. (2008). La percezione narrativa dello spazio: Teorie e rappresentazioni contemporanee (1a ed). Carocci.

Jakob, M. (2005). Paesaggio e letteratura. L. S. Olschki.

Jameson, F. (1989). Il postmoderno, o, La logica culturale del tardo capitalismo. Garzanti.

Jessop, M. (2004). The visualization of spatial data in the humanities. Literary and linguistic computing, *19*(3), 335-350. https://doi.org/10.1093/llc/19.3.335.

Krautter, B. (2024). The Scales of (Computational) Literary Studies: Martin Mueller's Concept of Scalable Reading in Theory and Practice. In F. Armaselu & A. Fickers (Eds.), Zoomland: De Gruyter, 261–286. https://doi.org/10.1515/9783111317779-011.

Lebo, T., Sahoo, S., & McGuinness, D. (2013). PROV-O: The PROV Ontology. W3C Recommendation. World Wide Web Consortium. https://www.w3.org/TR/prov-o/.

Lefebvre, H. (2018). La produzione dello spazio. PGreco. [1974].

Literary Atlas – Plotting English-Language Novels in Wales. http://www.literaryatlas.wales/en/.(cons. 18/04/2026).

MacEachren, A. M., Robinson, A., Hopper, S., Gardner, S., Murray, R., Gahegan, M., & Hetzler, E. (2005). Visualizing Geospatial Information Uncertainty: What We Know and What We Need to Know. Cartography and Geographic Information Science, 32(3), 139–160. https://doi.org/10.1559/1523040054738936.

MacEachren, A. M., Roth, R. E., O'Brien, J., Li, B., Swingley, D., & Gahegan, M. (2012). Visual Semiotics & Uncertainty Visualization: An Empirical Study. IEEE Transactions on Visualization and Computer Graphics, 18(12), 2496–2505. https://doi.org/10.1109/TVCG.2012.279.

Manzoni, A. (1840). I promessi sposi. Milano: Tipografia Guglielmini e Redaelli.

Martín-Rodilla, P., & Gonzalez-Perez, C. (2018). Representing imprecise and uncertain knowledge in digital humanities: a theoretical framework and ConML implementation with a real case study.

In *Proceedings of the Sixth International Conference on Technological Ecosystems for Enhancing Multiculturality*, 863–871. https://doi.org/10.1145/3284179.3284318.

Miles, A., & Bechhofer, S. (2009). SKOS Simple Knowledge Organization System – Reference. W3C Recommendation. World Wide Web Consortium. https://www.w3.org/TR/skos-reference/.

Moretti, F. (1997). Atlante del romanzo europeo: 1800-1900. Einaudi.

Morini, F., Garretón, M., Pomerance, J., Zeissig, N., de Guenther, S., Dörk, M., Thomet, F., Freyberg, L., Kyriazis, I., & Scholz, A. (2025). Critical Interactivity: Exploration and Narration in Data Visualization. IEEE Computer Graphics and Applications. https://doi.org/10.1109/mcg.2025.3544684.

Novel City Maps. UCLab. https://uclab.fh-potsdam.de/NCM/index.en.html. (cons. 18/04/2026).

Orlando Furioso Atlas. Link: https://www.furiosoatlas.com/atlas.html. (cons. 18/04/2026).

Otten, H., Hildebrand, L., Nagel, T., Dörk, M., & Müller, B. (2018). Shifted Maps: Revealing spatio-temporal topologies in movement data. In 2018 IEEE VIS Arts Program (VISAP) (pp. 1-10). IEEE. https://doi.org/10.1109/VISAP45312.2018.9046054.

Panagiotidou, G., Lamqaddam, H., Poblome, J., Brosens, K., Verbert, K., & Vande Moere, A. (2022). Communicating uncertainty in digital humanities visualization research. IEEE Transactions on Visualization and Computer Graphics, 29(1), 635-645. https://doi.org/10.1109/tvcg.2022.3209436.

Pannach, F., Cheng, L., & Pianzola, F. (2024). The GOLEM-Knowledge Graph and Search Interface: 2024 Computational Humanities Research Conference, CHR 2024. Proceedings of the Computational Humanities Research Conference 2024, CEUR Workshop Proceedings, 3834, 462–471. https://www.scopus.com/pages/publications/85210853888.

Papotti, D., & Tomasi, F. (Eds.). (2014). La geografia del racconto: Sguardi interdisciplinari sul paesaggio urbano nella narrativa italiana contemporanea. Lang.

Piatti, B. (2008). Die Geographie der Literatur: Schauplätze, Handlungsräume, Raumphantasien. Wallstein.

Piatti, B., Bär, H. R., Reuschel, A.-K., Hurni, L., & Cartwright, W. (2009). Mapping Literature: Towards a Geography of Fiction. In Cartography and Art: Springer Berlin Heidelberg, 1–16. https://doi.org/10.1007/978-3-540-68569-2_15.

Pratelli, N. (2025). A Framework for Representing Geospatial Knowledge in Narratives, Tesi di Dottorato. https://hdl.handle.net/20.500.14242/215721.

Progetto LEDA – For an Atlas of Italian Ecological Literature: From the Great Acceleration to the Pandemic. https://progettoleda2022prin.github.io/atlante/index.html. (cons. 18/04/2026).

Reuschel, A. K., Piatti, B., & Hurni, L. (2009). Mapping literature. The prototype of "A Literary Atlas of Europe". In Proceedings of the 24th International Cartographic Conference.

Reuschel, A.-K., & Hurni, L. (2011). Mapping Literature: Visualisation of Spatial Uncertainty in Fiction. The Cartographic Journal, 48(4), 293–308. https://doi.org/10.1179/1743277411Y.0000000023.

Reuschel, A.-K., Piatti, B., & Hurni, L. (2013). Modelling Uncertain Geodata for the Literary Atlas of Europe. In K. Kriz, W. Cartwright, & M. Kinberger (Eds.), Understanding Different Geographies: Springer Berlin Heidelberg, 135–157. https://doi.org/10.1007/978-3-642-29770-0_11.

Ryan, M. L. (2003). Cognitive Maps and the Construction of Narrative Space. Herman D., (Ed.). Narrative Theory and Cognitive Sciences. Stanford, CA, CSLI Publ.

Senaratne, H., & Gerharz, L. (2011). An assessment and categorisation of quantitative uncertainty visualisation methods for geospatial data. In 14th AGILE international conference on geographic information science-advancing geoinformation science for a changing world. AGILE.

Shifted Maps: Visualizing Personal Movement through Map Networks. UCLab. https://shifted-maps.com. (cons. 18/04/2026).

Soja, E. W. (1996). Thirdspace: Journeys to Los Angeles and other real-and-imagined places. Blackwell, Cambridge-(MA)-Oxford.

Sorrentino, F. (2010). Il senso dello spazio. Lo spatial turn nei metodi e nelle teorie letterarie. Armando Editore.

Stockhammer, R. (2007). Kartierung der Erde: Macht und Lust in Karten und Literatur. W. Fink.

Stockhammer, R. (2013). Exokeanismós: The (Un) Mappability of Literature. 36(2), 123–138.

Tally, R. T. (2013). *Spatiality*. Routledge/Taylor & Francis Group.

The Alexandria of C.P. Cavafy. Onassis Foundation. https://cavafy-map.onassis.org/map. (cons. 18/04/2026).

The Atlas of Human Literatures. Global Humanities. https://comparativeglobalhumanities.mit.edu/ongoing-work/the-atlas-of-human-literatures/. (cons. 18/04/2026).

Thomson, J., Hetzler, E., MacEachren, A., Gahegan, M., & Pavel, M. (2005, March). A typology for visualizing uncertainty. In Visualization and Data Analysis 2005 (Vol. 5669, pp. 146-157). SPIE. https://doi.org/10.1117/12.587254.

von Ungern-Sternberg, A. (2009). Dots, lines, areas and words: Mapping literature and narration (With some remarks on Kate Chopin's "The Awakening"). In Cartography and art (pp. 1-24). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-68569-2_19.

Weitin, T. (2017). Scalable Reading. Zeitschrift für Literaturwissenschaft und Linguistik, 47(1), 1–6. https://doi.org/10.1007/s41244-017-0048-4.

Westphal, B. (2009). Geocritica. Reale finzione spazio. Armando Editore. [2007].

Windhager, F., Salisu, S., & Mayr, E. (2019). Exhibiting Uncertainty: Visualizing Data Quality Indicators for Cultural Collections. Informatics, 6(3), 29. https://doi.org/10.3390/informatics6030029.

Verso un grande corpus diacronico dell'italiano: il contributo degli esempi del GDLI

Eva Sassolini[1], Sebastiana Cucurullo[2], Marco Biffi[3] e Simonetta Montemagni[4]

[1]CNR Istituto di Linguistica Computazionale "A. Zampolli", Italia - eva.sassolini@ilc.cnr.it;
[2]CNR Istituto di Linguistica Computazionale "A. Zampolli", Italia - sebastiana.cucurullo@ilc.cnr.it
[3]Università degli Studi di Firenze e Accademia della Crusca, Italia - marco.biffi@unifi.it
[4]CNR Istituto di Linguistica Computazionale "A. Zampolli", Italia – simonetta.montemagni@ilc.cnr.it

ABSTRACT[1] (ITALIANO)

La disponibilità di corpora di ampie dimensioni dedicati alle varietà storiche dell'italiano costituisce tuttora una criticità. Sebbene negli ultimi anni siano state avviate diverse iniziative in questa direzione, le risorse esistenti risultano ancora limitate per estensione cronologica e dimensioni. In tale quadro si colloca il progetto ILC-DiacrITA, corpus diacronico dell'italiano sviluppato presso l'Istituto di Linguistica Computazionale "A. Zampolli" del CNR (CNR-ILC), la cui copertura dei secoli più antichi risulta tuttavia ancora poco consistente. Un'opportunità decisiva di ampliamento è offerta dal progetto di informatizzazione del *Grande Dizionario della Lingua Italiana* di Salvatore Battaglia (GDLI). Oltre alla strutturazione del dizionario, il progetto ha previsto l'estrazione sistematica degli esempi citati nelle voci lessicografiche, con l'obiettivo di trasformarli in una risorsa testuale autonoma da valorizzare (corpus GDLI). Il corpus così ottenuto rappresenta un patrimonio linguistico di eccezionale rilevanza: esso comprende oltre due milioni di passi e più di quaranta milioni di occorrenze (dato attualmente in fase di validazione), coprendo un arco cronologico di oltre un millennio. Pur caratterizzato da una prevalente matrice letteraria, il corpus GDLI presenta una notevole varietà di registri e tipologie testuali, includendo, accanto a lirica e narrativa, testi agiografici e trattatistici, documenti pratici e legislativi, nonché materiali della stampa periodica. Questo contributo presenta il lavoro in corso volto all'integrazione del corpus GDLI all'interno di ILC-DiacrITA, discutendone le potenzialità come risorsa di riferimento per lo studio computazionale e storico della lingua italiana e non solo.
Parole chiave: Corpora; Linguistica Computazionale; Cultural Heritage; Digital Humanities

ABSTRACT (ENGLISH)

Towards a large diachronic corpus of Italian: the contribution of examples from the GDLI
The availability of large-scale corpora dedicated to historical varieties of Italian remains a major challenge. Although several initiatives have been launched in recent years to address this issue, existing resources are still limited in terms of both chronological coverage and size. Within this context, the ILC-DiacrITA project—a diachronic corpus of Italian developed at the Institute for Computational Linguistics "A. Zampolli" of the CNR (CNR-ILC)—aims to provide a broad and structured resource, although its coverage of the earliest centuries is currently still limited. A decisive opportunity for expansion is offered by the digitisation project of the *Grande Dizionario della Lingua Italiana* by Salvatore Battaglia (GDLI). In addition to the structural encoding of the dictionary, the project has involved the systematic extraction of the examples cited in the lexicographic entries, with the aim of transforming them into an independent and reusable textual resource (the GDLI corpus). The resulting corpus represents a linguistic heritage of exceptional relevance: it includes over two million textual excerpts and more than forty million tokens (a figure currently under validation), spanning a chronological range of over one millennium. Despite its predominantly literary nature, the GDLI corpus exhibits a remarkable diversity of registers and text types, encompassing—alongside lyric and narrative texts—hagiographic and treatise literature, practical and legal documents, as well as materials from periodical press. This contribution presents the ongoing

[1] Nel quadro di un'elaborazione comune, Marco Biffi, come referente delle attività per l'Accademia della Crusca, ha contribuito alla redazione del § 3 insieme a Simonetta Montemagni come referente CNR-ILC. Eva Sassolini, come responsabile delle attività di realizzazione del corpus per CNR-ILC, ha prodotto i §§ 1, 4 e parte del 2; Sebastiana Cucurullo, come responsabile della gestione dei dati, ha contribuito alla redazione del § 2.

work aimed at integrating the GDLI corpus into ILC-DiacrITA and discusses its potential as a reference resource for the computational and historical study of the Italian language and beyond.
Keywords: Corpora; Computational Linguistics; Cultural Heritage; Digital Humanities

1. INTRODUZIONE

Il nostro gruppo di ricerca ha maturato negli anni una grande esperienza nel trattamento e gestione di materiali testuali di varia natura (Picchi, 2003). I progetti che hanno caratterizzato le nostre attività hanno sempre riguardato l'intero processo di elaborazione del testo, dall'acquisizione, all'annotazione sino alla preservazione a lungo termine (Sassolini et alii, 2014; Favaro et alii, 2023; Cinini et alii, 2025). In questa *pipeline* abbiamo sviluppato strumenti e procedure per ogni fase, specializzando e modificando gli approcci a seconda delle esigenze del singolo progetto. La valorizzazione di tale patrimonio testuale, obiettivo per noi primario, rimane un'operazione complessa sia per la difformità dei formati e delle annotazioni, sia per il lavoro necessario all'elaborazione dei dati testuali. Molti dei materiali digitalizzati, spesso ottenuti tramite sistemi di OCR, sono stati prodotti senza una piena consapevolezza delle implicazioni del dato digitale, sovente corredati da errori di varia natura. Anche dopo l'introduzione di standard come XML-TEI, la digitalizzazione è stata di frequente intesa come semplice trasposizione del cartaceo, senza un adeguato ripensamento del formato e della struttura del testo. Tutto ciò ha comportato una marcata eterogeneità nelle scelte di rappresentazione, talvolta determinate da esigenze contingenti. Per contro, i tentativi di attenersi a regole redazionali, soprattutto in periodi in cui mancavano riferimenti normativi condivisi, hanno prodotto formati particolari, spesso privi di un'adeguata documentazione delle scelte adottate. Questa situazione ha inciso sull'implementazione del corpus diacronico. Proprio l'aspetto diacronico, per quanto un obiettivo, era finora carente nella parte di testi in varietà storiche della lingua. Adesso con la possibilità di inserire le porzioni di testi che derivano dagli esempi del GDLI si è aperta una prospettiva reale di colmare queste lacune. Il corpus degli esempi del GDLI ha infatti dimensioni tali da consentire un aggiornamento consistente di ILC-DiacrITA e l'obiettivo di produrre una risorsa per la comunità scientifica che studia e lavora sull'italiano è concreto. Anche se il processo di costruzione è configurato come un'attività di lungo periodo e al momento abbiamo integrato solo i primi dieci volumi del GDLI, abbiamo deciso di far conoscere il progetto con l'intenzione di promuoverne la possibile l'interoperabilità con altre risorse.

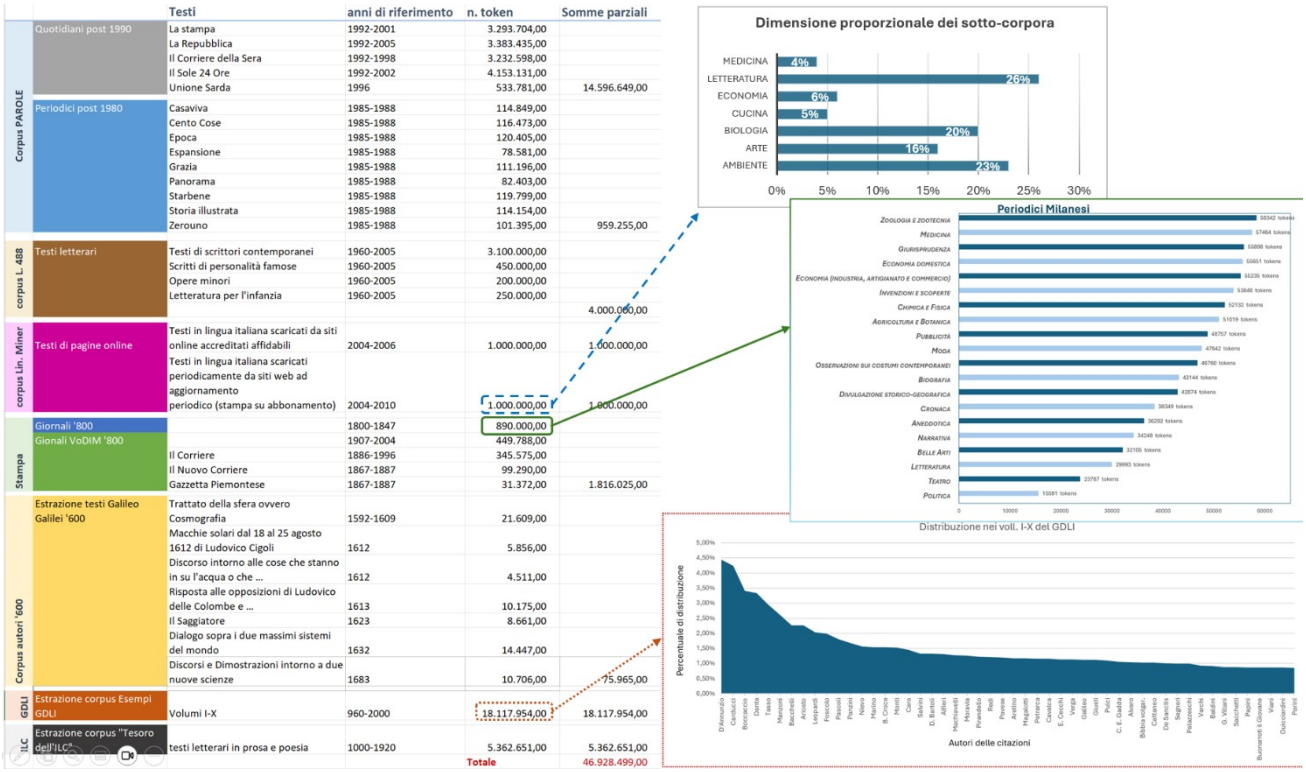
2. ILC-DIACRITA

Il corpus ILC-DiacrITA è attualmente composto da porzioni testuali[2] eterogenee per provenienza e consistenza e, per la sua natura "sedimentaria", ha progressivamente incorporato materiali testuali sia provenienti da progetti storici di CNR-ILC, sia da collaborazioni consolidate con istituzioni pubbliche e private specificamente dedicate alle varietà storiche dell'italiano (Sassi et alii, 2009; Picchi et alii, 2004).[3] Il processo di costruzione ha seguito un protocollo consolidati negli anni, che prevede procedure di omogeneizzazione / normalizzazione dei formati così come dei livelli di annotazione (Davidse & De Smet, 2021). Nel processo di migrazione ragioni di uniformità hanno obbligato a scegliere una versione dei testi privata delle annotazioni linguistiche, anche quando ciò ha significato rinunciare a preziose iniziative pionieristiche di lemmatizzazione (Onelli et alii, 2006). Tra l'altro tale scelta si è rivelata inevitabile, dal momento che alcuni testi presentavano formati ormai obsoleti e/o adottavano standard di annotazione non più in uso. Emblematico è il caso dei materiali lemmatizzati, nei quali le categorie di POS (*Part of Speech*) impiegate risultavano spesso difformi rispetto a quelle comunemente adottate nei principali sistemi di NLP, quali *Universal Dependencies*, *TreeTagger*, *spaCy* e analoghi. Stesso approccio è stato seguito per annotazioni di carattere filologico, rinunciando agli interventi e/o commenti che accompagnavano il testo per descriverne la tradizione, le varianti e le scelte editoriali. Con questi presupposti l'inserimento dei testi, sebbene condizionato da un pre-processing pressoché sistematico e impegnativo, ha consentito di comporre una grande quantità di testi affini (vd. Tab. 1). Per l'italiano contemporaneo il contributo quantitativamente più rilevante deriva dalla selezione effettuata sul corpus

[2] La decisione di non includere interi testi non è solo una caratteristica del nostro approccio ma è una pratica che svincola dalla necessità di evitare sbilanciamenti e rispettare le norme sul diritto d'autore.

[3] Una parte consistente dei dati proviene dall'implementazione dei progetti relativi al Cluster C18 – Legge 488/1999 – "Piano della Linguistica Computazionale: ricerche monolingui e multilingui". In epoca successiva anche dal progetto finanziato dal CNR come Ricerca Spontanea Tema Libero (RTL) dal titolo "Linguistic Miner".

prodotto nell'ambito del progetto PAROLE (*Preparatory Action for Linguistic Resources Organization for Language Engineering*) (Marinelli et al., 2003),[4] incentrato prevalentemente sulla lingua d'uso contemporaneo. Il progetto rappresentò una delle principali iniziative promosse dalla Comunità Europea nei primi anni Duemila per la costruzione di risorse linguistiche. Già allora era evidente come la mancanza di grandi lessici computazionali e la scarsa omogeneità delle risorse disponibili costituissero un ostacolo al progresso delle applicazioni NLP. Nel corso degli anni, ulteriori iniziative hanno contribuito all'ampliamento del corpus, quasi sempre nell'ambito di progetti o collaborazioni scientifiche. Tra queste si segnala la proficua collaborazione con l'Accademia della Crusca, che ha consentito di incrementare la parte di testi che presentano varietà storiche dell'italiano. Nell'ambito di attività e progetti comuni abbiamo potuto includere contributi selezionati all'interno del corpus del VoDim,[5] nonché del corpus de *I Periodici Milanesi*, rivelatosi di particolare interesse in quanto offre uno spaccato significativo dell'italiano preunitario non letterario (De Stefani et alii 1984). Con una dimensione prossima al milione di token, tale corpus è interamente costituito da articoli di giornale provenienti da 58 testate, pubblicate nel periodo compreso tra il 1800 e il 1847.[6]



		Testi	anni di riferimento	n. token	Somme parziali
Corpus PAROLE	Quotidiani post 1990	La stampa	1992-2001	3.293.704,00	
		La Repubblica	1992-2005	3.383.435,00	
		Il Corriere della Sera	1992-1998	3.232.598,00	
		Il Sole 24 Ore	1992-2002	4.153.131,00	
		Unione Sarda	1996	533.781,00	14.596.649,00
	Periodici post 1980	Casaviva	1985-1988	114.849,00	
		Cento Cose	1985-1988	116.473,00	
		Epoca	1985-1988	120.405,00	
		Espansione	1985-1988	78.581,00	
		Grazia	1985-1988	111.196,00	
		Panorama	1985-1988	82.403,00	
		Starbene	1985-1988	119.799,00	
		Storia illustrata	1985-1988	114.154,00	
		Zerouno	1985-1988	101.395,00	959.255,00
corpus L. 488	Testi letterari	Testi di scrittori contemporanei	1960-2005	3.100.000,00	
		Scritti di personalità famose	1960-2005	450.000,00	
		Opere minori	1960-2005	200.000,00	
		Letteratura per l'infanzia	1960-2005	250.000,00	
					4.000.000,00
corpus Lin. Miner	Testi di pagine online	Testi in lingua italiana scaricati da siti online accreditati affidabili	2004-2006	1.000.000,00	1.000.000,00
		Testi in lingua italiana scaricati periodicamente da siti web ad aggiornamento periodico (stampa su abbonamento)	2004-2010	1.000.000,00	1.000.000,00
Stampa	Giornali '800		1800-1847	890.000,00	
	Gionali VoDIM '800		1907-2004	449.788,00	
		Il Corriere	1886-1996	345.575,00	
		Il Nuovo Corriere	1867-1887	99.290,00	
		Gazzetta Piemontese	1867-1887	31.372,00	1.816.025,00
Corpus autori '600	Estrazione testi Galileo Galilei '600	Trattato della sfera ovvero Cosmografia	1592-1609	21.609,00	
		Macchie solari dal 18 al 25 agosto 1612 di Ludovico Cigoli	1612	5.856,00	
		Discorso intorno alle cose che stanno in su l'acqua o che ...	1612	4.511,00	
		Risposta alle opposizioni di Ludovico delle Colombe e ...	1613	10.175,00	
		Il Saggiatore	1623	8.661,00	
		Dialogo sopra i due massimi sistemi del mondo	1632	14.447,00	
		Discorsi e Dimostrazioni intorno a due nuove scienze	1683	10.706,00	75.965,00
GDLI	Estrazione corpus Esempi GDLI	Volumi I-X	960-2000	18.117.954,00	18.117.954,00
ILC	Estrazione corpus "Tesoro dell'ILC"	testi letterari in prosa e poesia	1000-1920	5.362.651,00	5.362.651,00
				Totale	46.928.499,00

Tabella 1: Panoramica del corpus ILC-DiacrITA con corpus GDLI (voll. I-X)

La disponibilità di materiali digitali ottocenteschi di carattere informativo risulta infatti estremamente limitata e, anche nei casi in cui tali risorse siano state conservate, l'adozione di formati spesso obsoleti ne ostacola un pieno sfruttamento. Un ulteriore incremento della porzione del corpus di testi più antichi è stato ricavato da sezioni di testi della *Biblioteca galileiana* nell'edizione Favaro,[7] nell'ambito della pluriennale collaborazione con il Museo Galileo di Firenze. Tali testi, di carattere scientifico e divulgativo,

[4] Il progetto europeo omonimo era finalizzato alla costruzione di banche dati testuali e lessicali, per le principali lingue europee, strutturate in modo uniforme. La composizione della parte in lingua italiana, conformemente al resto del corpus, comprende libri (17,91%), articoli di giornale (49.68%), periodici (24,58%) e miscellanea per il restante (7,83%). Nell'ambito del progetto fu reso disponibile un sotto-corpus di circa 3 milioni di token per la distribuzione alla comunità esterna. Si localizza temporalmente tra il 1970 e il 1997 ma ne esiste anche una versione più estesa di 4 milioni di token, costituita da ulteriori articoli di giornale che vanno dal 1997 al 2005.

[5] Il corpus è stato oggetto del progetto PRIN 2012 "Corpus di riferimento per un Nuovo Vocabolario dell'Italiano moderno e contemporaneo. Fonti documentarie, retrodatazioni, innovazioni" (Biffi, 2016).

[6] I testi furono lemmatizzati con il contributo tecnologico di CNR-ILC nei primi anni '80 in modo semi-automatico. Gli elementi del corpus coprono generi testuali diversi e i generi riguardano informazione di vari argomenti.

si collocano tra la fine del '600 e la metà del '700 e sono materiali di particolare rilevanza in quanto testimonianza di un italiano non letterario, difficilmente reperibile. Fanno parte di ILC-DiacrITA anche varie porzioni di testi che coprono un arco cronologico esteso dall'anno Mille ai primi decenni del Novecento, riunite sotto l'etichetta di *Tesoro dell'ILC:*[8] si tratta prevalentemente di testi letterari, in prosa e in poesia, riconducibili alla tradizione dei progetti legati al software DBT (Data Base Testuale).[9] Questo strumento, pioniere nell'ambito dell'analisi del testo, ha permesso di convogliare presso l'istituto gran parte delle iniziative che in passato hanno riguardato la digitalizzazione di risorse testuali. Tuttavia, alcune di queste risorse, caratterizzate da sistemi di codifica e annotazione fortemente personalizzati, sono ancora oggi in attesa di una piena restituzione alla comunità scientifica.[10] Il sistema software, caratterizzato da un'elevata granularità dell'indicizzazione e da un'ampia capacità di personalizzazione, ha da un lato supportato in modo significativo gli studi filologici sui primi formati digitali indicizzati, ma dall'altro ha risentito dei limiti tecnologici dell'epoca, che in alcuni casi hanno imposto soluzioni oggi non più recuperabili. Nella maggior parte dei casi, tuttavia, la capacità di svincolare il testo dalla sua forma sequenziale ha permesso di scomporne le unità logiche e di trasformarlo in un indice di parole. È attraverso tali indici che è stato possibile costruire un "testo dei testi", indipendente dalle fonti originarie, successivamente confluito in ILC-DiacrITA.

3. IL CORPUS DEGLI ESEMPI DEL GDLI

Un importante contributo al corpus diacronico è rappresentato oggi dall'insieme degli esempi del GDLI, unanimemente considerato il più importante dizionario storico dell'italiano[11]. Come nei dizionari che lo hanno preceduto, ogni voce è corredata da numerosi esempi selezionati per coprire il più ampio arco cronologico possibile e provenienti da una base documentaria eccezionalmente ampia: 6.226 autori e 13.848 opere (Biffi e Guadagnini, 2022). Le citazioni, ritagliate per mantenere autonomia sintattica e semantica, rendono le voci vere e proprie micro-antologie e fanno del corpus GDLI una risorsa diacronica rappresentativa per lo studio dell'italiano scritto (Favaro et al., 2022). Questa combinazione di ampiezza e cura nel ritaglio, rende l'insieme delle citazioni un patrimonio testuale coerente e sfruttabile come risorsa autonoma. Per queste ragioni, il corpus degli esempi del GDLI può essere considerato un corpus diacronico rappresentativo dell'italiano (Biffi, 2018), utile per estrapolare dati sull'uso delle parole entro i confini della tradizione letteraria documentata, anche attraverso la mediazione di precedenti dizionari (Burgassi e Guadagnini, 2017). È importante sottolineare che gli esempi sono stati estratti automaticamente da una versione digitale non standard del dizionario, ottenuta tramite sistemi di riconoscimento ottico dei caratteri (OCR), con tutte le criticità connesse (Sassolini e Biffi, 2020). Il GDLI, caratterizzato da un'impaginazione molto densa e da voci estese, rappresenta un oggetto complesso per i sistemi OCR; di conseguenza la versione digitale contiene inevitabilmente errori di diversa natura. Le sezioni più problematiche sono proprio gli esempi, dove la ridotta dimensione del carattere, la mancanza di chiari elementi strutturali di separazione e l'uso sistematico del corsivo aumentano la probabilità di errori. Queste criticità influiscono su tutte le fasi di strutturazione del dizionario, incidendo in modo significativo sulla qualità degli esempi estratti. Per mitigare il problema, sono state implementate, tramite funzionalità DBT, procedure di controllo incrociato tra esempi e testi di riferimento presenti in ILC-DiacrITA, al fine di individuare ed eliminare eventuali sovrapposizioni. Tuttavia, la dimensione e la frammentarietà del corpus rendono questa operazione particolarmente onerosa. Inoltre, anche se

[7] Si tratta di oltre 500 opere che Antonio Favaro, il massimo studioso di Galileo, ha individuato come appartenute alla biblioteca privata del grande scienziato. La raccolta contiene tutte le sue opere, ma anche quelle di scienziati contemporanei che Galileo stesso ha commentato.

[8] Da oltre quindici anni CNR-ILC ha avviato un progetto di recupero e valorizzazione di importanti risorse testuali, denominate "Tesoro dell'ILC", volto a raccogliere un grande patrimonio di testi che hanno caratterizzato la storia dei progetti implementati nella nostra linea di ricerca.

[9] Il DBT è un sistema software proprietario per l'analisi testuale, ideato da Eugenio Picchi e sviluppato presso CNR-ILC alla fine degli anni '80 del secolo scorso. Il sistema è caratterizzato da una procedura di indicizzazione molto flessibile e articolata e da una serie di funzionalità avanzate di ricerca sui testi indicizzati.

[10] Soprattutto per testi acquisiti prima del Duemila, mancando il supporto di chi aveva condotto tali annotazioni e seguito il processo di indicizzazione, non si riesce a ricostruire il formato sequenziale del testo.

[11] Nato per aggiornare il "Dizionario della lingua italiana" di Tommaseo e Bellini (1861–1879), a sua volta erede del "Vocabolario degli Accademici della Crusca", il GDLI prosegue la tradizione della lessicografia storica italiana, fondando la descrizione delle voci su un vasto apparato di citazioni d'autore in 21 volumi con oltre 23.000 pagine (GDLI, 2004).

attualmente il corpus GDLI estratto è relativo ai soli primi dieci volumi del GDLI[12], è tuttavia costituito da quasi un milione di passi, vi è quindi un'elevata probabilità che includa testi non presenti in ILC-DiacrITA. Per ridurre gli errori nel dettato sono state studiate attività di pre-processing, comprendenti strategie di correzione automatica basate su pattern manifestamente non plausibili[13] (Sassolini et al., 2021) e fasi progressive di controllo manuale, organizzate in campagne di revisione mirate a specifici fenomeni[14] (Kissos and Dershowitz, 2016; Kukich, 1992).

4. PROSPETTIVE DI UTILIZZO

Il processo d'integrazione del corpus GDLI con ILC-DiacrITA è ancora in corso e si concluderà con l'inclusione degli esempi di tutti i 21 volumi del GDLI. Le prospettive del lavoro di integrazione sono già evidenti, molteplici e differenziate per tipologia e ricadute, e riguardano attività tra loro in larga parte indipendenti. Di seguito sono sinteticamente riassunte le principali, senza tuttavia precludere in futuro ulteriori possibilità di utilizzo.

4.1 Valorizzazione dei patrimoni testuali

In una prospettiva orientata alla pubblica consultazione del corpus ILC-DiacrITA per lo studio della lingua italiana, l'ampliamento derivante dall'integrazione del consistente corpus GDLI contribuisce a colmare le carenze di testi relativi alle epoche più antiche e ad arricchire ciascun periodo storico in termini di autori e opere rappresentati. Questo filone di attività si colloca nell'ambito delle iniziative di promozione, diffusione e valorizzazione dei patrimoni testuali, in continuità con una lunga tradizione di progetti di consultazione e analisi avanzata dei testi, sviluppati nel tempo presso l'Istituto. Numerose banche dati, rese accessibili online mediante funzionalità di tipo DBT, sono state realizzate nel tempo in questo contesto e continuano tuttora a costituire uno strumento di riferimento per la comunità degli studiosi.

4.2 Corpus di addestramento

Un ulteriore scenario riguarda la possibilità di estrarre da ILC-DiacrITA un corpus destinato al fine-tuning per task di NLP. In questo caso, le attività di pre-processing assumono una natura diversa e prevedono la predisposizione di testi privi di riferimenti bibliografici, organizzati secondo archi temporali definiti. In letteratura sono già note esperienze di costruzione di corpora diacronici, ne è un esempio MIDIA[15] (*Morfologia dell'Italiano in DIAcronia*), ma anche UNITA[16] (*A Diachronic Italian Corpus based on "L'Unità"*) e DiaCORIS[17] (Onelli et alii, 2006), sebbene in questi ultimi la distribuzione temporale copra un arco più limitato: dall'epoca postunitaria al secondo dopoguerra per DiaCORIS e fino al primo decennio degli anni Duemila per UNITA (Basile et alii, 2020). Altri corpora di varietà storiche dell'italiano linguisticamente annotati, ma di dimensioni più contenute, sono *Voci della Grande Guerra* (VGG, Lenci et al., 2020), contenente testi relativi a diverse varietà (sia di generi testuali che di registri) di italiano del primo '900; e il corpus dei documenti pubblici del politico Alcide De Gasperi (Tonelli et al., 2019), un corpus multigenere che raccoglie testi tra il 1901 e il 1954. Nel caso di ILC-DiacrITA non è ancora disponibile la versione annotata linguisticamente. Al momento si è scelto di preservare le versioni già annotate dei testi che ne sono dotati, al fine di utilizzarle come riferimento per la valutazione delle annotazioni automatiche future.

[12] L'intenzione è di estendere il lavoro all'intero dizionario nelle fasi di lavoro successive di strutturazione del GDLI

[13] Ad esempio, la presenza di lettere maiuscole all'interno di parole o la sequenza di caratteri in cui la maiuscola è posizionata prima dell'apostrofo (es. deH'esemplare).

[14] Un caso emblematico riguarda la presenza di numeri in luogo di parole. Ad es.: "10" per "Io", "lo", "io"; "11" per "Il", "il", "li", ecc.

[15] Corpus di testi scritti in lingua italiana, completamente annotato con indicazione del lemma e della parte del discorso cui è ricondotta ciascuna occorrenza nei testi. Il corpus, che si estende per un arco cronologico che va dall'inizio del XIII alla prima metà del XX secolo ed è ripartito in cinque periodi temporali www.corpusmidia.unito.it (D'Achille, 2017).

[16] Il corpus copre 67 anni della storia del giornale omonimo, dal 1948 al 2014 (Basile et al., 2020).

[17] Corpus diacronico comprendente testi scritti italiani prodotti tra il 1861 e il 1945, ampliando la struttura e le possibilità di ricerca del corpus sincronico CORIS/CODIS di 100 milioni di parole.

4.3 Supporto alla correzione di testi

La possibilità di introdurre in ILC-DiacrITA errori di natura ortografica attraverso l'integrazione con il corpus del GDLI è concreta ed è stata considerata con piena consapevolezza. Per questo motivo oltre ad implementare strategie e procedure per operare campagne di correzione sui testi, abbiamo deciso di utilizzare le funzioni di indicizzazione e interrogazione DBT come ausilio all'individuazione di errori. Tale approccio consente di effettuare confronti automatici tra testi, filtrando attraverso criteri avanzati di selezione e di individuare corrispondenze attraverso la ricerca di famiglie complesse di parole co-occorrenti negli stessi contesti. Poiché una parte degli autori è condivisa tra i due corpora, le sofisticate funzionalità di DBT permettono di isolare numerosi casi di sovrapposizione anche in presenza di livelli di similarità inferiori all'80%. Considerate le dimensioni complessive dei due insiemi testuali, solo sistemi di reperimento automatico sono in grado di produrre selezioni accurate e mirate, sulle quali avviare successivi interventi di correzione manuale. Questo scenario apre dunque a una duplice ricaduta positiva: da un lato, il miglioramento della qualità del corpus degli esempi del GDLI come oggetto autonomo e indipendente; dall'altro, il perfezionamento della versione digitale del dizionario, attraverso la correzione mirata del campo degli esempi, laddove necessario.

BIBLIOGRAFIA

Basile, P., Caputo, A., Caselli, T., Cassotti, P. e Varvara, R. (2020). Overview of the EVALITA 2020 diachronic lexical semantics (DIACR-Ita) task. In Valerio Basile, Danilo Croce, Maria Di Maro e Lucia C. Passaro (eds.), Proceedings of the 7th evaluation campaign of natural language processing and speech tools for Italian (EVALITA 2020).

Biffi, M., Guadagnini, E. (2022). «Le citazioni riconducono il dizionario nell'ambito della letteratura e della vita»: un primo sguardo d'insieme sui citati del GDLI. Studi di Lessicografia Italiana. XXXIX. 351-386.

Biffi, M. (2018), Tra fiorentino aureo e fiorentino cinquecentesco. Per uno studio della lingua dei lessicografi. In La Crusca e i testi. Lessicografia, tecniche editoriali e collezionismo librario intorno al Vocabolario del 1612, a cura di Gino Belloni e Paolo Trovato. Padova. libreriauniversitaria.it edizioni. 543-560.

Biffi, M. (2016). Progettare il corpus per il vocabolario postunitario (all'interno del contributo collettaneo: Claudio Marazzini et alii, Progetto "Corpus di riferimento per un nuovo vocabolario dell'italiano moderno e contemporaneo. Fonti documentarie, retrodatazioni, innovazioni" (PRIN 2012), in L'italiano elettronico. Vocabolari, corpora, archivi testuali e sonori, Atti della "Piazza delle Lingue" dell'Accademia della Crusca. edizione 2014 (Firenze, 6-8 novembre 2014). a cura di Claudio Marazzini e Ludovica Maconi Firenze, Firenze, Accademia della Crusca. 259-80.

Burgassi, C., Guadagnini, E. (2017). La tradizione delle parole. Sondaggi di lessicologia storica. Strasbourg. ÉLiPhi.

Cinini, A., Marongiu, P., Sassolini, E. e Monachini, M. (2025). Preserving and enhancing cultural heritage:
the Digest project. in AIUCD 2025. Rebora, Simone; Rospocher, Marco; Bazzaco, Stefano (a cura di) (2025). Diversità, Equità e Inclusione: Sfide e Opportunità per l'Informatica Umanistica nell'Era dell'Intelligenza Artificiale, Proceedings del XIV Convegno Annuale AIUCD2025. Verona: AIUCD, p. 663. ISBN 978-88-942535-9-7. DOI 10.6092/unibo/amsacta/8380. In: Quaderni di Umanistica Digitale.

Davidse, K., & De Smet, H. (2021). Diachronic Corpora. In M. Paquot & S. Th. Gries (Eds.), A Practical Handbook of Corpus Linguistics (pp. 211–233). Springer Nature. https://doi.org/10.1007/978-3-030-46216-1_10

De Stefanis Ciccone, S., Bonomi, I., Masini, A. (1984). La stampa periodica milanese della prima metà dell'Ottocento: testi e concordanze. A cura di Remo Bindi e Eugenio Picchi. Orientamenti linguistici. Giardini editore. ISBN: 8842712167.

D'Achille, P. (2017). Per la storia della formazione delle parole in italiano: un nuovo corpus in rete (MIDIA) e nuove prospettive di studio. Grossmann M. (eds.). Firenze. Franco Cesati.

Favaro, M., Sassolini, E., Cinini, A., Haines, M. (2023). Un futuro per gli archivi de "Gli Anni della Cupola". in AIUCD 2023: Carbé, Emmanuela, Lo Piccolo, Gabriele, Valenti, Alessia, Stella, Francesco. La memoria digitale: forme del testo e organizzazione della conoscenza. Atti del XII Convegno

Annuale AIUCD. Siena: AIUCD, p. 424. ISBN 978-88-942535-7-3. DOI 10.6092/unibo/amsacta/7721. In: Quaderni di Umanistica Digitale.

Favaro, M., Guadagnini, E., Sassolini, E., Biffi, M., Montemagni, S. (2022). Toward the Creation of a Diachronic Corpus for Italian: a Case Study on the GDLI. in Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2022). Language Resources and Evaluation Conference (LREC 2022). Marseille, 25 June 2022. ed. R. Sprugnoli et alii, Association for Computational Linguistics (ACL). 94-100.

GDLI = Grande dizionario della lingua italiana. (2004). di Salvatore Battaglia (poi diretto da Giorgio Bàrberi Squarotti). Torino. UTET. 1961-2002, 21 voll.; con Supplemento 2004 e Supplemento 2009, diretti da Edoardo Sanguineti. Torino. UTET. 2004 e 2008. e Indice degli autori citati nei volumi I-XXI e nel Supplemento 2004, a cura di Giovanni Ronco. Torino. UTET. http://www.GDLI.it

Kissos, I., Dershowitz, N. (2016). OCR error correction using character correction and feature-based word classification. In Proceedings of the 12th IAPR International Workshop on Document Analysis Systems (DAS2016). Santorini. Greece.

Kukich, K. (1992). Techniques for automatically correcting words in text. ACM Computing Surveys (CSUR). 24(4):377-439.

Lenci, A., Montemagni, S., Boschetti, F., De Felice, I., De Rossi, F., Dell'Orletta, F., Di Giorgio, M., Miliani, M., Passaro, L. C., Puddu, A., Venturi, G., and Labanca, N. (2020), Voices of the Great War: A Richly Annotated Corpus of Italian Texts on the First World War. In Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020, 11– 16 maggio 2020), 911-918, Marseille, France, May. European Language Resource Association (ELRA).

Marinelli, R., Biagini, L., Bindi, R., Goggi, S., Monachini, M., Orsolini, P., Picchi, E., Rossi, S., Calzolari, N., Zampolli, A. (2003). The Italian PAROLE corpus: an overview. In A. Zampolli, N. Calzolari, L. Cignoni, (eds.), Computational Linguistics in Pisa - Linguistica Computazionale a Pisa. Linguistica Computazionale, Special Issue, XVI-XVII, (2003). Pisa-Roma, IEPI. Tomo I. 401-421.

Onelli, C., Proietti, D., Seidenari, C., & Tamburini, F. (2006). The DiaCORIS project: a diachronic corpus of written Italian. In N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odijk & D. Tapias (Eds.), Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06) (pp. 1371–1374). European Language Resources Association (ELRA).

Picchi, E., Ceccotti, M. L., Cucurullo, S., Sassi, M., Sassolini, E. (2004). Linguistic Miner: An Italian Linguistic Knowledge System. Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04). Lisbon, Portugal. European Language Resources Association (ELRA). Http://www.lrec-conf.org/proceedings/lrec2004/pdf/726.pdf.

Picchi, E. (2003). PiSystem: sistemi integrati per l'analisi testuale. In A. Zampolli, N. Calzolari, L. Cignoni, (eds.), Computational Linguistics in Pisa - Linguistica Computazionale a Pisa. Linguistica Computazionale, Special Issue, XVIII-XIX, (2003). Pisa-Roma. IEPI. Tomo II. 597-627.

Sassi, M. Cucurullo, S., Picchi, P. (2006). Valorizzazione del patrimonio letterario della lingua italiana.Il corpus italiano e la legge 488. Cluster C18 - Legge 488: Project Deliverable. http://pumax.isti.cnr.it/linkdoc.php?authority=cnr.ilc&collection=cnr.ilc&langver=it&idcode=2006-B4-003.

Sassolini, E., Cucurullo, S., Biffi, M. (2024) L'informatizzazione del GDLI: risultati, prospettive, sfide future, in AIUCD 2024: Di Silvestro, Antonio ; Spampinato, Daria (a.c.d) (2024). Me.Te. Digitali. Mediterraneo in rete tra testi e contesti. Proceedings. Catania. 281–286. DOI: 10.6092/unibo/amsacta/7927.

Sassolini, E., Biffi, M., De Blasi, F., Guadagnini, E. e Montemagni, S. (2021). La digitalizzazione del GDLI: un approccio linguistico per la corretta acquisizione del testo?, in AIUCD 2021: DH per la società: e-guaglianza, partecipazione, diritti e valori nell'era digitale. Raccolta degli abstract estesi della 10a conferenza nazionale. Pisa. a cura di A.M. Del Grosso, F. Boschetti ed E. Salvatori. 159-166. https://aiucd2021.labcd.unipi.it/book-of-abstracts/

Sassolini, E., Biffi, M. (2020). Strategie e metodi per il recupero di dizionari storici. Conferenza AIUCD 2020, 15-17 gennaio 2020, Università Cattolica del Sacro Cuore. Milano. ISBN 978-88-942535-4-2. DOI 10.6092/unibo/amsacta/6316. In: Quaderni di Umanistica Digitale: 235-239

Sassolini, E., Cucurullo, S., Sassi, M. (2014). Methods of textual archive preservation. In: CLiC-it 2014 – Prima Conferenza Italiana di Linguistica Computazionale (Università di Pisa e CNR, Pisa, Italia, 9-10 dicembre 2014). Proceedings, vol. I, ISBN 978-886741-472-7, 334–338. Pisa University Press. Pisa.

Mondiale: A Multi-Genre 20th Century Corpus in the Political Domain". In Proceedings of CLiC-it 2019.

Verso una classificazione funzionale dei prompt basata sulle intenzioni cognitive esplicitate dagli studenti nell'interazione con l'IA

Daniele Scala[1], Salvatore Varriale
[1]Università degli Studi di Genova, Dipartimento di Lingue e Culture Moderne (DLCM), Italy-DANIELE.SCALA@EDU.UNIGE.IT
[2]Università degli Studi di Genova, Dipartimento di Lingue e Culture Moderne (DLCM), Italy-salvatore.varriale@edu.unige.it

ABSTRACT (ITALIANO)

Il contributo propone una cornice teorico-metodologica per interpretare il prompt come formulazione linguistica che rende osservabile la funzione della richiesta nell'interazione con sistemi di intelligenza artificiale generativa. Superando sia una lettura esclusivamente tecnica del prompt engineering sia i tentativi di classificazione basati sulla tassonomia di Bloom, lo studio introduce una prospettiva funzionale incentrata su ciò che lo studente fa attraverso il prompt nel corso dell'interazione. In questa prospettiva, i prompt vengono analizzati non in base al livello cognitivo del compito, ma in relazione alla funzione che la richiesta svolge nella conversazione (ad esempio chiarire il task, esplorare idee, richiedere una valutazione o affinare progressivamente un output). Ne deriva una classificazione dei prompt nelle interazioni studente-IA basata su categorie funzionali, utile per analizzare come l'IA generativa venga utilizzata nei processi di apprendimento in ambito educativo. Il contributo presenta inoltre il disegno di una successiva fase di sperimentazione empirica finalizzata a mettere alla prova l'impianto proposto.
Parole chiave: prompt engineering; intelligenza artificiale generativa; interazione studente-IA; classificazione funzionale dei prompt; apprendimento mediato dall'IA

ABSTRACT (ENGLISH)

Towards a functional classification of prompts based on the cognitive intentions expressed by students in their interactions with AI
The contribution proposes a theoretical and methodological framework to interpret the prompt as a linguistic formulation that makes the function of a request observable within interactions with generative artificial intelligence systems. Moving beyond both a purely technical view of prompt engineering and classification attempts based on Bloom's taxonomy, the study introduces a functional perspective focused on what the student does through the prompt during the interaction. Within this perspective, prompts are analyzed not according to the cognitive level of the task, but in relation to the function that the request performs within the conversation (e.g., clarifying the task, exploring ideas, requesting evaluation, or progressively refining an output). This leads to a classification of prompts in student–AI interactions based on functional categories, which can be used to analyze how generative AI is employed in learning processes in educational contexts. The contribution also outlines the design of a subsequent empirical phase aimed at testing the proposed framework.
Keywords: prompt engineering; generative artificial intelligence; student–AI interaction; functional classification of prompts; AI-mediated learning

1. INTRODUZIONE

L'integrazione di modelli di Intelligenza Artificiale (IA) generativa nei contesti educativi sta rapidamente ridefinendo le modalità di insegnamento, valutazione e apprendimento, offrendo altresì nuove possibilità agli studenti nello studio autonomo (Lee & Palmer, 2025). In particolare, i modelli linguistici di grandi dimensioni (LLM) alla base dei chatbot conversazionali più comuni sono oggetto di crescente interesse per il loro potenziale nel supportare attività di scrittura, ricerca e in generale attività che richiedono problem solving (Barrett & Pack, 2023; Memmert & Bittner, 2022). Diventa quindi sempre più importante analizzare le implicazioni didattiche, pedagogiche e cognitive che derivano dal loro utilizzo all'interno dei processi educativi (Celik, 2023; Lee & Palmer, 2025). Per farlo, sono richiesti strumenti analitici capaci di

descrivere non solo gli output prodotti, ma anche le modalità attraverso cui tali output vengono sollecitati.
Una delle principali linee di ricerca riguarda il ruolo del prompt engineering nel favorire interazioni più efficaci tra studenti e sistemi di IA generativa. Con prompt *engineering* si intende la progettazione e l'ottimizzazione dei prompt di input per guidare le risposte dei LLM e migliorare accuratezza, pertinenza e coerenza dell'output. L'obiettivo, quindi, è individuare le strategie di prompting più efficaci attraverso l'analisi delle interazioni con l'IA, al fine di comprendere meglio i meccanismi di funzionamento di tali sistemi (Linardatos et al., 2021). Di conseguenza, il prompt può essere visto come strumento di mediazione di questa interazione, capace di orientare in modo intenzionale e strategico le risposte dei LLM (Mondal et al., 2024; Mezzadri & Paita, 2025).
Buona parte della letteratura attuale si concentra su aspetti strutturali e procedurali del prompting (ad esempio l'utilizzo di tecniche quali *zero-shot*, *few-shot*, *chain-of-thought* e *tree-of-thought*) per migliorare l'accuratezza delle risposte generate dai modelli (Choi & Chang, 2025), lasciando meno esplorata la dimensione interpretativa secondo cui il prompt formulato da uno studente può essere analizzato come traccia osservabile della funzione che la richiesta assume nel corso dell'interazione con l'IA.
Infatti, la formulazione di un prompt non è soltanto un atto tecnico, ma implica scelte esplicite rispetto al tipo di supporto, risposta o intervento che lo studente richiede al sistema di IA nello svolgimento del compito (Choi & Chang, 2025). In questa prospettiva, la scelta linguistica può essere interpretata come traccia di un'intenzione cognitiva, intesa come la funzione che il prompt svolge nel corso dell'interazione con l'IA in relazione allo svolgimento del compito. Interpretare il prompt in questi termini consente di assumerlo come un'unità analitica osservabile. Si propone quindi una classificazione dei prompt basata sulle diverse funzioni che le richieste possono assumere nell'interazione con l'IA, ad esempio chiarire, esplorare, valutare, affinare o generare elementi utili allo svolgimento del compito.
In una prima fase, tale prospettiva ha trovato un riferimento utile nella tassonomia di Bloom, nella versione rivisitata da Anderson e Krathwohl (2001), in quanto ha consentito di considerare il prompt non soltanto come input tecnico, ma anche come formulazione interpretabile in relazione al tipo di operazione richiesta. Tuttavia, proprio l'analisi delle interazioni studente-IA mostra la necessità di superare una classificazione fondata sui livelli cognitivi, orientandosi verso un modello più adeguato a descrivere la funzione che i prompt svolgono nelle diverse fasi dell'interazione.
L'obiettivo è duplice: da un lato, offrire una classificazione dei prompt utilizzabile come strumento analitico per l'osservazione delle pratiche di apprendimento che coinvolgono l'IA generativa; dall'altro, delineare una base concettuale per ricerche future volte a indagare le connessioni tra le modalità con cui gli studenti orientano le proprie richieste all'IA, lo sviluppo del pensiero critico e le forme di agency nell'interazione umano-IA.
La domanda di ricerca alla base del contributo è la seguente: quali intenzioni cognitive tendono a emergere nei prompt formulati dagli studenti nell'interazione con l'Intelligenza Artificiale generativa?

2. QUADRO TEORICO

2.1 DEFINIZIONE DI PROMPT

Poiché la qualità degli output dei Large Language Model non dipende esclusivamente dalla potenza del modello o dai dati di addestramento, la letteratura recente sottolinea il ruolo centrale delle modalità di formalizzazione delle richieste per un uso efficace di tali strumenti (White et al., 2023). In questo quadro si colloca il prompt engineering, inteso come progettazione e ottimizzazione sistematica dei prompt, finalizzata a orientare le risposte dei LLM verso output accurati, pertinenti e coerenti (Chen et al., 2023). Il prompting richiede una formulazione chiara e non ambigua, basata su scelte consapevoli di lessico e struttura. A differenza della comunicazione umana, nell'interazione con l'IA risultano prioritari chiarezza e concisione (Dornburg & Davin, 2024; Giray, 2023). In questa prospettiva, è consigliabile l'uso di verbi d'azione diretti (es. scrivi, classifica, riassumi) e di formulazioni affermative (Badino, D'Asaro, Pedrazzoli, 2024). Tra le strategie più efficaci rientrano anche l'assegnazione di un ruolo al modello e la definizione esplicita del contesto (livello, destinatari, obiettivi), che favoriscono un maggiore allineamento tra richiesta e output. Tuttavia, la natura probabilistica dei LLM rende spesso necessario un affinamento progressivo della richiesta.

2.2 COME VENGONO CLASSIFICATI I PROMPT OGGI: L'APPROCCIO STRUTTURALE

Dopo aver delineato i principi fondamentali del prompt engineering, è utile esaminare le principali tipologie di prompt da un punto di vista procedurale e strutturale, in quanto un prompt può assumere forme diverse in funzione della complessità del compito e del grado di controllo desiderato sull'output (Prompt Engineering Guide, 2026)[1]. Lo *zero-shot* prompting, ad esempio, consiste in una richiesta diretta, priva di esempi guida forniti durante l'addestramento, ed è generalmente utilizzato per compiti generali e ben definiti, come nel caso di una domanda del tipo "In che anno è scoppiata la rivoluzione francese?". Il *one-shot* e il *few-shot* prompting si collocano su un livello di maggiore elaborazione e prevedono l'inserimento di uno o più esempi all'interno del prompt, con la funzione di orientare il modello rispetto al formato o al tipo di risposta attesa, aumentando la prevedibilità e la coerenza dell'output. Una modalità distinta è rappresentata dal *chain-of-thought* prompting (Wei *et al.*, 2022), che invita il modello a esplicitare i passaggi intermedi del ragionamento che conducono alla soluzione di un problema. Un esempio frequentemente citato riguarda i problemi aritmetici: a fronte di una domanda come "Ho tre mele. Ne compro altre due. Quante mele ho?", il modello esplicita il ragionamento passo dopo passo ("All'inizio ho tre mele. Poi ne compro altre due. Tre più due fa cinque."), rendendo il processo decisionale più trasparente e riducendo il rischio di errori. Il *tree-of-thoughts* (Yao *et al.*, 2023; Long, 2023) costituisce un'ulteriore evoluzione di questo approccio, poiché il processo di generazione procede attraverso più percorsi di ragionamento alternativi, che vengono progressivamente valutati e selezionati. Nella formulazione proposta da Hulbert (2023), il modello simula più "esperti" che sviluppano soluzioni parallele, interrompendo i percorsi non ritenuti.

Poiché l'interazione con i modelli generativi si configura generalmente come un processo iterativo di prova e revisione, risulta spesso necessario ricorrere a una sequenza di follow-up per affinare progressivamente l'output iniziale, introducendo nuovi vincoli o precisazioni. Questa dinamica può essere letta alla luce dei modelli di collaborazione uomo-IA, in cui l'IA genera contenuti mentre il contributo umano resta centrale per orientarli, valutarli e adattarli agli obiettivi del compito (Ifenthaler & Schumacher, 2023; Ren et al., 2024). In questo quadro, il ruolo dell'utente non si limita alla formulazione del prompt iniziale, ma consiste nel guidare l'interazione nel tempo, attraverso più cicli iterativi, per allineare progressivamente gli output alle proprie esigenze. L'interazione con l'IA diventa così un processo di costruzione progressiva dell'output, in cui il controllo resta nelle mani dell'utente.

Nel presente paragrafo sono state descritte diverse tipologie di prompt da un punto di vista strutturale e procedurale; tuttavia, più recentemente si stanno diffondendo anche approcci di tipo tassonomico che mettono i prompt in relazione con livelli cognitivi predefiniti, ad esempio attraverso l'associazione tra verbi direttivi e categorie della tassonomia di Bloom rivisitata da Anderson e Krathwohl (2001).

2.3 TENDENZE EMERGENTI NELLA CLASSIFICAZIONE DEI PROMPT: DALL'APPROCCIO STRUTTURALE A QUELLO COGNITIVO-PEDAGOGICO

Accanto all'approccio strutturale, si stanno affermando tendenze che propongono uno spostamento da una classificazione prevalentemente tecnica dei prompt verso prospettive di tipo cognitivo e pedagogico. In queste prospettive, i prompt sono considerati come formulazioni linguistiche che riflettono processi cognitivi e obiettivi di apprendimento (Structural Learning, n.d.). In tale quadro, la tassonomia di Bloom offre un riferimento teorico particolarmente rilevante, in quanto consente di interpretare le richieste formulate dagli studenti in relazione ai diversi livelli di complessità cognitiva che esse implicano. Alcuni studi recenti, in questa direzione, hanno iniziato a mettere in relazione la formulazione linguistica dei prompt con specifici obiettivi cognitivi, proponendo associazioni sistematiche tra le richieste rivolte all'IA e le categorie della tassonomia.

Originariamente proposta da Bloom e colleghi nel contesto della didattica e della psicologia dell'educazione (Bloom, 1956), la tassonomia rappresenta un sistema di classificazione degli obiettivi educativi volto a descrivere una gerarchia di processi cognitivi, dalla semplice memorizzazione di informazioni fino alla capacità di giudizio critico e di creazione di nuovi significati (Krathwohl, 2002). In altre parole, essa classifica gli obiettivi di apprendimento in base a livelli crescenti di complessità nel dominio cognitivo (Adams, 2015). La versione originale della tassonomia suddivide il dominio cognitivo in sei categorie principali: conoscenza, comprensione, applicazione, analisi, sintesi e valutazione. Anderson e Krathwohl (2001) hanno poi proposto una versione che ha ridefinito i livelli attraverso verbi

[1] https://www.promptingguide.ai/introduction/elements (cons. 23/01/2026)

d'azione piuttosto che nomi (ricordare, comprendere, applicare, analizzare, valutare, creare), riflettendo più esplicitamente la modalità in cui essi vengono utilizzati e facilitandone quindi la comprensione in termini di azioni cognitive (Krathwohl, 2002).
Tuttavia, nel contesto dell'interazione studente-IA, una classificazione basata sui livelli tassonomici presenta alcuni limiti. In primo luogo, la tassonomia di Bloom nasce per descrivere obiettivi e processi di apprendimento del soggetto, mentre il prompt rivolto a un chatbot esprime una richiesta indirizzata a un sistema esterno, all'interno di un processo conversazionale e iterativo, caratterizzato da successive riformulazioni, specificazioni e affinamenti della richiesta (Mondal et al., 2024). In secondo luogo, un inquadramento centrato sul rapporto tra verbi e livelli cognitivi rischia di ridurre l'analisi a una corrispondenza lessicale solo apparentemente lineare: nei contesti mediati dall'IA, infatti, un singolo prompt può riguardare simultaneamente più livelli cognitivi, rendendo problematica una corrispondenza univoca tra formulazione linguistica e categoria tassonomica (Jackson, 2025).
Considerando questi limiti, il presente contributo assume che, nell'interazione studente-IA, i prompt siano più adeguatamente descrivibili non in base a una gerarchia di livelli cognitivi, ma in relazione al ruolo che svolgono nella conversazione ai fini del compito. In questa prospettiva, la sezione seguente introduce una classificazione funzionale dei prompt, volta a rendere analizzabili le diverse modalità con cui gli studenti utilizzano l'IA per l'apprendimento.

3. OPERAZIONALIZZAZIONE DELLA CLASSIFICAZIONE FUNZIONALE DEI PROMPT

Alla luce di quanto discusso nella sezione precedente, il presente contributo definisce con il termine "intenzione cognitiva" non una generica azione mentale inferita a partire dal testo, bensì la funzione che il prompt assume nel corso dell'interazione con l'IA ai fini dello svolgimento del compito. L'attenzione si concentra quindi sul modo in cui lo studente pone la richiesta al sistema per chiarire, esplorare, valutare, affinare o generare elementi utili al suo avanzamento. Questa definizione si può anche collocare in continuità con approcci che concepiscono il prompt engineering come una competenza legata all'uso dei LLM orientato a obiettivi specifici da parte di utenti non esperti (Knoth et al., 2024).
Partendo da questa ridefinizione, i singoli prompt possono essere classificati in base al ruolo che svolgono nel supportare lo svolgimento del compito. La loro interpretazione, tuttavia, avviene nel contesto delle sequenze interazionali, così da collocarne la funzione all'interno dello sviluppo della conversazione.
La classificazione proposta intende distinguere tra richieste di chiarimento del compito o dei suoi elementi, richieste orientate all'esplorazione di idee o possibilità, richieste di valutazione o verifica, richieste di affinamento progressivo di risultati della conversazione e richieste di generazione di un prodotto relativamente compiuto. Nel dettaglio, le cinque categorie funzionali del modello proposto sono le seguenti:

I. **Task clarification**: comprende i prompt il cui scopo principale è ridurre ambiguità o incertezza rispetto a un compito, un suo passaggio, un criterio, una consegna, un termine, un concetto o un output precedente della conversazione. In altre parole, nei prompt di Task clarification lo studente intende comprendere meglio che cosa fare, che cosa significa qualcosa, com'è formulato un problema (o un sottoproblema) o come interpretare un'informazione. Rientrano in questa categoria quindi: richieste di spiegazione, parafrasi, disambiguazione, esplicitazione di criteri, interpretazione di una consegna, comprensione di un termine, richieste di approfondimento per ulteriore chiarezza, chiarimenti circa una precedente risposta dell'IA.
II. **Exploration**: comprende i prompt orientati ad aprire il campo delle possibilità a disposizione, generare idee, alternative, piani di lavoro, prospettive diverse, strategie per proseguire la risoluzione del task. Lo studente, quindi, sfrutta prompt di Exploration quando necessita stimoli circa la ricerca delle possibilità utili per orientarsi nel task o in un suo sottoproblema. Rientrano in questa categoria: richieste di brainstorming, richieste di alternative (data una o più opzioni), ricerca di idee di partenza, individuazione di possibili percorsi risolutivi, suggerimenti preliminari o concetti chiave da cui partire.
III. **Evaluation**: comprende i prompt che hanno lo scopo di verificare, controllare, giudicare o confrontare qualità, correttezza, coerenza, adeguatezza o affidabilità di un output dell'IA, un'informazione, un elaborato dello studente o una serie di alternative formulate. Lo studente, quindi, si avvale di un prompt di Evaluation quando intende sottoporre a giudizio o verifica qualcosa che è già disponibile, includendo potenzialmente parametri o criteri (che possono essere

anche impliciti).Rientrano in questa categoria attività di fact-checking, controllo qualità, verifica di correttezza, richiesta di feedback argomentato, confronto tra opzioni, analisi di punti di forza e debolezza e controllo della coerenza.

IV. **Refinement**: comprende i prompt finalizzati a modificare, integrare, adattare o migliorare progressivamente un elemento già presente nella conversazione, come un prompt precedente, un output dell'IA o una direzione di lavoro già avviata. Si tratta quindi di richieste che intervengono su contenuti esistenti per rielaborarli, perfezionarli o aggiornarli in modo progressivo. Rientrano in questa categoria i prompt di follow-up, gli adattamenti progressivi, le aggiunte o le rimozioni di vincoli, i cambi di tono o formato, le integrazioni di istruzioni aggiuntive, gli affinamenti di prompt precedenti o output ritenuti insoddisfacenti.
V. **Item generation**: comprende i prompt orientati alla produzione di un output o artefatto che sia relativamente compiuto e di conseguenza potenzialmente utilizzabile in modo diretto rispetto a un task o un sottoproblema. In questa categoria lo studente richiede all'IA la generazione di un elemento nuovo, a partire da zero o quasi, sotto forma di testo, tabella, immagine, schema, bozza strutturata o altra tipologia di prodotto. L'appartenenza a questa categoria non implica assenza di controllo cognitivo da parte dello studente, che può successivamente proseguire l'interazione attraverso prompt di evaluation o refinement.

Di seguito si riporta un esempio concreto per ciascuna tipologia di prompt (cfr. Tabella 1)*:*

Categoria	**Esempio**	**Funzione specifica**
Task clarification	*Quando dici che Porto è una meta "comoda da girare a piedi", cosa intendi esattamente?*	Lo studente chiede di interpretare un output precedente dell'IA per comprenderlo meglio.
Exploration	*Quali approcci potrei usare per spiegare la fotosintesi a una classe poco motivata?*	Si richiede l'individuazione di possibili strategie.
Evaluation	*Secondo te, questo piano giornaliero è realistico o ci sono troppe attività nella stessa giornata?*	Il chatbot è chiamato a valutare l'adeguatezza e la fattibilità della proposta.
Refinement	*Ho creato questo itinerario per una scuola secondaria di secondo grado. Riadattalo per la scuola primaria, tenendo conto dell'età degli alunni.*	Il contenuto di base è già disponibile; si richiede un riadattamento.
Item generation	*Crea una rubrica di valutazione a quattro livelli per la correzione di testi informativi in inglese di livello A2. Tieni conto del Quadro Comune Europeo di Riferimento per le Lingue (QCER)*	Si richiede la produzione di un artefatto completo.

Tabella 1. Esempi di prompt per le cinque categorie funzionali

Le categorie non sono da intendersi come rigidamente mutuamente esclusive. Nei casi in cui un prompt presenti in modo esplicito più funzioni riconoscibili, esso può essere ricondotto a più categorie contemporaneamente, così da preservare la complessità funzionale della richiesta.
Nel loro insieme, le cinque categorie permettono di osservare e classificare i prompt in base al ruolo che assumono nel corso dell'interazione, superando sia una lettura strutturale sia una tassonomica. La classificazione proposta si configura quindi come uno strumento analitico utile per comprendere come gli

studenti utilizzano l'IA generativa nello svolgimento di un compito e quali forme di supporto richiedono nei diversi momenti della conversazione.

4. METODOLOGIA

Il contributo propone il design di un sistema di categorie funzionali per l'analisi dei prompt nell'interazione studente-IA. L'approccio è esplorativo, prevalentemente qualitativo, con una componente quantitativa limitata all'analisi delle frequenze. L'obiettivo è mappare i tipi di prompt all'interno dell'intero log conversazionale, per individuare quali richieste emergono con maggiore ricorrenza e descriverne le caratteristiche.
La fase di sperimentazione, prevista in un momento successivo, si basa su una breve attività in classe con un gruppo di studenti (6–7 partecipanti), impegnati in un task aperto ma vincolato, ad esempio la progettazione di un itinerario di viaggio sulla base di indicazioni relative a budget, durata e preferenze. L'attività si svolge senza una formazione preliminare specifica sull'uso del sistema, per osservare pratiche spontanee di interazione.
I dati saranno raccolti tramite trascrizione dell'intero log conversazionale. Il prompt costituisce l'unità minima di analisi ed è considerato sempre nel contesto della conversazione, evitando interpretazioni isolate. A uno stesso prompt possono essere associate più categorie analitiche (task clarification, exploration, evaluation, refinement, item generation), intese come etichette descrittive e non come classi rigide.
In questa fase, l'analisi si limita alla mappatura e descrizione dei prompt; un'analisi più approfondita delle dinamiche interazionali sarà oggetto di sviluppi successivi.

5. CONCLUSIONI

Il contributo propone una cornice teorico-metodologica per analizzare l'interazione con sistemi di intelligenza artificiale generativa in termini di intenzioni cognitive osservabili, superando sia una prospettiva centrata sugli aspetti tecnici del prompting, sia i tentativi di applicare in modo diretto alla formulazione dei prompt framework come la tassonomia di Bloom nella versione rivisitata di Anderson e Krathwohl (2001). In particolare, tali approcci tendono a leggere la formulazione linguistica della richiesta come indice relativamente diretto dell'operazione cognitiva coinvolta. Nei contesti mediati dall'IA, tuttavia, la possibilità di delegare la produzione al sistema rende meno trasparente la relazione tra formulazione del prompt e processo cognitivo sottostante. Di conseguenza, questa associazione può risultare poco fondata.
L'approccio adottato sposta quindi l'attenzione dal prompt considerato isolatamente al prompt interpretato nel contesto di sequenze interazionali, analizzate attraverso categorie funzionali quali task clarification, exploration, evaluation, refinement e item generation. Questo consente di descrivere in modo preciso come gli studenti orientano l'interazione con il sistema nel corso dello svolgimento di un compito e quali forme di supporto richiedono nei diversi momenti della conversazione. In questa prospettiva, la classificazione proposta offre uno strumento analitico utile per descrivere e confrontare le modalità d'uso dell'IA in contesti educativi.
Nel loro insieme, la ridefinizione del concetto di intenzione cognitiva, la proposta di una classificazione funzionale dei prompt e la relativa impostazione metodologica costituiscono il nucleo del contributo presentato. L'impianto descritto sarà oggetto di una successiva fase di sperimentazione empirica, finalizzata a verificarne la tenuta analitica in contesti disciplinari differenti e su campioni più ampi. Ciò consentirà anche di esplorare ulteriori dimensioni dell'interazione studente-IA, come le forme di agency, le dinamiche di controllo e delega e le possibili relazioni tra configurazioni del prompting e processi di apprendimento.

RINGRAZIAMENTI

Il contributo è frutto anche delle attività svolte nell'ambito del corso di dottorato in Digital Humanities dell'Università di Genova. Desideriamo pertanto ringraziare i nostri colleghi, i nostri supervisori e la coordinatrice. Un sentito ringraziamento va anche alle nostre famiglie.

BIBLIOGRAFIA

Adams N. E. (2015). Bloom's taxonomy of cognitive learning objectives. Journal of the Medical Library Association: JMLA, 103(3), 152-153.

Anderson, L. W., & Krathwohl, D. R. (2001). A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives: complete edition. Addison Wesley Longman, Inc.

Badino, M., D'Asaro, F. A., & Pedrazzoli, F. (2024). Educare all'IA. La sfida didattica dell'intelligenza artificiale: ChatGPT e Gemini. Milano: Sanoma.

Barrett, A., & Pack, A. (2023). Not quite eye to A.I.: Student and teacher perspectives on the use of generative artificial intelligence in the writing process. International Journal of Educational Technology in Higher Education, 20. https://doi.org/10.1186/s41239-023-00427-0

Bloom, B. S. (1956). *Taxonomy of educational objectives: The classification of educational goals (Vol. 1).* Longmans, Green.

Celik, I. (2023). Towards Intelligent-TPACK: An empirical study on teachers' professional knowledge to ethically integrate artificial intelligence (AI)-based tools into education. Computers in Human Behavior, 138.

Chen, B., Zhang, Z., Langrené, N., & Zhu, S. (2023). Unleashing the potential of prompt engineering in large language models: A comprehensive review. arXiv preprint arXiv:2310.14735.

Choi, W. C., & Chang, C. I. (2025). A survey of techniques, key components, strategies, challenges, and student perspectives on prompt engineering for large language models (LLMs) in education. Preprints. https://doi.org/10.20944/preprints202503.1808.v1

Dornburg, A., & Davin, K. J. (2024). ChatGPT in foreign language lesson plan creation: Trends, variability, and historical biases. ReCALL, 1-16.

Giray, L. (2023). Prompt engineering with ChatGPT: A guide for academic writers. Annals of Biomedical Engineering, 51(12), 2629-2633.

Hulbert, D. (2023). dave1010/tree-of-thought-prompting: First version (v0.1). Zenodo. https://doi.org/10.5281/zenodo.10323452

Ifenthaler, D., & Schumacher, C. (2023). Artificial intelligence and human intelligence in education. Journal of Research on Technology in Education, 55, 1-6. https://doi.org/10.1080/15391523.2022.2154511

Jackson, J. (2025). *Higher order prompting: Applying Bloom's revised taxonomy to the use of large language models in higher education*. Studies in Technology Enhanced Learning, 4(1). https://doi.org/10.21428/8c225f6e.0915c17e

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. Theory into practice, 41(4), 212-218.

Knoth, N., Tolzin, A., Janson, A., & Leimeister, J. M. (2024). *AI literacy and its implications for prompt engineering strategies*. *Computers and Education: Artificial Intelligence, 6*, 100225. https://doi.org/10.1016/j.caeai.2024.100225

Lee, D., & Palmer, E. (2025). Prompt engineering in higher education: A systematic review to help inform curricula. International Journal of Educational Technology in Higher Education, 22(7). https://doi.org/10.1186/s41239-025-00503-7

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2021). Explainable AI: A review of machine learning interpretability methods. Entropy, 23(1), 18. https://doi.org/10.3390/e23010018

Long, J. (2023). Large language model guided tree-of-thought. arXiv preprint arXiv:2305.08291.

Mezzadri, M., & Paita, M. (2025). Integrare il prompt engineering tra le abilità trasversali per l'educazione linguistica: Uno studio di caso. Italiano LinguaDue, 17(2), 614-638.

Memmert, L., & Bittner, E. (2022). Complex problem solving through human-AI collaboration: A literature review.

Mondal, S., Bappon, S. D., & Roy, C. K. (2024). Enhancing user interaction in ChatGPT: Characterizing and consolidating multiple prompts for issue resolution. In Proceedings of the 21st International Conference on Mining Software Repositories (pp. 222-226).

Prompt Engineering Guide. (2026). Elements of a prompt. https://www.promptingguide.ai/introduction/elements

Ren, C., Pardos, Z., & Li, Z. (2024). Human-AI collaboration increases skill tagging speed but degrades accuracy. arXiv preprint arXiv:2403.02259.

Structural Learning. (n.d.). *AI prompts for every level of Bloom's taxonomy: A teacher's guide*. Structural Learning. https://www.structural-learning.com/post/ai-prompts-blooms-taxonomy-teachers-guide

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). Tree of thoughts: Deliberate problem solving with large language models. In Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023).

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., et al. (2023). A prompt pattern catalog to enhance prompt engineering with ChatGPT. arXiv preprint arXiv:2302.11382.

Il videogioco come strumento di mediazione interculturale

ALESSIA STOCCO
UNIVERSITÀ DI PISA, ITALY - ALESSIASTOCCO01@GMAIL.COM

ABSTRACT (ITALIANO)

Il presente elaborato indaga come il videogioco possa diventare uno strumento propedeutico al Diritto Interculturale, ponendo al centro dell'analisi la figura dell'Avatar e la necessità di comprendere il punto di vista altrui attraverso un approccio metodologico qualitativo-procedurale. Partendo da un'introduzione all'Effetto Proteo e alla Retorica Procedurale, la ricerca ipotizza che l'immersione in un sistema di regole diverse costringa il giocatore a una mediazione attiva tra il proprio ordinamento valoriale e quello imposto dal software. Attraverso l'analisi di tre videogiochi, il lavoro mostra come le loro meccaniche facciano vivere all'utente tre esperienze diverse: il bisogno di collaborare con altre culture per sopravvivere, la difficile scelta tra seguire la burocrazia o la propria coscienza e la messa in discussione dei dogmi religiosi imposti. I risultati ottenuti hanno dimostrato che il "decentramento" avviene come un processo procedurale in cui l'interazione tra l'architettura narrativa del gioco e l'utente, obbliga quest'ultimo ad una mediazione costante.
Parole chiave: Videogiochi; Diritto Interculturale; Effetto Proteo; Game Design; Empatia;

ABSTRACT (ENGLISH)

Video games as a tool for intercultural mediation.
This paper investigates how video games can become a propaedeutic tool for Intercultural Law, placing at the center of the analysis the figure of the Avatar and the need to understand the other's point of view through a qualitative-procedural methodological approach. Starting from an introduction to the Proteus Effect and Procedural Rhetoric, the research hypothesizes that immersion in a different rule system forces the player to actively mediate between their own value system and the one imposed by the software. Through the analysis of three video games, the work shows how their mechanics enable the user to live three distinct experiences: the need to collaborate with other cultures to survive, the difficult choice between following bureaucracy or one's own conscience, and the questioning of imposed religious dogmas. The results showed that "decentering" occurs has a procedural process in which the interaction between the narrative architecture of the game and the player, forces the latter to constant mediation.
Keywords: Video games; Intercultural Law; Proteus Effect; Game Design: Empathy;

1. INTRODUZIONE

I giochi sono spesso considerati uno strumento puramente orientato all'intrattenimento, ma esiste oggi un interesse crescente nello sfruttare il loro potere per cause prosociali. Una manifestazione di questo interesse è l'emergere di progetti di ricerca e organizzazioni dedicate allo sviluppo di risorse videoludiche. Recenti studi hanno dimostrato come i giochi sociali operano offrendo ambienti di apprendimento flessibili e realistici mitigando rischi associati al mondo reale (Facchino et al., 2025): *iThrive Sim* (Rivers & Bertoli, 2024) è un gioco sviluppato per supportare studenti nella comprensione di concetti civici e storici, sviluppando al contempo le loro competenze socio-emotive. Lo *status quaestionis* vede i videogiochi come particolarmente adatti a sostenere programmi educativi in cui la promozione dell'empatia è un obiettivo chiave. Ciò accade perché questi strumenti permettono ai giocatori di abitare i ruoli di altre persone in un modo univocamente immersivo (Belman & Flanagan 2010). Il concetto di Narrative Architecture di Henry Jenkins (2004) ha permesso di superare la dicotomia tra ludologi (focalizzati sulle meccaniche e sulle regole di gioco) e narratologi (interessati alla trama e allo storytelling). Oggi, infatti, il videogioco è inteso come spazio in cui la narrazione è incastonata nell'architettura del mondo e nelle sue regole, trasformando l'ambiente in un testo visivo che il giocatore deve decifrare muovendosi nello spazio (Naul & Liu 2020). Il compito del Narrative Designer è quindi quello di arricchire lo spazio disseminandolo di indizi visivi e strutturali che attendono di essere attivati dal giocatore: non si tratta di raccontare una storia, ma di progettare la possibilità che essa venga scoperta. Attraverso la spazialità, il videogioco può

trasformare qualsiasi elemento narrativo in mezzo di trasmissione culturale e valoriale; ciò avviene perché nel videogioco gli elementi culturali sono oggetti fisici con cui è necessario interagire per proseguire. All'interno di questi spazi narrativi emerge la Gaming Religion: per abitare il mondo di gioco, l'utente deve accettarne i presupposti religiosi, compiendo rituali digitali che generano nel giocatore un senso di appartenenza e significato (Geraci 2012). Se questo approccio abitua i giocatori a muoversi tra sistemi di regole e valori diversi, il videogioco diventa il terreno ideale per sperimentare le sfide del Diritto Interculturale. Esso nasce dalla necessità di passare dall'istituzione alla relazione, superando la rigidità delle norme giuridiche per venire a contatto con altre culture in modo concreto (Consorti, 2011). Il Diritto Interculturale, infatti, richiede un cambio di prospettiva: il "decentramento". In tal senso, i videogiochi consentono ai giocatori di assumere diverse identità e di essere esposti a differenti contesti (Stamou et al., 2026), costringendo a un'immersione normativa diversa da quella dell'utente. Alla luce di queste premesse teoriche, il contributo intende rispondere alla seguente domanda di ricerca: in che modo l'esperienza videoludica, e in particolare l'uso dell'Avatar, abilita quel processo di "decentramento" necessario al Diritto Interculturale, trasformando il giocatore da osservatore passivo a mediatore attivo tra sistemi valoriali diversi? L'ipotesi di fondo è che il "decentramento" venga abilitato dalla Retorica Procedurale dei videogiochi, perché costringe l'utente a performare attivamente la mediazione tra il proprio sistema di valori e le regole del software come condizione necessaria per l'avanzamento nell'esperienza.

2. STATO DELL'ARTE E METODOLOGIA

Se l'Effetto Proteo (Yee & Bailenson, 2007) descrive come la rappresentazione digitale modifichi il comportamento dell'utente, la Retorica Procedurale di Ian Bogost (2007; Ciaurriz Velasco, 2010) spiega come le regole del gioco persuadano attraverso l'azione. In generale, l'utente tende a conformarsi a ciò che il suo avatar rappresenta: osservandone l'aspetto, ne deduce le caratteristiche e finisce per comportarsi di conseguenza; questo fenomeno è definito come Effetto Proteo. Il termine deriva dal Dio mitologico greco Proteo, che possedeva il potere della metamorfosi al fine di evitare di rivelare la sua conoscenza del passato, del presente e del futuro (Yee et al., 2009). Sono state condotte alcune sperimentazioni al fine di dimostrare come tale effetto si trasferisca in situazioni sociali del mondo esterno: Nel *Money Sharing Task* (Yee et al., 2009), i partecipanti che utilizzavano Avatar più alti tendevano a essere più egoisti nelle trattative economiche rispetto a quelli con Avatar più bassi. Altre ricerche, come quella su *FearNot!* (Belman & Flanagan, 2010), hanno dimostrato come l'interazione con personaggi virtuali vittime di bullismo potesse aumentare l'empatia nei bambini. In questo caso, essa non nasceva dalla semplice osservazione, ma dalla responsabilità che il bambino sentiva nei confronti del personaggio virtuale. Tuttavia, l'effetto non è sempre positivo: uno studio comparativo tra *Papers, Please* e *The Westport Independent* (Szolin et al., 2023) ha evidenziato come l'assunzione di ruoli di controllo possa talvolta ridurre la sensibilità verso l'Altro, portando a comportamenti più cinici per conformarsi alle regole del sistema. Questo porta ad approfondire il ruolo dell'empatia nel Game Design che può essere utilizzata come mezzo per modellare un videogioco che ha l'obiettivo di sensibilizzare gli utenti (Belman & Flanagan, 2010). Ricerche recenti suggeriscono che l'efficacia di tali strumenti risieda nell'azione procedurale in ambienti immersivi, poiché hanno un impatto più persistente sulla morale dell'utente (Lorenzo & Okahata, 2026b). Affinché l'esperienza videoludica possa generare un'empatia efficace, è necessario seguire specifiche direttive strutturali. Innanzitutto, poiché i giocatori non adottano spontaneamente un atteggiamento empatico, questa deve essere sollecitata intenzionalmente fin dalle fasi iniziali del gioco. Inoltre, stimolare la sofferenza condivisa senza offrire strumenti pratici di risoluzione rischia di generare un dolore controproducente; pertanto, è cruciale fornire raccomandazioni operative che permettano all'utente di agire per migliorare il contesto rappresentato. Infine, tale processo deve fondarsi sull'evidenziazione delle somiglianze tra il giocatore e il gruppo target, mantenendo però un equilibrio delicato per non innescare meccanismi di evitamento difensivo di fronte a una vicinanza percepita come troppo minacciosa. Infine, la Retorica Procedurale, definita come l'arte di persuadere attraverso le rappresentazioni e interazioni, costringe l'utente a internalizzare le regole del sistema per poter avanzare, rendendo il videogioco uno strumento valido anche per la simulazione delle dinamiche proprie del Diritto Interculturale. In questa prospettiva, l'esperienza ludica diventa un esercizio di "decentramento", permettendo di superare la semplice analogia formale tra norme giuridiche e regole di gioco. Il software, infatti, non si limita solo a comunicare funzioni culturali, ma spinge l'utente a una negoziazione attiva tra il proprio ordinamento valoriale e le norme del gioco. Questo processo di

"decentramento" trova il suo perno nell'Avatar: un ente[1] di mediazione tramite il quale il giocatore è obbligato a sospendere temporaneamente il proprio giudizio morale e normativo, appartenente al mondo reale, per adottare la logica giuridica ed etica imposta dal software. Coerentemente con questo impianto teorico, lo statuto metodologico del presente contributo si definisce come una analisi qualitativa di tipo procedurale. La ricerca applica una close reading procedurale a tre casi studio, utilizzando il framework MDA (Hunicke et al., 2004): un approccio formale applicato al Game Design tramite il quale è possibile comprendere come le decisioni formali sull'architettura del gioco influenzino l'esperienza finale dell'utente. La fenomenologia suggerisce infatti che il significato dell'esperienza emerga proprio dall'interazione tra l'utente e i valori culturali incarnati nel sistema di regole (Mota et al., 2025).
L'acronimo rivela tre componenti distinte, ma legate tra di loro, sotto cui i videogiochi vengono analizzati:

- Meccaniche (Mechanics), che rappresentano l'insieme delle varie azioni, comportamenti e meccanismi di controllo resi disponibili all'utente nel contesto di gioco.
- Dinamiche (Dynamics), che descrivono il comportamento delle meccaniche che emergono nel momento in cui esse interagiscono con l'utente.
- Estetiche (Aesthetics), che descrivono le risposte emozionali del giocatore quando interagisce con il sistema di gioco.

Per la ricerca, l'Estetica è il punto di ingresso da cui le Dinamiche e le Meccaniche vengono decostruite; questo percorso a ritroso permette, infatti, di svelare la Retorica Procedurale dei giochi, ovvero come il sistema di regole obblighi effettivamente il giocatore al processo di "decentramento".
I giochi selezionati sono *The Legend of Zelda: Tears of the Kingdom*, *Papers, Please* e *Indika* in quanto le loro meccaniche di gioco permettono di indagare rispettivamente tre dimensioni del Diritto Interculturale:

- La dimensione antropologica dell'interdipendenza culturale, declinata come mediazione in condizioni di asimmetria. In *The Legend of Zelda*, l'avatar non può progredire senza integrare i poteri delle popolazioni che abitano il gioco, trasformando la cooperazione in una necessità strutturale per la sopravvivenza.
- La dimensione normativa del conflitto tra norme, ovvero la tensione tra diritto positivo e giustizia sostanziale. In *Papers, Please* la meccanica di gioco impone all'utente la rigida applicazione della norma burocratica, nonostante la presenza di situazioni in cui l'empatia emerge mettendo in difficoltà l'utente.
- La dimensione teologica della traduzione dei sistemi di valore, cioè come viene gestita la collisione tra sistemi normativi diversi. In *Indika*, la mediazione avviene tramite la comprensione di concetti universali declinati nei vari mondi, mettendo alla prova la capacità di discernimento del giocatore.

L'unità di analisi è l'interazione tra il giocatore e il sistema di regole a cui è sottoposto, ovvero come le regole influenzano il modo in cui il giocatore usa l'Avatar; ciò permette di indagare come l'esperienza videoludica possa abilitare il processo di "decentramento" richiesto dal Diritto Interculturale.
Per limitare l'arbitrarietà interpretativa, l'analisi dei giochi non si è basata sull'esperienza diretta dell'autore ma è stata integrata dall'osservazione non partecipante di gameplay su YouTube. Questa scelta metodologica permette di documentare le reazioni spontanee di utenti inconsapevoli alle sollecitazioni del software, fornendo evidenze empiriche utili per l'indagine.

3. ANALISI DEI CASI STUDIO

In *The Legend of Zelda: Tears of the Kingdom* il tema della diversità è centrale[2]. È un videogioco sviluppato da Nintendo per Nintendo Switch rilasciato ufficialmente il 12 maggio 2023 come sequel diretto di *Breath of the Wild*. Il gioco inizia con una cutscene in cui la principessa Zelda e il suo cavaliere

[1] Treccani. (2026). Persona. In *Vocabolario Treccani on line*. https://www.treccani.it/vocabolario/persona/ (cons. 09/02/2026)
[2] Baldetti, S. (2025, 17 maggio). Zelda costruisce ponti. Giochi Sacri. https://giochisacri.substack.com/p/zelda costruisce-ponti (cons. 15/01/2026)

Link esplorano i sotterranei del castello di Hyrule per scoprire l'origine del "miasma", una sostanza tossica che ammala gli abitanti del regno. Lungo il percorso, i murales narrano di un periodo oscuro in cui Ganondorf rubò una Gemma Segreta trasformandosi nel Re dei Demoni. Nelle profondità, i due protagonisti trovano il cadavere mummificato di Ganondorf sigillato dalla mano di Rauru, primo Re di Hyrule; al loro arrivo, la mummia si risveglia sprigionando un'ondata di miasma che frantuma la Spada Suprema di Link e consuma il suo braccio destro. Le vibrazioni provocate dal potere di Ganondorf generano un terremoto: il castello di Hyrule inizia a fluttuare nell'aria staccandosi dalle fondamenta, mentre il suolo sotto Zelda crolla e lei precipita nell'abisso. Link, invece, viene salvato dalla stessa mano spettrale che sigillava la mummia. Da questo momento il giocatore ha pieno controllo sull'Avatar per esplorare un mondo profondamente mutato (Gabby16bit, 2023).
Seguendo il vocabolario del framework MDA, il gioco soddisfa diverse risposte emotive nel giocatore:

- **Fantasy**, l'utente viene trasportato in un nuovo mondo abitato da figure immaginarie, vestendo i panni di Link. Nel gameplay di Gabby16bit (2023) al minuto 04:02, l'utente manifesta l'immedesimazione rivolgendosi a Link in prima persona: *"Tutti questi cuori li avevo ottenuti per te Zelda, per dimostrarti il mio amore"*.
- **Discovery**, nonostante l'incipit narrativo, l'esplorazione è costante.
- **Narrative**: il gioco alterna l'azione a cutscenes che informano l'utente sul contesto, su risorse da trovare e raccogliere e su possibili nemici da sconfiggere durante l'esplorazione.
- **Challenge**, per ottenere uno dei poteri del braccio di Rauru, l'utente deve completare una sfida user-crafting. Nel gameplay (Gabby16bit, 2023, minuto 27:37), l'utente afferma: *"E' una sfida … Va bene, sono fatto per le sfide"*.
- **Fellowship**, questo aspetto emerge soprattutto nelle mini-sfide basate sullo scambio di aiuti tra Link e le comunità locali del regno di Hyrule. Al minuto 20:02 del gameplay di QuelTaleAle (2023, Ep. 12), Link interagisce con l'uccellina Momo al Borgo dei Rito: donandole della gelatina Chuchu bianca, ottiene nuove risorse (frecce x10) e informazioni per congelare i mostri.

Per analizzare la mediazione in condizioni di asimmetria, è stato identificato un sistema di feedback così strutturato:

- **Stato del sistema**, al minuto 01:12:00 del gameplay di QuelTaleAle (2023, Ep. 12) l'Avatar deve recuperare l'arco di Tulin da un mostro volante, ma l'utente si trova in uno stato di incompletezza funzionale poiché la distanza e le correnti d'aria rendono l'obiettivo irraggiungibile tramite le sole meccaniche di base.
- **Azione del giocatore**, per modificare lo stato del sistema, l'utente deve attivare intenzionalmente la meccanica della raffica di vento di Tulin mentre Link è in volo (meccanica della paravela di Link); questo mostra come l'azione del giocatore non è più individuale, ma mediato dalla cooperazione con l'alleato.
- **Feedback**, il correttore del sistema è costituito dall'interazione sinergica tra le meccaniche dei due personaggi, poiché il soffio di Tulin agisce come acceleratore che modifica il comportamento della paravela, generando un riscontro positivo immediato.

Questo ciclo genera la dinamica di **cooperazione necessaria**, resa esplicita nel gameplay di QuelTaleAle (2023, Ep. 14) al minuto 1:02:00 nello scambio di battute tra il padre di Tulin e il figlio: *"Ragazzi … ce l'avete fatta!"* – *"Si! E ci siamo riusciti perché abbiamo lavorato tutti assieme!"*.
Successivamente, sono state individuate le meccaniche che supportano la dinamica di cooperazione:

- **Paravela di Link**, meccanica di navigazione aerea che si attiva premendo il taso [A] "Vai!" del controller Nintendo Switch in prossimità di burroni o situazioni di volo assistito. Essa permette la planata ma richiede un supporto esterno per coprire grandi distanze orizzontali.
- **Raffica di Tulin**, meccanica attivabile in volo dall'utente tramite il taso [A] "Folata"; l'input produce un flusso d'aria più potente che aiuta il volo dell'Avatar.

L'analisi di *The Legend of Zelda: Tears of the Kingdom* ha dimostrato come il "decentramento" avviene quando Link (un uomo della Terra) accetta l'impossibilità di imporre la propria natura fisica al mondo di gioco, trovandosi costretto a integrare le abilità dell'Altro (Tulin, un uomo-uccello) per proseguire con l'esperienza ludica. Sospendendo la propria autonomia per integrare le tecniche delle diverse popolazioni, l'utente sperimenta attivamente il passaggio dall'istituzione alla relazione.

Papers, Please è un videogioco sviluppato da Lucas Pope e pubblicato l'8 agosto 2013, inizialmente su Windows e MacOS e successivamente reso disponibile per altre piattaforme. Il titolo è ambientato nel paese fittizio di Arstotzka, nella città confine di Grestin Est, separata dalla vicina Kolechia. L'azione si svolge nel 1982, in seguito ad un conflitto sanguinoso durato sei anni che ha reso i rapporti diplomatici estremamente tesi e la situazione alla frontiera instabile[3]. Nel corso di 31 giornate lavorative, l'ispettore dell'immigrazione deve esaminare i documenti delle persone alla frontiera per scegliere chi potrà entrare e chi dovrà essere respinto. Alla fine di ogni giornata, il giocatore riceve uno stipendio basato esclusivamente sul numero di persone esaminate correttamente; questo denaro è vitale per pagare l'affitto, il riscaldamento e le medicine necessarie alla sopravvivenza della famiglia dell'ispettore.
Nel titolo di Lucas Pope sono stati individuati i seguenti modelli estetici:

1. **Fantasy**, l'utente viene da subito inserito in un nuovo contesto nel quale deve ricoprire il ruolo di un ispettore sottomesso alle norme burocratiche del gioco. L'immedesimazione è immediata poiché al minuto 4:40, FaviJ (2018) esclama: *"Dai che tra tre minuti devo andare a casa, non posso stare qui tutto il giorno!"*.
2. **Sensation**, l'uso di una colonna sonora introduttiva cupa e di un italiano impoverito di articoli e preposizioni, la presenza dei comandi "Cammina al lavoro", "Dormi" e scritte come "Gloria ad Arstotzka", agiscono come stimoli sensoriali che spingono l'utente ad interiorizzare il proprio ruolo nell'apparato statale.
3. **Narrative**, la trama è di tipo procedurale poiché evolve aumentando la difficoltà delle regole a cui l'utente deve sottostare. All'inizio, l'attività di controllo è limitata ai soli cittadini di Arstotzka e solo dopo, il gioco richiede anche la validazione di passaporti, cedolini d'entrata e permessi di soggiorno per gli stranieri. L'esperienza è anche intervallata dalla lettura di giornali che hanno lo scopo di fornire la cornice contestuale necessaria per legittimare la crescente severità delle procedure. Per analizzare il conflitto tra la norma giuridica e la giustizia sostanziale, è stato isolato un caso osservabile nel gameplay di Gabby16bit (2016) dal minuto 11:11:
4. **Stato del sistema**, il giocatore riceve una richiesta d'aiuto informale (una nota scritta) da una ragazza che teme di essere vittima di tratta da parte di Dari Ludum, un uomo che quel giorno è in attesa di essere ammesso ad Arstotzka. Poco dopo, l'utente si presenta allo sportello e i suoi documenti sono formalmente perfetti.
5. **Azione del giocatore**, l'utente deve ora scegliere se approvare l'ingresso dell'uomo oppure negarglielo per proteggere la vittima, consapevole di commettere un errore burocratico che lo penalizzerà sullo stipendio. Al minuto 12:02 emerge nell'utente un conflitto interiore: *"Cosa fare? Cosa fare? Adesso ho paura e mi batte forte il cuore perché se vedo quel tizio devo prendere una scelta … forse morale forse immorale … non lo so"*. Alla fine, il giocatore sceglie di aiutare la vittima rifiutandogli così l'ingresso nel Paese.
6. **Feedback**, il sistema di gioco risponde con un ammonimento immediato nei confronti dell'utente (FaviJ 2018, minuto 25:00); questo riscontro rappresenta la volontà del gioco di riportare l'utente al suo ruolo di mero esecutore delle regole, sanzionando il decentramento etico del giocatore.

Il ciclo di feedback individuato genera la dinamica della **scelta tragica**, poiché la mediazione interculturale e l'empatia verso l'Altro sono costose.
Successivamente, sono state individuate le meccaniche a supporto di questa dinamica:

- **Timbro Rifiuta/Approva**, meccanica di azione primaria che l'utente utilizza per prendere una decisione, attivabile cliccando il tasto sinistro del mouse.
- **Ammonimento M.A.**, meccanica di feedback sanzionatorio che scatta quando l'azione del giocatore non rispetta le norme del software.
- **Gestione delle risorse**, questa meccanica lega l'esito dei turni lavorativi alla sopravvivenza dei familiari dell'utente poiché sa che il prezzo del suo "decentramento" morale ricade direttamente sul benessere del proprio nucleo familiare.

[3] Università Cattolica del Sacro Cuore. (2024). *Papers, Please*. Atlante delle distopie mediali. https://www.unicatt.it/uc/atlantedistopiemediali-tutte-le-schede-papers-please (cons. 09/02/2026)

L'analisi di *Papers, Please* ha dapprima evidenziato che il giocatore viene assorbito dal ruolo burocratico, percependo la norma di Arstotzka come l'unico sistema valoriale possibile. Tuttavia, il "decentramento" avviene quando l'utente smette di essere un semplice esecutore, per dare priorità al sistema valoriale dell'Altro. Il passaggio dall'istituzione alla relazione avviene quando l'utente decide di disobbedire, accettando il costo economico della propria scelta. Questo passaggio riflette una maturazione del ragionamento morale: l'utente abbandona una logica basata sulla ricompensa immediata per accogliere principi morali e di giustizia, simulando la transizione verso stadi superiori di giudizio morale dei dilemmi narrativi complessi (Holl et al., 2020).
Indika è un videogioco sviluppato dallo studio Odd Meter e rilasciato il 17 maggio 2024 su PS5[4]. Ambientato in una Russia alternativa del XIX secolo, il titolo segue le vicende della giovane suora Indika tormentata dalla voce del Diavolo. Il gioco si configura come un'opera di decostruzione religiosa[5], le cui meccaniche smantellano il concetto di fede basato sulla ricompensa, creando una dissonanza cognitiva nel giocatore[6] attraverso una "pedagogia dell'inutilità". Nel gioco, l'utente è invitato a compiere azioni devote che incrementano un contatore di "Punti Fede" visibile a schermo; tuttavia, questa accumulazione è semanticamente vuota perché i punti non sbloccano abilità, non migliorano le statistiche e non influenzano la narrazione. Questa inutilità viene esplicitata nel gameplay di Prizio (2024) al minuto 32:53, quando una schermata recita: *"Don't waste time collecting points. They are pointless."*.
In questo gioco sono stati individuati i seguenti modelli estetici:

1. **Challenge**, all'utente viene chiesto di svolgere dei doveri, come prendere una cesta di verdure dal magazzino da consegnare a Suor Smaragda (Prizio, 2024, minuto 3:30), che permettono di progredire nella storia del gioco anche se l'accumulazione dei punti è inutile.
2. **Discovery**, l'utente è inserito in una ambientazione da esplorare per trovare oggetti utili nel superamento delle prove; il giocatore è spesso guidato dalla voce della stessa Indika che suggerisce, ad esempio, la direzione corretta per proseguire nell'esperienza, come avviene al minuto 1:28:17 del Gameplay di Prizio (2024): *"The stairs are broken, we'll have to go through the mill."*.
3. **Narrative**, l'esperienza videoludica è intervallata da cutscenes cronologiche e dialoghi in cui la storia viene rivelata progressivamente. Per analizzare la traduzione di sistemi valoriali diversi, è stato individuato questo ciclo di feedback:
4. **Stato del sistema**, dal minuto 1:27:50 il software pone l'utente di fronte ad un improvviso inseguimento da parte di un lupo gigante. In questo contesto, emergono tre punti di vista differenti sulla figura della bestia: per Ilya, compagno di viaggio di Indika, il lupo è un peccatore che merita di morire; Indika ribatte che solo gli umani possono peccare; il Diavolo poi interviene smontando entrambe le visioni, affermando che ciò che per l'uomo è peccato, in realtà è solo un impulso biologico condiviso da ogni specie.
5. **Azione del giocatore**, l'utente attiva le meccaniche necessarie per eliminare il lupo e proseguire nell'esperienza assieme a Ilya, sospendendo temporaneamente ogni giudizio sui tre punti di vista.
6. **Feedback**, il sistema risponde con la sopravvivenza dell'Avatar; tuttavia, l'esito positivo non convalida nessuno dei punti di vista proposti, ma conferma il sistema procedurale del software: davanti alle necessità, le spiegazioni religiose crollano.

Il ciclo di feedback individuato genera la dinamica della **traduzione di valori** nella quale sono state individuate le seguenti meccaniche:

- **Spostamento di oggetti**, l'utente trascina un barile con ruote, utilizzandolo come piattaforma per salire sull'ascensore di mattoni e superare così l'ostacolo architettonico.
- **Incastro meccanico**, l'utente deve individuare il perno corretto e raccoglierlo, così da inserirlo nell'incavo tra due componenti meccaniche; ciò garantisce la trasmissione della coppia motrice, attivando la rotazione della ruota.

[4] Sony Interactive Entertainment. (2024). *INDIKA*. PlayStation Store. https://www.playstation.com/it-it/games/indika/ (cons. 28/01/2026)
[5] Baldetti, S. (2025, 19 febbraio). Gaming con le suore. *Giochi Sacri*. https://giochisacri.substack.com/p/gaming-con-le-suore (cons. 15/01/2026)
[6] Baldetti, S. (2025, 28 giugno). La fabbrica dei pixel. Giochi Sacri. https://giochisacri.substack.com/p/la-fabbrica-dei-pixel (cons. 15/01/2026)

- **Controllo dell'equilibrio**, il giocatore deve controllare l'equilibrio di Indika su superfici precarie per attirare l'attenzione del lupo e condurlo verso la ruota in movimento per eliminarlo.

L'analisi di *Indika* ha rivelato che il "decentramento" avviene quando l'utente è costretto a sospendere il proprio giudizio su tre punti di vista diversi per rispondere all'urgenza di sopravvivere nel gioco. Il passaggio dall'istituzione alla relazione avviene quando l'utente sceglie di seguire la logica procedurale del software per salvare la vita non solo a sé stesso, ma anche all'Altro (Ilya).

4. CONCLUSIONI

La presente indagine è l'inizio di un percorso di ricerca da approfondire, poiché ha dimostrato che l'esperienza videoludica abilita il "decentramento" richiesto nelle dimensioni antropologica, normativa e teologica del Diritto Interculturale. Attraverso la Retorica Procedurale emersa nell'analisi dei casi studio, il giocatore non si limita ad osservare la diversità, ma è costretto a performarla per sopravvivere nel sistema. Tuttavia, poiché la sola analisi dei gameplay non è sufficiente, le ricerche future dovranno coinvolgere un campione di utenti per verificare i risultati tramite sessioni di gioco e la compilazione di questionari. Tali studi dovranno anche analizzare la persistenza dell'Effetto Proteo nel lungo periodo e il rischio che la simulazione di ruoli negativi possa desensibilizzare l'utente anziché educarlo, esplorando come queste dinamiche mutino negli ambienti multiplayer, dove l'utente incontra diversità reali e non simulate.

RINGRAZIAMENTI

L'autore ringrazia la Professoressa Enrica Salvatori dell'Università di Pisa per il supporto e la guida. In conformità con i principi di trasparenza, si dichiara che sono stati utilizzati strumenti di Intelligenza Artificiale Generativa (LLM) esclusivamente per il supporto nella revisione stilistica e nella sintesi dei testi originali dell'autore. L'ideazione, l'analisi critica e i contenuti scientifici sono opera originale dell'autore.

BIBLIOGRAFIA

Belman, J., & Flanagan, M. (2010). Designing Games to Foster Empathy. *International Journal of Cognitive Technology*, *14*(2), 11-21.

Bogost, I. (2007). *Persuasive Games: The Expressive Power of Videogames*. MIT Press.

Ciaurriz Velasco, F. (2010). Persuasive Games. The Expressive Power of Videogames [Reseña]. *Comunicación: Revista Internacional de Comunicación Audiovisual, Publicidad y Estudios Culturales*.

Consorti, P. (2011). Per un diritto interculturale. In S. Gianfaldoni (A cura di), *Esperimenti interculturali. Prove d'autore* (pp. 9-12). Plus.

Facchino, A. P., Marchetti, D., Colasanti, M., Fontanesi, L., & Verrocchio, M. C. (2025). The use of serious games for psychological education and training: a systematic review. *Frontiers in Education*, *10*. https://doi.org/10.3389/feduc.2025.1511729

FavijTV. (2018, 2 febbraio). *PROVO PAPERS PLEASE per LA PRIMA VOLTA* [Video]. YouTube. https://www.youtube.com/watch?v=_ZdPnOeU7eE

Gabby16bit. (2016, 3 gennaio). *DANNATI TERRORISTI! - Papers, Please Ep.2* [Video]. YouTube. https://www.youtube.com/watch?v=TrW5W1ALe98

Gabby16bit. (2023, maggio). *GIOCO DELL' ANNO. The Legend of Zelda: Tears of the Kingdom #1* [Video]. YouTube. https://www.youtube.com/watch?v=sD0dCtX-PxI

Geraci, R. M. (2012). Theological Productions: The Role of Religion in Video Game Design. In A. L. Brackin & N. Guyot (A cura di), *Cultural Perspectives of Video Games: From Designer to Player* (pp. 101-114). Inter-Disciplinary Press.

Holl, E., Bernard, S., & Melzer, A. (2020). Moral decision-making in video games: A focus group study on player perceptions. *Human Behavior and Emerging Technologies*, *2*(3), 278–287. https://doi.org/10.1002/hbe2.189

Hunicke, R., LeBlanc, M., & Zubek, R. (2004). MDA: A formal approach to game design and game research. In *Proceedings of the AAAI Workshop on Challenges in Game AI* (Vol. 4, No. 1, p. 1722).

Jenkins, H. (2004). Game Design as Narrative Architecture. In P. Harrigan & N. Wardrip-Fruin (A cura di), *First Person: New Media as Story, Performance, and Game* (pp. 118-130). MIT Press.

Lorenzo, S. B., & Okahata, L. (2026b). Empathy in action: cultivating altruism through immersive game experiences. *Frontiers in Virtual Reality*, *6*. https://doi.org/10.3389/frvir.2025.1716138

Mota, V. F., Duarte, E. F., Instituto de Computação, Universidade Estadual de Campinas, & Colégio Técnico, Universidade Federal de Minas Gerais. (2025). Phenomenology and Digital Games: perspectives on why and directions on how. In *PhenoHCI'25: Phenomenological Concepts and Methods for HCI Research Workshop*. https://ceur-ws.org/Vol-1234/paper-5.pdf

Naul, E., & Liu, M. (2020). Why Story Matters: A Review of Narrative in Serious Games. *Journal of Educational Computing Research*, *58*(3), 687-707.

Prizio. (2024, 17 maggio). *INDIKA - Gameplay ITA - #1* [Video]. YouTube. https://www.youtube.com/watch?v=bnOyO7VOxho

QuelTaleAle Games (2023, 23 maggio). *The Legend of Zelda: Tears of the Kingdom - Gameplay ITA - 12 - IL VILLAGGIO DEI RITO* [Video]. YouTube. https://www.youtube.com/watch?v=rLIMqUaoD7w

QuelTaleAle Games (2023, 25 maggio). *The Legend of Zelda: Tears of the Kingdom - Gameplay ITA - 14 - SFIDA BOSS TEMPIO DEL VENTO* [Video]. YouTube. https://www.youtube.com/watch?v=Ds3VrGYeSYg

Rivers, S. E., & Bertoli, M. C. (2024). Using games to ignite teens' civic and social and emotional learning. *Frontiers in Education*, *9*. https://doi.org/10.3389/feduc.2024.1322721

Stamou, S., Apostolakis, K. C., & Stephanidis, C. (2026). An Integrative review of empathy, attitude change, and historical consciousness in Games: Mapping Gaps and Opportunities in Game-Based Digital Heritage Research. *Heritage*, *9*(1), 13. https://doi.org/10.3390/heritage9010013

Szolin, K., Kocur, M., Soucek, R., & Tiede, K. (2023). Exploring the user-avatar relationship in videogames: A systematic review of the Proteus effect. *Human-Computer Interaction*, *38*(5-6), 374-399. https://doi.org/10.1080/07370024.2022.2106474

Yee, N., Bailenson, J. N., & Ducheneaut, N. (2009). The Proteus Effect: Implications of Transformed Digital Self-Representation on Online and Offline Behavior. *Communication Research*, *36*(2), 285-312. https://doi.org/10.1177/0093650208330254

Edizione digitale commentata di *Le libere donne di Magliano* di Mario Tobino: primi passi, metodi e prospettive

FABIO ZARROLI
UNIVERSITÀ DEGLI STUDI DI PARMA, ITALY - FABIO.ZARROLI@UNIPR.IT

ABSTRACT (ITALIANO)

Il contributo ha per obiettivo la presentazione del progetto di edizione digitale commentata di *Le libere donne di Magliano* di Mario Tobino (Vallecchi, Firenze 1953). Il progetto si inserisce nell'ambito della letteratura a tema manicomiale e degli psichiatri scrittori e coniuga alle nuove opportunità offerte dalle Digital Humanities una intensa attività di ricerca archivistica e documentale, condotta sia sulle carte del fondo autoriale che su documenti di carattere istituzionale. L'edizione mira perciò, da un lato, a fornire gli strumenti interpretativi e documentali necessari a sciogliere nodi e questioni interne al testo e/o alla sua stesura, dall'altro, ad aprire possibilità di analisi linguistica (confronto tra testo clinico e testo letterario), stilistiche (ricorrenze formali, tematiche e linguistiche) e di contenuto (documentazione dei fatti reali), permesse da precise scelte di codifica. Per quest'ultima si utilizza il linguaggio standard *de-facto* nell'ambito delle edizioni scientifiche digitali XML-TEI, mentre per la visualizzazione e per la fruizione del testo si è ricorso al sistema web TEI publisher.
Parole chiave: Edizione digitale commentata; Mario Tobino; Le libere donne di Magliano; Medical Humanities.

ABSTRACT (ENGLISH)

Annotated digital edition of Le libere donne di Magliano *by Mario Tobino: first steps, methods, and prospects.*
The aim of this contribution is to present the project for a digital annotated edition of Le libere donne di Magliano by Mario Tobino (Vallecchi, Florence 1953). The project falls within the scope of literature on the theme of mental hospitals and psychiatrist writers and combines the new opportunities offered by Digital Humanities with intensive archival and documentary research, conducted both on the author's papers and on institutional documents. The edition therefore aims, on the one hand, to provide the interpretative and documentary tools necessary to resolve issues within the text and/or its drafting and, on the other, to open up possibilities for linguistic analysis (comparison between clinical and literary texts), stylistic (formal, thematic and linguistic recurrences) and content analysis (documentation of real events), made possible by precise coding choices. For the latter, the *de-facto* standard language used in digital scientific editions, XML-TEI, is used, while the TEI publisher web system is used for the visualisation and function of the text.
Keywords: Annotated digital edition; Mario Tobino; The mad women of Magliano; Medical Humanities.

1. INTRODUZIONE: CONTESTUALIZZAZIONE DEL PROGETTO

L'Edizione digitale de *Le libere donne di Magliano*, testo dello psichiatra viareggino Mario Tobino, è un progetto inerente a un percorso dottorale che tuttavia si inserisce in una dimensione di ricerca più ampia, di interessi e prospettive condivisi con la Fondazione Mario Tobino (di cui è Presidente Isabella Tobino, nipote dell'autore), il Laboratorio Seminariale di Ricerca dell'Università di Pisa *Le scritture della malattia mentale: tra cartella clinica e psicopatografia* coordinato dalla professoressa Marina Riccucci e con il CNR di Pisa, in particolare con il professor Angelo Mario Del Grosso (Riccucci, Zamperini, Di Carlo, 2023). La valorizzazione del patrimonio documentale, letterario e archivistico di argomento psichiatrico e manicomiale è un elemento di grande rilevanza per l'approfondimento, la conoscenza e la divulgazione di una questione di elevato interesse storico, culturale e sociale (Chiaravallotti, Taverniti, Dovetto, 2023; Serratore, Portaro, Donato, Chiaravallotti, 2025). Si avverte a ogni modo che ricerca e divulgazione avvengono (e devono avvenire) nel rispetto della sensibilità dei dati e delle norme sulla privacy vigenti (Bagnai Losacco, 2024).

A metà strada tra romanzo e diario, *Le libere donne di Magliano* è un testo che ben si presta a essere interpretato e analizzato con gli strumenti del commento. Anzi, si può affermare che la lettura de *Le libere donne di Magliano*, non semplice in primo luogo per la frammentarietà della struttura, necessiti di essere accompagnata da un apparato critico e da un commento capaci di coglierne e debitamente spiegarne le peculiarità stilistiche e i fattori di originalità.
L'edizione digitale permette di collocare ai piedi del testo un ampio commento corredato dalla documentazione originale e in gran parte inedita rinvenuta durante un minuzioso lavoro di ricerca condotto su tre archivi: il fondo autoriale (Archivio Bonsanti del Gabinetto Vieusseux), l'archivio dell'ex Ospedale Psichiatrico Provinciale di Maggiano (Archivio di Stato di Lucca) e l'Archivio Storico della Provincia di Lucca, ai quali si aggiungono i materiali a stampa e bibliografici conservati presso la Fondazione Ragghianti di Lucca e la Fondazione Mario Tobino.
L'ipotesi del lavoro che si presenta nasce dunque dall'esigenza di rendere disponibile in un'edizione commentata uno dei testi più autorevoli in materia di narrazione della malattia mentale e della vita manicomiale, inserendolo all'interno di una tradizione che recenti pubblicazioni hanno contribuito a consolidare (Riccucci, 2025). Dal punto di vista strettamente letterario, l'aspetto più innovativo consiste proprio nella pubblicazione dell'apparato critico, costituito da un'introduzione e da un commento inediti, data anche l'irreperibilità dei 13 quaderni su cui era stato vergato il manoscritto (Italia, 2007).
La versione digitale consente inoltre di ricorrere a strumenti di analisi e risorse afferenti all'ambito delle Digital Humanities che permettono al lettore, grazie a oculate scelte nella strutturazione della codifica (XML), l'accesso a sistemi di analisi linguistica (confronto tra testo clinico e testo letterario), stilistica (occorrenze formali, tematiche e retoriche) e contenutistica o tematica (documentazione dei fatti reali) che si svilupperanno nel corso del progetto di edizione.
La scelta del formato digitale permette infine di includere fonti secondarie (traduzioni, recensioni d'epoca uscite su giornali) e ovviare i limiti di spazio imposti dalle tradizionali versioni cartacee: le note di commento si collocano a piè di pagina attraverso l'utilizzo dell'elemento *<note place="bottom" type="editorial">*, per creare un luogo nel foglio di visualizzazione nel quale inserire le informazioni paratestuali indispensabili per una piena ed esauriente comprensione del testo.

2. STATO DELL'ARTE

Il testo che si prende a riferimento è quello utilizzato anche per le edizioni cartacee moderne, ossia quello della prima edizione Vallecchi 1953 (Tobino, 1953). Tra le edizioni de *Le libere donne di Magliano* edite sino a oggi, non ve ne sono né digitali, né che abbiano un commento al testo. Tra le edizioni disponibili si elencano: la prima, edita da Vallecchi nel 1953 (Tobino, 1953); l'edizione Mondadori del 1963 (Tobino, 1963) che aggiungeva al testo vallecchiano la Prefazione di Tobino dal titolo *Dieci anni dopo*; dopo varie riedizioni o ristampe, nel 2007 il Meridiano (Mondadori) curato da Paola Italia (Italia, 2007) inseriva in Appendice uno scritto in forma di diario dal titolo *I° quaderno del II° Maggiano*; infine, l'edizione del 2024 curata da Paola Italia (Italia, 2024) in cui appaiono un'Avvertenza della curatrice, la ristampa di *Dieci anni dopo* (con aggiunta di una nota a piè di pagina della curatrice) e uno scritto di Paola Italia dal titolo *Mario Tobino e* Le libere donne di Magliano.
Sul versante delle edizioni digitali, un modello preso a riferimento è l'edizione de *I promessi sposi* inerente al progetto Leggo Manzoni ideato da Paola Italia e Francesca Tomasi (https://projects.dharc.unibo.it/leggomanzoni/), in cui il testo del romanzo è connesso ai testi dei quaranta commenti, anche se con necessità di uno «stand-off markup approach» (Levchenko, Nava, Russo, 2025, p. 265) e quindi di identificazione dei singoli elementi <note> tramite l'attributo @xml:id, in ragione proprio della pluralità dei commenti inclusi. Altri elementi di interesse di questo modello sono l'inclusione delle sezioni "Traduco" e "Vedo", che integrano le traduzioni straniere e le edizioni a stampa (nel nostro caso sostituibili da ritagli di giornale delle recensioni d'epoca). Infine, si guarda con interesse alla creazione di un repository GitHub che mette a disposizione dell'utente la documentazione tecnica del progetto.
Altri progetti che si prendono a modello dal punto di vista della restituzione dei dati e dell'indicizzazione delle sezioni sono l'Edizione digitale delle Lettere di Vespasiano da Bisticci (https://vespasianodabisticciletters.unibo.it/) e la Semantic Digital Edition of Paolo Bufalini's notebook (https://projects.dharc.unibo.it/bufalini-notebook/) curate da Francesca Tomasi.

3. METODOLOGIE ADOTTATE

Il linguaggio di codifica utilizzato è XML-TEI, per cui nella scelta degli elementi ci si avvale delle Linee guida TEI per la rappresentazione standard delle risorse testuali (Burnard, 2014; Pierazzo, 2015; Driscoll, Pierazzo, 2016). Il programma di visualizzazione scelto è TEI Publisher, un framework open-source progettato per la realizzazione, pubblicazione e condivisione di edizioni scientifiche digitali. La piattaforma segue infatti gli standard internazionali per la codifica dei dati testuali e bibliometrici e consente di immettere diversi tipi di presentazione testuale e annotazioni. La scelta è ricaduta su TEI Publisher in quanto si ritiene il più flessibile e ready-to-use tra quelli esaminati, con possibilità di ampie personalizzazioni ed elaborazione dei dati.
L'edizione che si prepara riproduce fedelmente il testo dell'edizione vallecchiana del 1953, ma si segnalano in apparato le pur poche varianti intercorse nell'edizione Mondadori del 1963 (si prevede l'utilizzo dell'elemento <app> con <rdg> e attributo "source", trattandosi di testo a stampa – "wit" nei casi in cui la comparazione avviene con un documento manoscritto –), oltre alle variazioni del testo che si sono generate in fase di stesura, rilevabili dal confronto con la documentazione d'archivio.
Come criterio generale, in fase di codifica si descrivono tutti quegli elementi testuali (nel testo autoriale e nell'apparato critico) che si prestano ad analisi di carattere tematico e linguistico. Per fare solo alcuni esempi e riferendoci ai tag più utilizzati, si etichettano: i dati bibliometrici con il tag <ref type="bibl">es_bibl</ref> (i diversi <ref> si distinguono per attributi), le date con il tag <date when="es_date">es_date</date>, i nomi delle istituzioni in cui si è rinvenuto il documento con il tag <orgName role="repository">es_archive</orgName>, i nomi di persona con il tag <name type="person">es_name_person</name>, i toponimi con il tag <placeName>es_namePlace</placeName>.
Dal punto di vista strutturale, si mantiene la divisione in paragrafi del testo originale, i quali si individuano in codifica come singole unità di analisi mediante l'elemento , all'interno della sezione <head></head>. A ognuna di queste unità è attribuito un numero visualizzabile (all'interno dell'elemento <label>), il quale però è da considerarsi afferente all'apparato critico e non una modifica al testo autoriale. I paragrafi sono in totale 144 e sebbene per la maggior parte siano irrelati fra loro, molti di essi, benché dislocati o non continuativi, hanno una sequenzialità tematica che l'autore del commento ha il dovere di ricostruire e debitamente segnalare a fini esegetici ed ermeneutici. Per questa ragione i vari che si riferiscono a una stessa vicenda, a una stessa paziente, a una stessa patologia ecc., si mettono in relazione mediante l'elemento <relation> (preferibile, in questo caso, a <link>).
In prospettiva, l'interesse analitico ricade perciò sul problema della rappresentazione delle risorse da mettere a disposizione dell'utente finale, soprattutto per quanto riguarda la presenza sistematica, nel testo, di elementi rilevanti dal punto di vista esegetico. Anche in questo caso, valgano pochi esempi particolarmente suggestivi: elementi del linguaggio medico-scientifico, utilizzo della metafora zoologica, tematizzazione dell'erotismo, impiego di figure del mito o sacre.
La codifica viene salvata di volta in volta sul portale GitHub, un sistema che gestisce efficacemente il versionamento e il tracciamento dei dati, usato per archiviare, gestire, condividere e collaborare su progetti software. L'utilizzo di GitHub permette infatti anche di condividere l'accesso alle risorse e ai documenti di codifica sia in modalità di sola visualizzazione che abilitando alla modifica. L'utilizzo di GitHub risulta di estrema utilità in quanto permette il salvataggio delle varie modifiche che di volta in volta si effettuano al testo di codifica. In questo modo è possibile restituire, a lavoro ultimato, la cronologia delle modifiche apportate al testo di codifica, la quale a sua volta si presta a essere oggetto di eventuale analisi e valutazione.
Si segnala infine che allo stato attuale il lavoro manca di una sezione di elenco bibliografico strutturata, la cui integrazione è prevista in una fase successiva del lavoro.

```
<p>
    Oggi
        <note place="bottom" type="editorial"> Stando al <ref type="bibl">I° quaderno del II°
            Magliano</ref> il <distinct xml:lang="lat" rend="italic">terminus post quem</distinct
            >per la stesura del manoscritto è il <date when="1950-04-19">19.04.1950</date>. Con
            buona probabilità il manoscritto è stato redatto in forma di diario; infatti, il
            testo contiene numerosi riferimenti temporali specifici probabilmente espunti in fase
            di trascrizione dattiloscritta. Nella filza dell' <orgName role="repository">Archivio
            Storico</orgName> della <settlement type="prov">Provincia di Lucca</settlement>
            contenente i fascicoli con le singole pratiche di ammissione dei malati del <date
            >1950</date>, alla data <date when="1950-04-19">19.04.1950</date> è presente soltanto
            un fascicolo relativo a <name type="person">Ds=-  .  M=  .</name> <ref type="prot"
            >(Prot. N° 137 del 1950, fasc. n° 62)</ref>. Dai documenti contenuti nell'articolo si
            può soltanto risalire al luogo e alla data di nascita di <name>Deidamia Marchi</name>
            (<placeName>Lunata</placeName>, <date when="1910-08-27">27.8.1910</date>), e non all'
            <orgName> Ospedale </orgName> o al luogo di provenienza. All'interno della filza è
            presente anche un fascicolo contenente fogli sparsi e bifogli riguardanti pazienti
            trasferiti da altri manicomi (come si vedrà più avanti per la malata <name type
            ="person">Belaglia</name>) o liste di pazienti a carico della <settlement>provincia
            di Lucca</settlement>, cioè incapaci di sostenere in proprio le spese di retta
            manicomiale. Da questi fogli non si è in grado di individuare la paziente.</note>
    è arrivata, proveniente da Firenze, una malata, una matta, giovane, fresca, alta, con lo
        stampo della salute fisica. Quando sono entrato nel reparto
        <note place="bottom" type="editorial"> <name>Tobino</name> nel <date> 1948 </date> (come
            documentano i diari) vinse il ruolo di <roleName> primario </roleName> della
            divisione femminile dell' <orgName> Ospedale Provinciale Psichiatrico </orgName> di
            <placeName> Maggiano </placeName>, dopo aver tentato diversi concorsi in altri
            ospedali psichiatrici. Molti appunti di diario testimoniano incertezza riguardo alle
            volontà di restare a <placeName> Lucca </placeName> o cambiare luogo in cui lavorare.
            La divisione femminile era divisa in reparti, nominati in tutto il romanzo. A seconda
            dell'intensità dei deliri le malate venivano divise come segue: reparto <hi
            >"tranquille"</hi>, reparto <hi>"vigilanza"</hi>, reparto <hi>"agitate"</hi>.</note>
    era seduta a letto e mangiava con golosità. Aveva la camicia aperta sì che si vedeva
        comodamente un seno. Non aveva alcun pudore, neppure la finzione del pudore. È affetta da
        "schizofrenia", quella malattia mentale che scompone la persona umana rendendola senza
        senso e senza scopo.
```

Figura 1. Esempio di codifica di un estratto dell'apparato

4. PRIMI ESITI DELLA RICERCA, RISULTATI ATTESI E PROSPETTIVE

Nel tentativo di introdurre il lavoro e di spiegare le scelte di funzione e di codifica a oggi adottate, si è accennato, seppur *en passant*, ai primi risultati conseguiti e a quelli che si attendono dal progetto in questione. Si desidera però rendere noto con più specificità lo studio condotto sul materiale consultato e della sperimentazione che in sede di implementazione delle risorse digitali questo può apportare.
La documentazione inedita rinvenuta è ingente e di considerevole significatività: sono stati studiati, presso l'Archivio Bonsanti del Gabinetto Vieusseux, la serie I del Fondo Mario Tobino (FMTb), ossia le scatole contenenti materiale manoscritto, in relazione alla produzione autoriale di argomento manicomiale e gli oltre 180 manoscritti di Diario. È stata studiata interamente la serie II (Corrispondenza), consistente in 1866 unità archivistiche. Inoltre, sono state studiate, in relazione al materiale afferente a composizione e circolazione di *Le libere donne di Magliano* la serie III (Materiale a stampa – e relative sottoserie: scritti su e di Mario Tobino –) e la serie IV (Carte varie). Per quanto riguarda l'archivio ospedaliero, conservato presso l'Archivio di Stato di Lucca, sono state consultate le Rubriche uomini e le Rubriche donne contenenti l'elenco degli ingressi e delle dimissioni con relative date, e circa 100 Cartelle cliniche, tra le quali sono state individuate circa 20 cartelle corrispondenti ai personaggi de *Le libere donne di Magliano*. Queste cartelle contengono, nella sezione *Diario clinico*, testi manoscritti di Tobino indispensabili per l'individuazione dell'avantesto e della datazione di diversi paragrafi de *Le libere*, nonché per approfondimenti di ordine scientifico riguardanti il ruolo di Tobino all'interno dei vari orientamenti psichiatrici novecenteschi (Babini, 2009). Presso l'archivio provinciale ubicato all'Archivio Storico della Provincia di Lucca sono stati invece consultati circa 60 faldoni (oltre a varie rubriche e schede archivistiche) contenenti documentazione (delibere delle assemblee della giunta provinciale, corrispondenza con l'Ospedale psichiatrico, elenchi dei ricoveri ecc.) relativi agli anni di composizione del testo (1950, 1951, 1952).
La ricerca archivistica è tutt'ora in corso e in fase di costante aggiornamento. Non si esclude l'eventuale sperimentazione di tecniche di Handwritten Text Recognition (HTR) per la gestione del contenuto dei documenti manoscritti. Infine, si sperimentano tecniche di International Image Interoperability Framework (IIIF) per collegare elementi del testo di Mario Tobino e/o dell'apparato critico con le fonti documentarie su cui il testo letterario o il commento stesso si basano. In tal senso, si prospetta anche la creazione di sezioni fotografiche per la consultazione dei ritagli a stampa (soprattutto recensioni su giornale) conservati presso la Fondazione Ragghianti di Lucca e delle traduzioni straniere conservate presso la Fondazione Tobino.

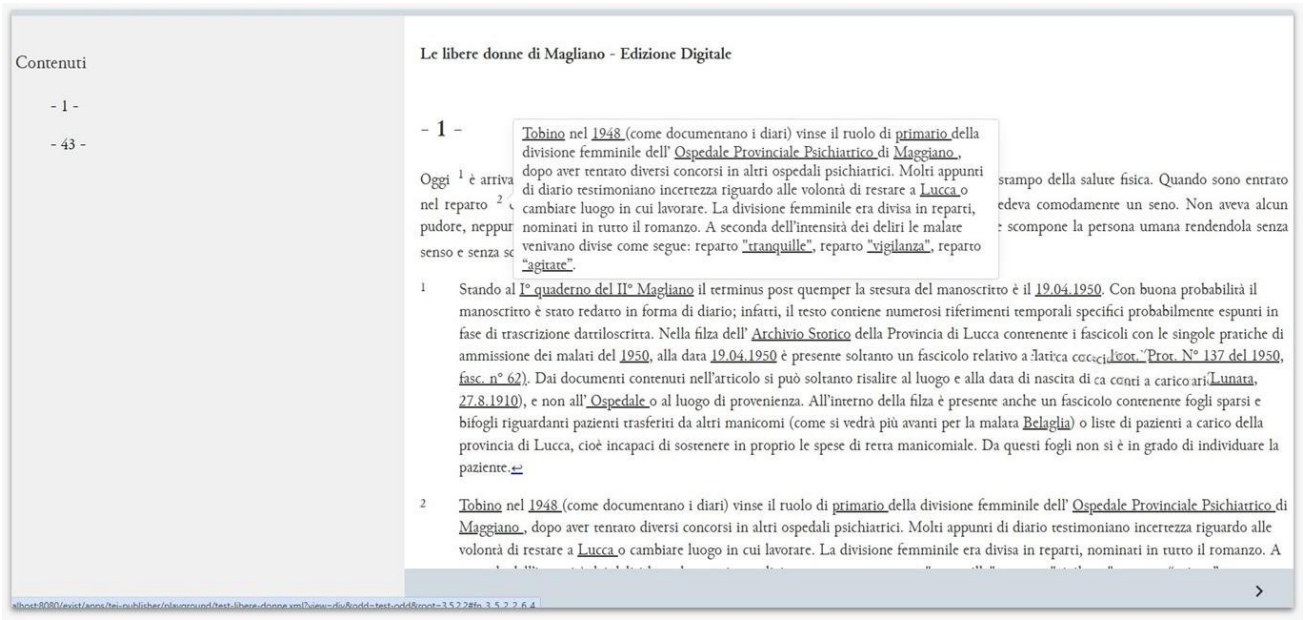


Figura 2. Esempio di visualizzazione in TEI publisher di un estratto dell'apparato (cfr. figura 1)

BIBLIOGRAFIA

Babini, V.P. (2009). Liberi tutti. Manicomi e psichiatri in Italia: una storia del Novecento. Il Mulino.

Bagnai Losacco, V. (2024). Guida pratica all'accesso e alla consultazione delle cartelle cliniche. In I. Tobino (a cura di), *Gli umani documenti. Le cartelle cliniche dell'Ospedale Psichiatrico di Maggiano*. maria pacini fazzi editore, pp. 35-52.

Burnard, L. (2014). What Is the Text Encoding Initiative?. OpenEdition Press. https://doi.org/10.4000/books.oep.426.

Chiaravallotti, M.T., Taverniti, M., Dovetto, F.M., (2023). Preserving Cultural Heritage: Digitizing the Historical Archive of the Former Psychiatric Hospital of Girifalco (South Italy) [Conference presentation] 7th IEEE Congress on Information Science and Technology (CiSt), pp. 627-633. https://ieeexplore.ieee.org/document/10409873.

Driscoll, M. J., & Pierazzo E. (Eds.). (2016). Digital scholarly editing. Theories and Practices. Lightning Source for Open Book Publishers.

Edizione digitale delle Lettere di Vespasiano da Bisticci. n.d. Accessed April 20, 2026. https://vespasianodabisticciletters.unibo.it/

Leggo Manzoni. n.d. Accessed April 20, 2026. https://projects.dharc.unibo.it/leggomanzoni/

Levchenko, M., Nava B., Russo E. (2025). TEI encoding as a Unified Structure for Multilingual Digital Editions: the Leggo Manzoni Case Study. In S. Rebora, M. Rospocher, S. Bazzaco (a cura di), *Diversità, Equità e Inclusione: Sfide e Opportunità per l'Informatica Umanistica nell'Era dell'Intelligenza Artificiale*, Atti del XIV Convegno Annuale AIUCD, Verona 11-13 giugno 2025, Università di Verona, pp. 264-269. https://amsacta.unibo.it/id/eprint/8380/1/AIUCD2025_Proceedings.pdf

Pierazzo, E. (2015). Digital Scholarly Editing: Theories, Models and Methods. Ashgate.

Riccucci, M., (a cura di). 2025. Gli psichiatri-scrittori italiani. Una storia di quasi due secoli. Pisa University Press.

Riccucci, M., Zamperini E., Di Carlo S., (2023, febbraio 10-11). Le parole per dirlo: luoghi, persone, narrazioni [Atti di Convegno]. Pisa-Maggiano, Italia. Pisa University Press.

Semantic Digital Edition of Paolo Bufalini's notebook. n.d. Accessed April 20, 2026. https://projects.dharc.unibo.it/bufalini-notebook/

Serratore, G., Portaro, A., Donato, E.N., M., Chiaravallotti, M.T., (2025). A pipeline for the digital preservation of Girifalco psychiatric hospital's medical records [Conference presentation] IEEE 8th

Congress on Information Science and Technology (CiSt), pp. 463-468. https://api.semanticscholar.org/CorpusID:282823275.

Tobino, M. (with Italia, P., & Borgna, E., & Magrini, G.). (2007). Opere scelte. Mondadori.

Tobino, M. (1953). Le libere donne di Magliano. Vallecchi.

Tobino, M. (with Italia, P.). (2024). Le libere donne di Magliano. Mondadori.

Il restauro testuale dei papiri nel progetto GreekSchools: CoPhiEditor e l'integrazione di modelli linguistici

Simone Zenzaro[1], Federico Boschetti[2], Angelo Mario Del Grosso[3], Graziano Ranocchia[4]

[1]CNR Istituto di Linguistica Computazionale "A. Zampolli", (CNR-ILC), Italia - simone.zenzaro@cnr.iT
[2]CNR Istituto di Linguistica Computazionale "A. Zampolli", (CNR-ILC), Italia - federico.boschetti@ilc.cnr.it
[3]CNR Istituto di Linguistica Computazionale "A. Zampolli", Italia - angelomario.delgrosso@cnr.it
[4]Dipartimento di Filologia, Letteratura e Linguistica, Università di Pisa, Italia - graziano.ranocchia@unipi.it

ABSTRACT (ITALIANO)

Il contributo presenta un approccio automatizzato per il restauro testuale dei papiri di Ercolano basato su modelli linguistici di tipo *Transformer*. L'integrazione dell'intelligenza artificiale nella piattaforma CoPhiEditor tramite *API RESTful* abilita un workflow collaborativo in cui i suggerimenti automatici di supplementi testuali supportano, senza sostituirla, la discussione filologica. L'approccio propone una filologia assistita da tecnologie di AI generativa che bilancia analisi computazionale e controllo editoriale umano, ponendo particolare attenzione alla sostenibilità tecnologica dell'edizione scientifica digitale.
Parole chiave: text restoration; filologia computazionale; intelligenza artificiale; edizione scientifica digitale; modelli linguistici

ABSTRACT (ENGLISH)

Textual restoration of papyri in the GreekSchools project: CoPhiEditor and the integration of linguistic models.
This paper presents an automated approach for the text restoration of the Herculaneum papyri based on Transformer-based language models. The integration of artificial intelligence into the CoPhiEditor platform via RESTful APIs enables a collaborative workflow where automatic suggestions for textual supplements support, rather than replace, philological discussion. The approach proposes a philology assisted by generative AI technologies that balances computational analysis and human editorial control, with a focus on the technological sustainability of digital scholarly editions.
Keywords: text restoration; computational philology; artificial intelligence; digital scholarly edition; language models

1. INTRODUZIONE

Lo stato frammentario dei testi papiracei comporta la scomparsa non solo materiale, ma anche concettuale e culturale del discorso dell'autore. Nel campo della filologia, le lacune, e la loro estensione, pongono un importante problema di ricostruzione del testo in cui la perdita del contesto e la natura indiretta della trasmissione rendono ogni integrazione ipotetica. Come osservato in (Berti, 2019) le tecnologie digitali non eliminano l'incertezza filologica, ma offrono nuovi strumenti per analizzarla e rappresentarla. Gli oltre 1800 papiri del corpus dei papiri di Ercolano rappresentano un caso emblematico di testi (principalmente) filosofici frammentari. Molti di questi papiri, scritti in greco antico, sono stati attribuiti al filosofo epicureo Filodemo Di Gadara e contengono la più antica storia della filosofia greca che ci è pervenuta: la *Rassegna dei Filosofi*.
In questo contesto, uno degli obiettivi del progetto ERC Advanced Grant 885222 GreekSchools, riguarda l'edizione della *Rassegna dei Filosofi* avvalendosi di tecniche non invasive per la lettura del testo nascosto e di una piattaforma software *web-based* per la filologia che permette la creazione di edizioni scientifiche digitali *open-access*. Questo contributo si innesta su (Zenzaro et al., 2025), all'interno del progetto GreekSchools. In questa fase ci siamo concentrati sul potenziamento degli strumenti di restauro testuale a disposizione del filologo, con particolare attenzione a due aspetti complementari: accuratezza e usabilità. Da un lato, abbiamo realizzato dei modelli linguistici in grado di generare proposte di supplemento al testo dei papiri, dall'altro abbiamo integrato le proposte di supplemento automatico nella sezione di discussione del testo della piattaforma CoPhiEditor che viene utilizzata dagli esperti di dominio per l'edizione dei testi. Una Application Programming Interface (API) RESTful (Fielding, 2000), funge da mediatore fra i modelli linguistici e la *user interface* (UI) di CoPhiEditor.

2. STATO DELL'ARTE

Nel contesto della filologia computazionale e dello sviluppo di strumenti per la produzione di edizioni scientifiche digitali, il restauro testuale di documenti danneggiati (*text restoration*), cioè l'attività scientifica di integrazione di lacune mediante *supplementa* testuali, sta vivendo un rinnovato impulso, trainato soprattutto dagli avanzamenti nel campo dell'Intelligenza Artificiale (Battezzato et al., 2025). La tecnologia alla base dei recenti avanzamenti è l'architettura *Transformer* (Vaswani et al., 2017), costruita su complesse reti neurali artificiali profonde (Suissa, 2022). In effetti, testi conservati su supporti degradati come papiri frammentari mal conservati, iscrizioni deteriorate, manoscritti corrosi et similia, presentano spesso perdite di testo sostanziali che gli studiosi ricostruiscono mediante un'operazione congetturale complessa, vincolata da fattori materiali, ad esempio lo spazio disponibile, le tracce paleografiche, la topografia del supporto, ma anche da vincoli linguistici e testuali quali la morfologia, la sintassi, la semantica, la formularità e i passi paralleli (Riemenschneider & Frank, 2023; Sommerschield et al., 2023).
L'IA generativa e, in particolare, i modelli linguistici pre-addestrati (Qiu et al., 2020) consentono oggi di ottenere predizioni spesso molto accurate. Tali predizioni non sostituiscono l'atto editoriale interpretativo, argomentato e di responsabilità del filologo-editore, ma possono supportarlo molto efficacemente, stimando la plausibilità delle integrazioni a partire da molteplici regolarità di varia natura apprese dai corpora e dai paralleli testuali durante la fase di addestramento del modello (Riemenschneider & Frank, 2023; Sommerschield et al., 2023; Battezzato et al., 2025). Attualmente gli approcci maggiormente adottati per il compito di text-restoration si articolano in due grandi famiglie architetturali: (i) modelli encoder-only (Rogers et al., 2020), tipicamente addestrati con *Masked Language Modeling* (MLM), e (ii) modelli encoder–decoder (*Sequence-to-Sequence* oppure *Text-to-Text*), più adatti per compiti di *text infilling* e generazione condizionata (Raffel et al., 2020; Lewis et al., 2020). La famiglia di modelli encoder-only rappresenta il cuore computazionale per il task di *fill-mask*[1] in cui una o più porzioni di testo vengono sostituite con un *token* speciale (ad es. [MASK]) e il modello propone una lista ordinata (sulla base di una metrica) di candidati plausibili per ciascuna posizione mascherata. Questo paradigma deriva direttamente dall'architettura BERT (Bidirectional Encoder Representations from *Transformers*) e dalla sua capacità di modellare il contesto bidirezionale (Devlin et al., 2019). Per le lingue classiche, la letteratura recente mostra come tale impostazione sia efficace tanto per lacune brevi (una parola o pochi morfemi) quanto come componente di strategie iterative per lacune più ampie, ed è stata adottata anche in scenari a bassa disponibilità di risorse, grazie a scelte accurate di pre-addestramento e tokenizzazione. In quest'area si collocano, ad esempio, i modelli encoder-only che includono varianti come RoBERTa (Liu et al., 2019) per il greco antico o LatinBERT (Bamman & Burns, 2020) per il latino e configurazioni multilingui greco-latino, come GrεBERTa, PhilBerta, SPhilBerta (Riemenschneider & Frank, 2023) e PhiloBerta (Allbert & Allbert, 2025). Il progetto Logion (Cowen-Breen et al., 2023) mostra l'utilità di un modello BERT addestrato su un ampio corpus greco premoderno per supportare attività filologiche "in senso lato" (rilevazione/correzione di errori, suggerimenti congetturali, ispirazione interpretativa), insistendo sul principio che l'output del modello richiede sempre interpretazione umana. In questi sistemi le strategie di tokenizzazione *subword* (Sennrich et al., 2016) sono essenziali per supportare efficacemente lingue morfologicamente molto ricche. Progetti spartiacque come Ithaca (Assael et al., 2022) dimostrano inoltre come la ricostruzione del testo può essere proficuamente integrata con altri compiti filologici quali l'attribuzione geografica e quella cronologica, configurando la predizione come un sistema di supporto allo studio del testo, e non come meccanismo sostitutivo del giudizio del filologo-editore (Assael et al., 2022). Se consideriamo, invece, la ricostruzione di lacune molto estese o di lunghezza ignota, l'approccio *fill-mask* risulta meno adatto, poiché vincola per costruzione a predefinire numero e granularità delle "maschere". In questi casi risultano più appropriati modelli encoder–decoder orientati a compiti di *text infilling* (Zhu et al., 2019; Zhang et al., 2025), alla ricostruzione cioè di segmenti di testo arbitrari. Nello specifico del greco antico e del latino, modelli come GrεTa[2] e PhilTa[3], varianti del modello T5 monolingui/multilingui, sono stati proposti proprio per gestire più efficacemente supplementi *multi-token* e integrazioni sequenziali. In questa direzione va anche il progetto Philo-1-preview (Ferrara, 2025), che si dedica anche all'emendazione e non solo al supplemento in lacuna. Un

[1] https://huggingface.co/tasks/fill-mask
[2] https://huggingface.co/bowphs/GreTa
[3] https://huggingface.co/bowphs/PhilTa

avanzamento recente è rappresentato da Aeneas[4], che affronta integrazione e contestualizzazione integrando modalità diverse (anche visive), gestendo esplicitamente lacune di lunghezza arbitraria e introducendo meccanismi di recupero da passi paralleli per informare la generazione e ridurre il rischio di supplementi plausibili ma storicamente infondati (Assael et al., 2025). Il modello sviluppato da Aeneas implementa meccanismi predittivi per *missing character* e *unknown-length segment* e adotta una architettura derivata dal modello T5. Il recupero di passi paralleli con la tecnica del *retrieval-augmented generation* (Lewis et al., 2020) per guidare la generazione del testo riduce il rischio di allucinazioni, garantendo una maggiore coerenza semantica globale. Nell'ambito del progetto GreekSchools, il task *fill-mask* rappresenta l'opzione più praticabile, perché consente di generare suggerimenti puntuali e controllabili (a livello locale) e di incorporare vincoli lessicali e morfosintattici. L'efficacia dei task descritti dipende dalla qualità dei dati di addestramento e, soprattutto, da un'adeguata formalizzazione delle lacune che rispecchi la pratica congetturale del dominio filologico. In questa direzione, il dataset MAAT (Machine-Actionable Ancient Text Corpus) si propone come risorsa di riferimento perché converte e predispone testi provenienti da edizioni digitali in un formato strutturato *machine-actionable* che preserva testo, lacune e integrazioni editoriali, generando casi di test allineati alle attività reali del filologo (Fitzgerald & Barney, 2024). Il corpus MAAT include un numero elevato di integrazioni (centinaia di migliaia), con distribuzioni di lunghezza non casuali e presenza di alternative (lezioni varianti), rendendo necessarie metriche che tengano conto di più soluzioni accettabili (ad es. valutazioni *top-k* e criteri "*best-of*" rispetto a insiemi di varianti). Nel progetto Ithaca, l'addestramento deriva da un dataset costruito a partire dalla collezione del Packard Humanities Institute (https://inscriptions.packhum.org) e utilizza strategie di *data augmentation* e mascheramento coerenti con il danno materiale, dimostrando come la qualità della simulazione delle lacune influisca direttamente sulle prestazioni e sulla credibilità delle proposte.

Affinché questi modelli siano effettivamente utili nella produzione di edizioni scientifiche digitali, non basta la disponibilità di un modello predittivo basato su IA generativa, ma occorrono ambienti integrati di editing avanzato che garantiscano (i) coerenza delle convenzioni, (ii) verificabilità e versionamento, (iii) coerenza semantica e tracciabilità delle decisioni. In questa prospettiva, l'ecosistema di scholarly editing include numerose piattaforme con punti di forza differenti: SoSOL / Papyri.info (Papyrological Editor, https://github.com/sosol/sosol) è un ambiente collaborativo per creazione e revisione di trascrizioni papirologiche in Leiden+ oppure in TEI/EpiDoc, con workflow collaborativo e pubblicazione integrata nella piattaforma di papyri.info (https://papyri.info/); Perseids (https://github.com/perseids-project) è una piattaforma per annotazione ed editing di risorse del mondo classico, orientata a contributi distribuiti e pubblicazione con versionamento di dati di ricerca; Textual Communities (https://github.com/DigitalResearchCentre/tc) è un sistema per edizioni collaborative con strumenti di collazione automatizzata (integrazione con lo strumento CollateX, https://github.com/interedition/collatex.git) e gestione del lavoro collaborativo e distribuito in formato XML/TEI; LEAF-Writer (https://github.com/LEAF-VRE) è un editor web semantico e schema-aware pensato per la comunità TEI, con supporto a schemi personalizzati e integrazione con Linked Open Data/Web Annotation; Classical Text Editor (https://cte.oeaw.ac.at/) è uno strumento proprietario orientato alla produzione tipografica di edizioni critiche (apparati, note, resa camera-ready); TEI Publisher (https://github.com/eeditiones/tei-publisher-app.git) è un framework per costruire applicazioni di pubblicazione per edizioni TEI-based e sostenere pipeline XML/TEI-to-Web in ottica di *single-source publishing* (Fauchié & Audin, 2023). Nonostante ciò, l'adozione di servizi di integrazione automatizzata all'interno di piattaforme di scholarly editing rimane limitata: il filologo tende ancora a operare con workflow frammentati (editing in un ambiente, inferenza in un altro, pubblicazione altrove), con costi elevati di trasformazione e perdita di tracciabilità. Per affrontare tale criticità, il modulo in sviluppo in CoPhiEditor integra nel ciclo di editing i task di *fill-mask* e *text-infilling*, con l'approccio DSL-Based Digital Scholarly Editions (Zenzaro et al., 2025).

3. ARCHITETTURA

In questa sezione discutiamo principalmente le scelte architetturali e di design sia dal punto di vista tecnico che da quello della *user interface* ed *experience* (UI/UX). Il primo punto di snodo nelle scelte architetturali ha riguardato la separazione delle responsabilità (Dijkstra, 1982) delle funzionalità software

[4] https://predictingthepast.com/aeneas

in modo da migliorare gli aspetti di manutenibilità, riusabilità e testabilità che incidono direttamente sullo sforzo di sostenibilità a lungo termine dello strumento software, mitigando la rapida obsolescenza tecnologica. Per questo motivo, abbiamo individuato tre componenti principali: i modelli linguistici, la UI/UX all'interno di CoPhiEditor e una API come elemento di orchestrazione delle due componenti.
Per quanto riguarda i modelli linguistici, ci siamo concentrati sull'adattamento tramite fine-tuning di modelli pre-addestrati al fine di fornire un supporto ai filologi nel compito di integrazione delle lacune testuali. Focalizzandoci sulle esigenze del progetto GreekSchools, in una prima fase abbiamo utilizzato i modelli disponibili in letteratura per il greco antico come GreBerta (Riemenschneider & Frank, 2023), AristoBERTo[5], e Logion allo scopo di produrre una lista di suggerimenti e, dopo averli testati su una selezione lacunosa dei papiri di Ercolano, abbiamo deciso di applicare una strategia di *domain adaptation* tramite *fine tuning* a partire dai dati del corpus MAAT ed estendendoli con i dati inediti delle edizioni pubblicate durante i primi anni del progetto GreekSchools, oltre che con i testi vicini per stile a quelli di interesse. Abbiamo così ottenuto i modelli `CNR-ILC/gs-aristoBERTo`, `CNR-ILC/gs-GreBerta` e `CNR-ILC/gs-Logion` che sono disponibili sulla piattaforma HuggingFace[6]. Come *baseline* di valutazione, abbiamo deciso di implementare anche due modelli a n-grammi basati su *Maximum Likelihood Estimation* (MLE) e *Lidstone Smoothing* (Jurafsky, 2009). Il modello basato su MLE offre la massima fedeltà ai dati reali senza distorsioni statistiche, mentre con lo Smoothing di Lidstone si garantisce la robustezza del modello assegnando una probabilità minima anche agli eventi mai visti, evitando di fallire nel calcolo della probabilità su testi nuovi. In ogni caso, come prevedibile, i modelli basati su n-grammi offrono un'accuratezza sensibilmente ridotta rispetto ai modelli basati su BERT. Tutti i modelli sono stati valutati su un test set costruito a partire dalle edizioni del progetto GreekSchools tramite top_k accuracy.
Abbiamo dunque introdotto un livello di astrazione tramite una *API RESTful* che, indipendentemente dal modello linguistico utilizzato, uniforma la richiesta di una lista di suggerimenti per una lacuna. In questo modo, l'architettura permette aggiornamento, aggiunta e combinazione di modelli linguistici in maniera trasparente al papirologo. I modelli linguistici disponibili sono accessibili tramite l'*endpoint* `GET /models` che restituisce una lista di oggetti JSON contenenti le informazioni sui modelli, in particolare un identificatore univoco `_id` (vedi Figura 1 a sinistra).



Figura 1. *API RESTful*: JSON dei modelli linguistici disponibili e esempio di risultato per i suggerimenti

Una volta ottenuto l'identificativo di un modello, l'*endpoint* `GET /predictions` può essere utilizzato per ottenere una lista di proposte di supplemento al testo. I parametri della chiamata sono: l'id del modello da utilizzare, il numero di predizioni, il contesto e per i soli modelli basati su n-grammi il numero di *token* da generare. Il contesto è una porzione di testo che contiene la lacuna, ad esempio ὀρθῶς [...ṭαι] in cui il testo fra parentesi quadre indica la lacuna. Ciascuna lacuna contiene un punto sublineare per ogni carattere mancante e le lettere leggibili per quel frammento di testo. Nel caso dei papiri Ercolanesi, infatti, l'ampiezza di una lacuna non è quasi mai un'incognita assoluta; essa può essere determinata con discreta approssimazione attraverso un'analisi paleografica oltre allo spazio occupato dalle singole lettere in rapporto alle tracce (o vestigia) visibili. Il risultato è una lista di oggetti (vedi Figura 1 a destra) che rappresentano il supplemento in lacuna (`token_str`), un numero che ordina i supplementi secondo la metrica scelta (`score`), il supplemento in contesto (`sentence`).

[5] https://huggingface.co/Jacobo/aristoBERTo
[6] https://huggingface.co/CNR-ILC/models

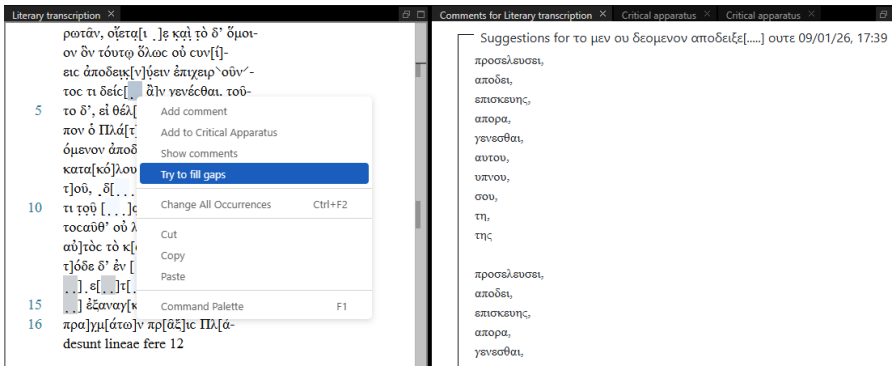


Figura 2. UI/UX per la consultazione dei supplementi in lacuna

Per quanto riguarda CoPhiEditor, è necessario progettare una soluzione di UI/UX integrata nel flusso di lavoro dei papirologi, facile da utilizzare e da consultare. I tempi di risposta dei modelli linguistici possono, in alcuni casi, essere anche lunghi (dell'ordine delle decine di secondi), questo dato ci ha fatto propendere per una soluzione asincrona rispetto alla richiesta in modo da non bloccare il lavoro di edizione. Inoltre, la sezione "Comments" dello spazio di lavoro viene utilizzata dal team di progetto per lavorare sui papiri, suggerendo modifiche e congetture al testo nella forma di *thread* di discussione. Perciò, abbiamo ritenuto opportuno applicare il meccanismo dei commenti anche agli agenti software tramite chiamate agli end-point dell'*API RESTful* per i modelli linguistici. La Figura 2 mostra due fasi della richiesta di supplementi in lacuna. A sinistra (Figura 2 a sinistra) l'utente seleziona una porzione di testo e, tramite un menù contestuale, avvia la richiesta di supplementi ("Try to fill gaps"). In Figura 2 a destra, la sezione dei commenti al testo mostra un nuovo commento generato a partire dalle risposte dei modelli linguistici. Il testo del commento generato dall'agente software sarà successivamente vagliato dal gruppo di lavoro che deciderà se accogliere o rifiutare i suggerimenti. La scelta dei modelli da utilizzare può essere configurata in base alle esigenze del progetto e indicata dal filologo durante la fase di richiesta.

4. ALCUNI ESEMPI

Si prendano in considerazione i seguenti passi lacunosi: PHerc. 1004 col. 69, 2-3 καὶ τὰ μ̣[..... / .]νείδη e PHerc. 1004 col. 85, 5-6 ὀρθῶς [....]ṭαι. Nel primo caso, la proposta d'integrazione di Janko μ̣[εγάλα ὀ]νείδη trova μεγάλα fra i primi dieci suggerimenti sia con il modello `CNR-ILC/gs-GreBerta`, sia con il modello `CNR-ILC/gs-Logion`. Il sostantivo neutro plurale ὀνείδη (*oneídē*), da ὄνειδος (*óneidos*), indica atto linguistico di rimprovero (rimprovero, biasimo, insulto) o valore sociale negativo (oltraggio, vergogna). καὶ τὰ μεγάλα ὀνείδη (*kaì tà megála oneídē*) si può quindi tradurre con "e i grandi rimproveri", dove μεγάλα, da μέγας (*mégas*: grande), qualifica la dimensione o l'intensità del sostantivo che segue. Fra i primi dieci suggerimenti compaiono altre proposte interessanti compatibili con i vincoli paleografici del papiro: μέγιστ' (*mégist'*: grandissimi), superlativo di μεγάλα, attestato in Plat. R. 344b con il nostro sostantivo: ὀνείδη ἔχει τὰ μέγιστα (*oneídē échei tà mégista*: ottiene le più grandi note di biasimo). Molto meno probabili ma non del tutto incompatibili col contesto sintattico e semantico sono i suggerimenti dell'avverbio μεταξύ (*metaxú*: in mezzo) e dell'aggettivo neutro plurale μάταια (*mátaia, da μάταιος*: vano). μεταξύ in posizione attributiva tra articolo e sostantivo (*τὰ μεταξύ ὀνείδη*) può significare "mediano, intermedio", anche se occorre soprattutto in espressioni dove indica un tempo o un luogo intermedio, più che una gradazione, come potrebbero essere dei "rimproveri intermedi", se contrapposti ipoteticamente a quelli piccoli e grandi. μάταια ὀνείδη potrebbe indicare invece dei "rimproveri vani", semanticamente accettabili nel contesto più ampio del passo, dove si discute dell'utile (συμφέρον, sumphéron). Per prendere visione del testo si veda https://viewer.greekSchools.eu/editions/col-69-pherc-1004.
Nel secondo caso, la proposta d'integrazione di Cirillo ὀρθῶς [εἴρη]ṭαι (*orthôs eírētai*: sono state dette correttamente [avverbio + terza persona singolare del perfetto indicativo medio-passivo di λέγω, légō con precedente soggetto neutro plurale]) trova εἴρηται fra i primi suggerimenti di `CNR-ILC/gs-Logion`, seguito da λέγεται (*légetai*, terza persona singolare del presente indicativo medio-passivo di λέγω). Nei primi dieci compaiono altre forme verbali semanticamente compatibili con ὀρθῶς, per quanto inadatte o molto meno compatibili con il contesto generale, come νοεῖται (*noeítai*: sono pensate) φέρεται (*féretai*:

sono portate) o δίδοται (*dídotai*: sono date). Per prendere visione del testo si usi ancora una volta il viewer: https://viewer.greekSchools.eu/editions/col-85-pherc-1004.
Anche quando conferma una congettura già formulata, il sistema propone alternative degne di attenzione che potrebbero non emergere immediatamente, consentendo di ragionare sulle scelte e di avere a disposizione soluzioni compatibili dal punto di vista semantico e paleografico.

5. CONCLUSIONI

L'integrazione di modelli linguistici avanzati nel workflow di edizione dei papiri di Ercolano segna un passaggio significativo verso una filologia digitale e computazionale sempre più interattiva, cooperativa e collaborativa, dove gli attori non sono solo filologi, ma anche agenti software esperti.
Tuttavia, il valore principale di questa ricerca non risiede esclusivamente nell'accuratezza del modello, ma nell'architettura di integrazione. L'aver separato la logica computazionale (i modelli) dall'interfaccia utente tramite un'*API RESTful* garantisce la sostenibilità e la modularità dello strumento, contrastando l'obsolescenza tecnologica. L'implementazione della UI/UX sotto forma di thread di discussione (commenti), all'interno di un ambiente software integrato, trasforma l'output dell'IA in una proposta dialogica, preservando, secondo un approccio human-in-the-loop, la centralità del filologo come decisore ultimo e responsabile scientifico dell'edizione. Questo lavoro mostra inoltre che l'intelligenza artificiale non deve essere intesa come un sostituto dell'atto congetturale, ma come un amplificatore delle capacità analitiche del filologo, capace di riempire i vuoti della tradizione superstite ed accessibile con ipotesi fondate su regolarità statistiche di varia natura altrimenti inafferrabili. Il lavoro fin qui svolto dimostra la fattibilità dell'integrazione tra strumenti di AI generativi e ambito filologico, aprendo la strada non solo al miglioramento dell'accuratezza dei modelli supportati, ma anche ad una nuova metodologia ecdotica, che impone una riflessione epistemologica all'interno della comunità DH.

RINGRAZIAMENTI

Questa ricerca ha ricevuto finanziamenti dal Consiglio Europeo della Ricerca (ERC) nell'ambito di Horizon 2020 dell'Unione Europea, Excellent Science (Accordo di sovvenzione n. 885222).

BIBLIOGRAFIA

Allbert, R. A., & Allbert, M. L. (2025). PhiloBERTA: A Transformer-Based Cross-Lingual Analysis of Greek and Latin Lexicons. https://arxiv.org/abs/2503.05265.

Assael, Y., Sommerschield, T., Cooley, A., Shillingford, B., Pavlopoulos, J., Suresh, P., Herms, B., Grayston, J., Maynard, B., Dietrich, N., Wulgaert, R., Prag, J., Mullen, A., & Mohamed, S. (2025). Contextualizing ancient texts with generative neural networks. Nature, 645(8079), 141–147. https://doi.org/10.1038/s41586-025-09292-5.

Assael, Y., Sommerschield, T., Shillingford, B., Bordbar, M., Pavlopoulos, J., Chatzipanagiotou, M., Androutsopoulos, I., Prag, J., & de Freitas, N. (2022). Restoring and attributing ancient texts using deep neural networks. Nature, 603(7900), 280–283. https://doi.org/10.1038/s41586-022-04448-z.

Battezzato, L., Graziosi, B., & Haubold, J. (A c. Di). (2025). Artificial Intelligence and Greek Philology: An Experiment. Edizioni della Normale. https://doi.org/10.2422/978-88-7642-815-9.

Bamman, D., & Burns, P. J. (2020). Latin BERT: A contextual language model for classical philology. arXiv. https://doi.org/10.48550/arXiv.2009.10053.

Berti, M. (2019). Historical fragmentary texts in the digital age. In M. Berti (Ed.), Digital classical philology: Ancient Greek and Latin in the digital revolution (pp. 257–276). De Gruyter Saur. https://doi.org/10.1515/9783110599572-015.

Fielding, R. T. (2000). Architectural styles and the design of network-based software architectures [Doctoral dissertation, University of California, Irvine]. UC Irvine Electronic Theses and Dissertations.

Cowen-Breen, C., Brooks, C., Graziosi, B., & Haubold, J. (2023). Logion: Machine-Learning Based Detection and Correction of Textual Errors in Greek Philology. In A. Anderson, S. Gordin, B. Li, Y. Liu, & M. C. Passarotti (A c. Di), Proceedings of the Ancient Language Processing Workshop (pp. 170–178). INCOMA Ltd., Shoumen, Bulgaria. https://aclanthology.org/2023.alp-1.20.

Dijkstra, E. W. (1982). On the role of scientific thought. In Selected Writings on Computing: A Personal Perspective (pp. 60–66). Springer. https://doi.org/10.1007/978-1-4612-5695-3_12.

Fauchié, A., & Audin, Y. (2023). The Importance of Single Source Publishing in Scientific Publishing. Digital Studies / Le champ numérique, 13(3), 1–16. https://doi.org/10.16995/dscn.9655.

Ferrara, G. (2025). Philo-1-preview: Un modello T5-Base per l'emendazione dei testi antichi. In S. Rebora, M. Rospocher, & S. Bazzaco (A c. Di), Diversità, Equità e Inclusione: Sfide e Opportunità per l'Informatica Umanistica nell'Era dell'Intelligenza Artificiale. Proceedings del XIV Convegno Annuale AIUCD2025 (pp. 404–410). AIUCD. https://doi.org/10.6092/unibo/amsacta/8380.

Fitzgerald, W., & Barney, J. (2024). A new machine-actionable corpus for ancient text restoration. In J. Pavlopoulos, T. Sommerschield, Y. Assael, S. Gordin, K. Cho, M. Passarotti, R. Sprugnoli, Y. Liu, B. Li, & A. Anderson (A c. Di), Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024) (pp. 56–60). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.ml4al-1.7.

Jurafsky, D., & Martin, J. H. (2009). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (2nd ed.). Prentice Hall.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (A c. Di), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 7871–7880). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.703.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2021). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. https://arxiv.org/abs/2005.11401.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint. https://arxiv.org/abs/1907.11692.

Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., & Huang, X. (2020). Pre-trained Models for Natural Language Processing: A Survey. arXiv preprint. https://arxiv.org/abs/2003.08271.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Journal of Machine Learning Research, 21(140), 1–67.

Riemenschneider, F., & Frank, A. (2023). Exploring large language models for classical philology. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 15181–15199). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.846.

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A Primer in BERTology: What We Know About How BERT Works. Transactions of the Association for Computational Linguistics, 8, 842–866. https://doi.org/10.1162/tacl_a_00349.

Sennrich, R., Haddow, B., & Birch, A. (2016). Neural Machine Translation of Rare Words with Subword Units. In K. Erk & N. A. Smith (A c. Di), Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1715–1725). Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1162.

Sommerschield, T., Assael, Y., Pavlopoulos, J., Stefanak, V., Senior, A., Dyer, C., Bodel, J., Prag, J., Androutsopoulos, I., & de Freitas, N. (2023). Machine Learning for Ancient Languages: A Survey. Computational Linguistics, 49(3), 703–747. https://doi.org/10.1162/coli_a_00481.

Suissa, O., Elmalech, A., & Zhitomirsky-Geffet, M. (2022). Text analysis using deep neural networks in digital humanities and information science. Journal of the Association for Information Science and Technology, 73(2), 268–287. https://doi.org/10.1002/asi.24544.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All You Need. Advances in Neural Information Processing Systems, 30.

Zenzaro, S., Del Grosso, A. M., Boschetti, F., & Ranocchia, G. (2025). CoPhiEditor: The DSL-Based DSE Methodology within the ERC Advanced Grant 885222-GreekSchools. Umanistica Digitale, 9(20), 31–56. https://doi.org/10.6092/issn.2532-8816/21231.

Zhang, A., Sivakumar, A., Tang, C.-W., & Thomas, C. (2025). Flexible-length Text Infilling for Discrete Diffusion Models. Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, 31344–31359. https://doi.org/10.18653/v1/2025.emnlp-main.1597.

Zhu, W., Hu, Z., & Xing, E. (2019). Text Infilling. https://arxiv.org/abs/1901.00158

1AIUCD0
102026
1
0101100

UNICA
UNIVERSITÀ
DEGLI STUDI
DI CAGLIARI

D.H.
UNICA.IT
Centro interdipartimentale per
l'Umanistica Digitale

ASSOCIAZIONE per
l'INFORMATICA UMANISTICA
e la CULTURA DIGITALE