The FAITH project: integrated tools and methodologies for digital humanities

Alfio Ferrara, Sergio Picascia, Elisabetta Rocchetti, Gaia Varese

Abstract Integration of many different sources and expertise is a key factor to solve specific research problems, especially in areas such as social sciences. The FAITH (Fight Against Injustice Through Humanities) project's main objective is to provide common tools and methodology for the collection, digitization and integration of different historical sources. In particular, the proposed solution involves the employment of an unique meta-model gathering information from different artifacts. Moreover, the FAITH project aims at applying data analysis techniques (e.g. bayesian networks, natural language processing techniques, image processing techniques) to provide insights about social issues in a diachronic perspective.
Abstract *Abstract in Italian*

Key words: digital humanities, data integration, data analysis.

1 Introduction to the FAITH Project

Social sciences are characterized by the need of an interplay between different expertise to support research activities. In particular, anthropological studies lean on many types of historical artifacts such as ancient manuscripts and skeletons. In such cases, integration between different sources is fundamental, but its technical implementation could represent an obstacle hindering research activities, especially when there is a lack of data management skills. Moreover, each discipline has its own methodology to tackle research problems: this approach naturally leads to the separation among studies carried by different specializations.
In order to provide common tools and techniques to approach anthropological research scenarios, the FAITH (Fighting Against Injustice Through Humanities) project aims at providing a methodology for the collection, digitization and integration of sources that can allow scientists to address crucial social issues through a

Università degli Studi di Milano, e-mail: name.surname@unimi.it

real interdisciplinary research. In particular, the project focuses on the extrapolation and digitization of archaeological, anthropological, medical, genetic, environmental, geological, documentary, literary, legal and artistic data, across different historical periods (Roman, Middle Ages, modern and contemporary), in a diachronic perspective. To show the potential and usefulness of such approach, the FAITH project is developing a case study considering signs of violence, abuse and discrimination in Milan in the XIII century.

2 Collecting and Integrating Data in the Humanities

Ease the interaction between researchers of different fields and allowing them to have an exhaustive perspective on the historical period of interest, are the main reasons for having a central storage where all the data converges. Unfortunately, these circumstances rarely arise in practise, especially in the humanities, where a considerable amount of resources and time is required in order to perform even simple analysis. This is mainly due to the difficulties encountered in collecting and integrating such kind of data.
When it comes to collecting data, there are two main procedures followed, depending on the format they are in: usually the data acquisition process is still performed with analog tools or, when the data is already digitized, it is organized following a very rigid and domain oriented schema, which makes the integration with other data sources difficult. We are in strict collaboration with experts of the different fields in order to assist them in digitizing already collected information and providing them specific tools in order to facilitate the acquisition of new data.
The fundamental step performed towards the accomplishment of these goals has been the definition of a flexible and general meta-model, capable of adapting and supporting the heterogeneity that distinguishes data in humanities. The model revolves around the idea of an abstract *ENTITY*, which then materializes in one of the following main sub-classes:

- *SOURCE*: it represents an historical data source from which are extracted all the pieces of information described in the database. Apart from the title, we are interested in characteristics, such the conservation status, the completeness or the material of which it is made;
- *EVENT*: it represents an historical event in which other entities appear. It can be described by a short summary and include a degree of certainty;
- *PERSON*: it is used to record data regarding people;
- *PLACE*: it is used to record data regarding geographical places, with the possibility of specifying also their coordinates;
- *ANIMAL*: it is used to record data regarding animals;
- *INSTITUTION*: it is used to record data regarding institutions;
- *OBJECT*: it is used to record data regarding objects;
- *APPELLATION*: it is used to register the names (appellations) of other entities, namely: *PLACE*, *PERSON*, *ANIMAL*, *INSTITUTION* and *OBJECT*;

- *GROUP*: it used to define a set of entities showing common characteristics.

Each of the previous entities can be in relationship with any other entity, even one of the same class. These relationships are shaped using the *TYPE* entity, determining which is the kind of relationship between the objects, together with *PLACE* and *TIME SPAN* entities, specifying where and when the relationship holds. Entities can be associated with one or more attachments, defined by the *ATTACHED* entity, that can be represented by any kind of media file, such as images, video or audio. Finally, it is possible to report the measurement of an aspects of an entity using a *METRIC*, which is associated to the value it assumes and its error, recorded in a certain *UNIT OF MEASURE*.

A practical example is shown in Figure 1. One textual *SOURCE* we imported in the database is the *Liber Mortuorum*, which contains records of dead people in the area of Milan around the XV century. A single record from this source, the 'Death of a PERSON', is inserted as an *EVENT* in which the dead person appears as an instance of the *PERSON* entity. The dead *PERSON* has her name recorded using an *APPELLATION* (i.e. *Petra*), her biological sex registered using the corresponding *GROUP* (i.e. *Female*) of a specific *TYPE* (i.e. *Biological Sex*) and her age measured with the *METRIC* 'Age' using 'Years' as *UNIT OF MEASURE*. The place in which the event takes place is registered as an instance of the *PLACE* entity having its own APPELLATION (i.e., S. Bartolomeo intus), while the date in which it happens is recorded with a *TIME SPAN*.

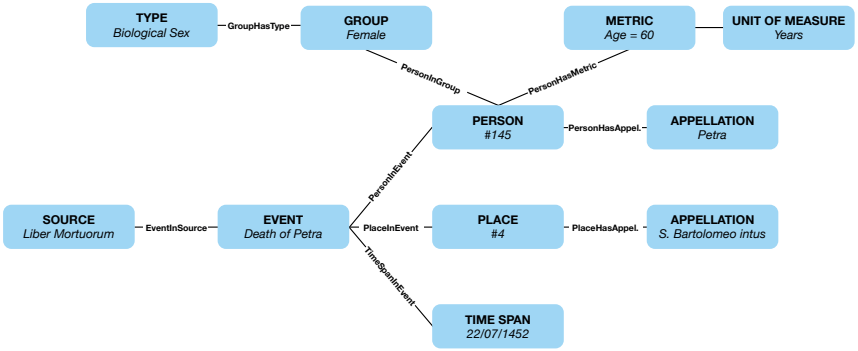

Fig. 1 The FAITH meta-model in practice: an example taken from the *Liber Mortuorum*.

3 Analysis Methods for the Digital Humanities

The application of data analysis techniques to digital humanities has a dual objective in our scenario: first of all, it allows experts to have an immediate high-level glimpse at the data at disposal, performing simple aggregation queries; on the other hand, complex statistical models can be built in order to model the intrinsic domain knowledge. Insights resulting from this last approach may also be included in the database, enriching even more the current available information.
Given the data at our disposal, we chose to focus mainly on probabilistic approaches. Indeed, our attention has been directed at integrating different data sources regarding the lifestyle of people in the area of Milan in the Late Middle Ages, in the attempt to reconstruct this specific historical context. We are trying to achieve this result employing Bayesian Networks, probabilistic graphical models that are able to exploit both tabular data and experts' knowledge in order to represent the information available. They are extremely effective at modeling the relationships elapsing between variables, computing the likelihood of certain scenarios, i.e. combination of evidences, and studying phenomena evolving over time with dynamic BNs [1].
They are particularly suitable for our task since the number of available tabular data is, at time, not sufficiently large for building an exhaustive network. BNs allow to support the phases of structure and parameter learning from data with the help of experts' knowledge: it is possible to define white and black lists in order to draw or avoid drawing edges between nodes, i.e. allowing for or denying dependencies relationships between variables; it is also possible to operate on the parameters, asking experts to estimate the values of conditional probabilities.
The expected outcome is a BN that allows us to generate 'what if' scenarios; in particular, we are in strict collaboration with anthropologists, studying how living conditions affected the modality with which people died at the time. This kind of analysis is interesting not only for inference purposes, but also in order to compare different time periods: for instance, we are currently observing how violent deaths distribute in Late Middle Ages with respect to now.

4 Ongoing Work and Future Prospects

At the present time, the FAITH project has built two data sources including, comprehensively:

- more than 40000 records representing events, people and relations between people occurred between 1452 and 1485 in Milan extracted from the *Liber Mortuorum*; specifically, these tuples record diseases (e.g. person X's disease), traumatic events (e.g. person Y is found dead at a specific place and time) and relationships (e.g. person X is married with person Y);
- 600 events from the *Liber sententiarum potestatis Mediolani* [2], which was written in 1385, and reports the criminal sentences pronounced by the chief magis-

trate of Milan Carlo Zen during his mandate. Each tuple stores information about different crimes decorated with the respective type of crime, people involved and type of condemnation or absolution.

We are currently working on an automated methodology to gather data from different sources and to organize it in the proposed meta-model. An experiment has been performed on partially structured data (e.g. Excel spreadsheets) to build the first mentioned data collection.
Based on the sources currently gathered, we have identified some research activities that will be pursued in the following months:

- we are going to employ probabilistic models, such as Bayesian Networks, in order to study the information recorded in the *Liber Mortuorum*. We would like to study how variables regarding personal data may effect the cause of death of an individual, and compare these relationships with the ones extracted analysing a dataset of a different time period;
- NLP techniques may be applied on the events extracted from the *Liber sententiarum potestatis Mediolani*: also in this case, a temporal comparison with data from another time period can be made, considering, for example, the different type of crimes committed and the characteristics of the people involved in them;
- finally, when it comes to image processing, we are considering the possibility of applying CNNs to medical images, such as CT scans and RMI, with the objective of developing unsupervised strategies for clustering and retrieving images.

References

1. Ghahramani, Z. (2006). Learning dynamic Bayesian networks. Adaptive Processing of Sequences and Data Structures: International Summer School on Neural Networks “ER Caianiello” Vietri sul Mare, Salerno, Italy September 6–13, 1997 Tutorial Lectures, 168-197.
2. Milano, Archivio Storico Civico e Biblioteca Trivulziana, Cimeli, 146.