

Francesco Tissoni

Pour un corpus numérique comparatiste des traductions d'Homère

Avertissement

Le contenu de ce site relève de la législation française sur la propriété intellectuelle et est la propriété exclusive de l'éditeur.

Les œuvres figurant sur ce site peuvent être consultées et reproduites sur un support papier ou numérique sous réserve qu'elles soient strictement réservées à un usage soit personnel, soit scientifique ou pédagogique excluant toute exploitation commerciale. La reproduction devra obligatoirement mentionner l'éditeur, le nom de la revue, l'auteur et la référence du document.

Toute autre reproduction est interdite sauf accord préalable de l'éditeur, en dehors des cas prévus par la législation en vigueur en France.

revues.org

Revues.org est un portail de revues en sciences humaines et sociales développé par le Cléo, Centre pour l'édition électronique ouverte (CNRS, EHESS, UP, UAPV).

Référence électronique

Francesco Tissoni, « Pour un corpus numérique comparatiste des traductions d'Homère », *Corpus Eve* [En ligne], Homère en Europe à la Renaissance. Traductions et réécritures, mis en ligne le 31 décembre 2015, consulté le 02 janvier 2016. URL : <http://eve.revues.org/1273>

Éditeur : Université de Savoie

<http://eve.revues.org>

<http://www.revues.org>

Document accessible en ligne sur :

<http://eve.revues.org/1273>

Document généré automatiquement le 02 janvier 2016.

© Tous droits réservés

Francesco Tissoni

Pour un corpus numérique comparatiste des traductions d'Homère

Les raisons de la proposition

- 1 Dans le domaine des études consacrées à l'histoire des textes classiques, et plus généralement à l'histoire de la culture, les traductions d'Homère jouent un rôle fondateur. Elles permettent d'abord la diffusion des textes homériques auprès d'un public de non spécialistes ; ensuite, elles influencent des générations de poètes et d'hommes de lettres qui reprennent souvent des formules, des métaphores ou des épisodes homériques à travers les mots des traducteurs. Pour ces deux niveaux, la réalisation d'un *corpus* numérique consacré aux traductions homériques se révèle fort utile, puisque le *corpus* rend disponibles des textes souvent rares, qui appartiennent à des traditions linguistiques et littéraires différentes et dont la consultation et la lecture ne sont pas toujours évidentes pour des spécialistes intéressés à Homère qui, normalement, travaillent dans un seul domaine linguistique.
- 2 Un *corpus* numérique des traductions d'Homère devient, cependant, incontournable à un troisième niveau, celui qui pourrait permettre de comprendre, à travers une analyse comparatiste, de quelle façon des traducteurs de différents pays et de différentes époques ont interprété, chacun dans son contexte, des concepts homériques¹. Dans ce cas, la traduction n'est plus un phénomène littéraire ou de vulgarisation, mais elle devient un puissant instrument d'analyse de la société dans laquelle le traducteur vivait et de sa culture.
- 3 Dans ce dernier cas, le *corpus* s'élève au niveau d'un outil de recherche dans la mesure où il pourrait permettre de retrouver et de comparer, à partir du terme grec, ses traductions multiformes et ses interprétations, à l'intérieur de toutes les traductions considérées.

Aspects théoriques et techniques : exemples

- 4 D'un point de vue technique et organisationnel, un *corpus* numérique pourrait être conçu de plusieurs manières. Afin de clarifier ce concept, je donnerai ici des exemples que j'examinerai en détail : les collections de livres numériques (Bibliothèque numérique), la base de données de textes et le codage sémantique. Il faut préciser que le choix ecodotique au niveau numérique a des répercussions importantes, même dans le domaine d'une recherche en sciences humaines : comme c'est toujours le cas, la qualité de l'outil influence en effet inévitablement la qualité de la recherche.

Collection de livres numériques : bibliothèque numérique

- 5 Par collection de livres numériques², nous voulons indiquer la reproduction en format numérique d'une série d'ouvrages choisis sur la base de critères cohérents, comprenant des métadonnées techniques et descriptives.
- 6 Du point de vue technique une collection de livres numériques présente des caractéristiques précises. Pour les décrire au mieux, nous prendrons en considération l'exemple d'un projet déjà réalisé et présent sur le web, celui de la Bibliothèque numérique Beic (BeicDL : Beic Digital Library), en introduisant quelques remarques³.

Le projet BeicDL – Beic Digital Library

- 7 Préalablement à la réalisation concrète de la BeicDL, il a été décidé que tout composant serait conforme à des standards internationaux garantissant la facilité de gestion dans le temps et l'interopérabilité⁴.
- 8 Au-delà des avantages pratiques immédiatement identifiables (qui vont de l'automatisation des contrôles à une simplification dans la production et la gestion des données), il y en a d'autres, qui se révèlent payants sur le moyen et le long terme : l'adoption des standards représente, à ce jour, la meilleure « assurance-vie » de n'importe quel système numérique puisqu'elle en garantit la survie dans le temps.

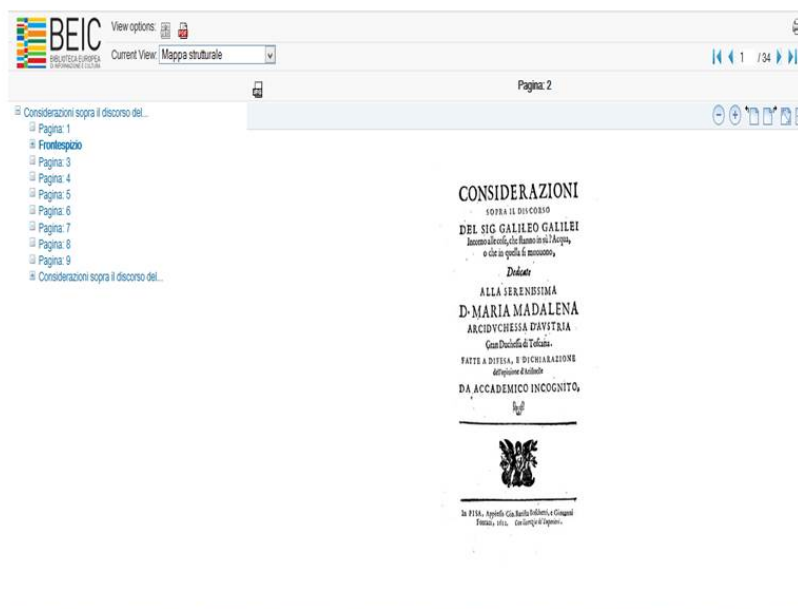
- 9 Du point de vue structurel, une bibliothèque numérique consiste en une architecture complexe, composée de plusieurs modules⁵ : une archive des collections numériques (DAMS : Digital Assets Management System), une application qui permet la gestion des collections numériques, un système de métadation, un outil de découverte comprenant un *viewer*⁶.
- 10 L'archivage des collections numériques est une infrastructure technologique qui a pour fonction de conserver les matériels dans le temps, en garantissant leur parfait entretien⁷. Elle peut être composée de plusieurs applications : l'une capable de vérifier la qualité des matériels avant leur archivage, l'autre capable de convertir les images du format TIFF à un format adapté à l'utilisation sur Internet (Jpeg o Jpeg 2000), de relier dans le même fichier PDF les différentes images qui relèvent d'un seul document, et d'effectuer la reconnaissance optique des caractères tant des images que du fichier PDF⁸ :





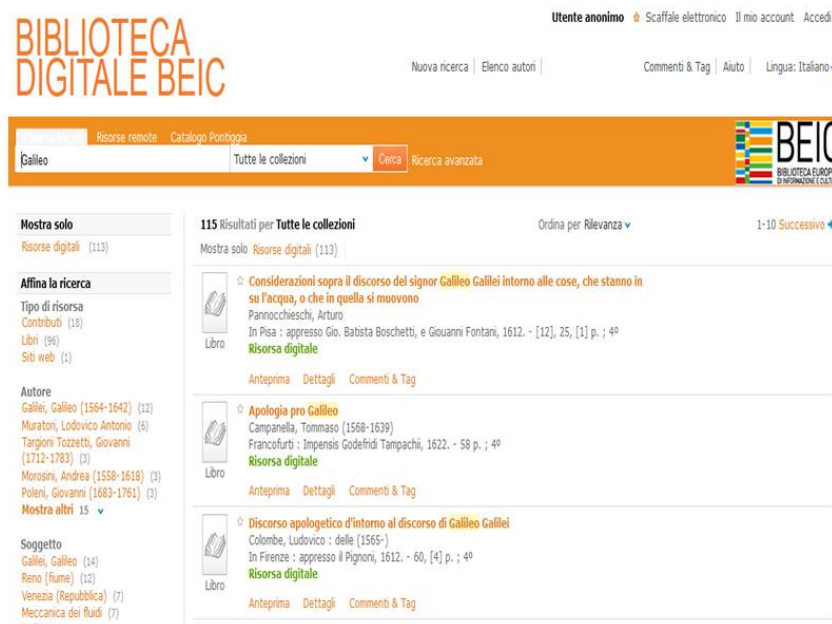
La Bibliothèque numérique BEIC : <http://www.beic.it>. Le catalogue numérique : <http://www.beic.it/it/articoli/biblioteca-digitale>

- 11 L'application utilisée par la BeicDL pour la gestion des collections numériques est DigiTool d'Ex Libris⁹. DigiTool exige que les images relatives à un document soient regroupées et accompagnées, pour la publication, d'un fichier XML conforme au schéma METS, qui contient les métadonnées. Au sens propre, les métadonnées sont des ressources d'informations sur les données, c'est-à-dire des données concernant les données (data about data) et devraient garantir les trois fonctions de base assurées par tout système informatique : identification, utilisation et conservation des contenus numériques¹⁰. Pour le contrôle complet d'une ressource numérique, il est nécessaire de disposer de métadonnées différentes qui peuvent être réparties en grandes catégories : métadonnées administratives, descriptives, structurelles, techniques et d'utilisation. Ces différentes typologies de métadonnées sont présentes dans le schéma XML METS¹¹.
- 12 Le fichier METS est divisé en sections, dont les plus importantes sont celles qui contiennent les enregistrements bibliographiques (*Descriptive metadata section*, introduite par l'élément <dmdSec>), et la description de la structure physique et logique du document numérisé (*Structural map section*, introduite par l'élément <structMap>).
- 13 La première section présente les enregistrements bibliographiques tirés d'un autre des modules qui composent le système d'une bibliothèque numérique, et de la BeicDL en particulier, celui consacré au catalogage des documents. Pour ce module BeicDL utilise une application *open source* : Koha, basée sur le standard MARC¹².
- 14 Dans la deuxième section du fichier METS est insérée la description de la structure logique et physique du document numérisé, tirée également d'un des modules qui composent le système de la bibliothèque numérique. Dans cette deuxième section du fichier – outre les indications concernant la couverture, les pages de garde, les tables et toutes les subdivisions logiques du document –, on signale aussi la présence d'*ex-libris* et de marques de possession et on indique fidèlement la numérotation des pages.
- 15 En outre, BeicDL met à la disposition de ses lecteurs une fonctionnalité supplémentaire : la reconnaissance optique des caractères (OCR)¹³.
- 16 Une fois que les documents sont publiés à l'intérieur de l'archive numérique, on peut les afficher à travers un lecteur (*viewer*). On a réalisé un lecteur spécifique pour chacune des typologies de matériels :



Exemple d'une édition numérisée sur la BEIC en format image avec la table des matières détaillée à gauche.

- 17 Un autre module fondamental est l'outil de découverte, qui permet de mettre en valeur, au maximum, les données d'un catalogage de qualité¹⁴. On devra prêter une attention particulière à la présentation des données et à la possibilité d'affiner la recherche au moyen de filtres convenables :



Exemple de classification à facettes sur la BEIC.

- 18 Pour faciliter le travail des chercheurs et des utilisateurs en général, on a disposé un espace personnel où l'utilisateur peut conserver et organiser les résultats de sa session de travail ; cette fonction lui offre la possibilité de conserver des documents considérés comme importants, d'organiser les objets en dossiers, d'ajouter des commentaires, de préserver les critères de

recherche pour chaque section, de personnaliser la présentation des données et de gérer son profil personnel.

- 19 Ce qui, plus que tout, prouve l'utilité d'avoir recours à des standards est l'interopérabilité, c'est-à-dire la capacité d'un système informatique d'interagir avec d'autres systèmes, en offrant un partage des données. L'interopérabilité n'est pas garantie par un module séparé, mais par une série de configurations de l'application qui gère les collections numériques.
- 20 Tout d'abord, on a assigné à toutes les entités numériques (la somme des images et du fichier XML auquel elles sont associées) un identifiant unique et persistant à travers le système *Handle*¹⁵.
- 21 Ensuite, l'activation du module OAI-PMH a rendu possible la collecte des métadonnées de la part d'autres institutions. OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting ou Protocole pour la collecte des métadonnées de l'Open Archive Initiative) est un protocole développé par l'Open Archive Initiative comme infrastructure de communication pour le libre accès (*open access*) et il est utilisé pour collecter les métadonnées dans une archive afin que les services puissent être construits en employant des métadonnées provenant de plusieurs archives¹⁶.
- 22 Le résultat de cette activité est la présence des enregistrements bibliographiques de la BeicDL en Incunabola Short Title Catalogue, Gesamtkatalog der Wiedrucke, The European Library et en Europeana.
- 23 Le choix de la BeicDL de numériser exactement les livres (avec toutes les images, même des feuilles blanches, et un catalogage soigné) se présente comme ecdotiquement neutre, mais pleinement conforme aux fonctions d'une bibliothèque : la tâche d'un bibliothécaire et, par conséquent, d'une bibliothèque numérique, consiste à offrir au public des reproductions numériques de haute qualité des livres, non pas celle d'utiliser la technologie pour faciliter l'étude du texte.

Bases de données de textes

- 24 Tandis qu'une collection de livres numériques met l'accent sur la reproduction numérique des livres tels quels, en restituant les images, en préservant la forme du livre et éventuellement en rendant possible la recherche du texte à travers un OCR, les bases de données de textes obéissent à une logique différente. L'objectif principal, en effet, est la numérisation d'une série de textes, de manière à rendre possible la recherche du *corpus* et à favoriser l'accès des chercheurs au texte.
- 25 Reproduire dans un *corpus* numérisé une série de textes à partir de l'édition imprimée pose des problèmes ecdotiques considérables, qui ont des implications importantes dans l'utilisation du *corpus*. Pour mieux les comprendre, il est utile de considérer quelques exemples.
- 26 La première base de données de textes littéraires a été le TLG, Thesaurus Linguae Graecae, conçu en 1972 par Marianne McDonald et David Packard, qui avaient eu l'intuition de créer une base de données de textes grecs de l'époque archaïque et classique (à l'origine, depuis Homère jusqu'à 200 après J.-C.). L'idée était tout à fait novatrice et posait de nombreux problèmes qui n'avaient jamais été abordés auparavant : le codage des caractères grecs, par exemple, constituait, à lui seul, un problème compliqué, vu que le système Unicode¹⁷ – par lequel on peut actuellement représenter plus de cent mille caractères différents, parmi lesquels les caractères grecs avec accents et esprits – n'était pas encore disponible¹⁸ ; en outre, comme le TLG était le premier exemple de ce genre, McDonald, Packard et le directeur de l'entreprise, le professeur Theodore F. Brunner, ont été contraints d'élaborer spécifiquement des outils *hardware* et *software* intégrés, dénommés *Ibycus System*¹⁹.
- 27 Ce qui depuis le début constitue le trait distinctif du TLG est l'idée de reproduire numériquement le texte d'éditions critiques faisant autorité, choisies par un comité scientifique, et consultables à travers un moteur de recherche.
- 28 La diffusion grandissante du TLG – qui demeure encore aujourd'hui un point de référence pour qui souhaite réaliser une base de données de textes numériques dans le domaine d'études des antiquisants – n'a donné naissance à aucun débat de caractère critique ; et même chez

ses utilisateurs les plus fidèles, on observe une étrange indifférence quant aux critères avec lesquels l'outil a été réalisé²⁰.

29 Apparemment, l'idée d'introduire dans la base de données une « image électronique spéculaire » de toutes les éditions imprimées utilisées en tant que modèles (« *each TLG text should be an electronic mirror image of the source edition from which it derives* ») semble irréprochable, mais en réalité pose des problèmes au niveau de la recherche et de la récupération des informations.

30 Le premier problème concerne l'orthographe des textes grecs : d'une part, il existe des différences dialectales qui intéressent les textes archaïques et classiques de toute époque ; de l'autre, il n'y a pas de définition scientifique concernant l'orthographe des textes d'époque tardive et byzantine. On sait que ces textes présentent des oscillations significatives par rapport à la norme. Mais, en l'absence d'une étude systématique, chaque éditeur se comporte comme il l'entend. Les éditeurs du TLG ont sous-estimé les effets du problème et si, dans la première édition de 1972, cela était parfaitement compréhensible, le choix de laisser le système inchangé aujourd'hui suscite bien des perplexités.

31 Le deuxième problème, peut-être plus important encore, est la présence de doublons dans la base de données, doublons qui peuvent consister en variantes graphiques du même mot²¹, portions identiques de textes²², fantômes lexicaux²³ et même œuvres entières présentes deux fois. Ce dernier cas est particulièrement grave. Il suffit de parcourir la liste des poètes de l'*Anthologie Palatine* pour s'apercevoir que la plupart d'entre eux apparaissent dans le TLG à deux reprises et dans des éditions différentes, si bien que le texte peut avoir des variantes parfois significatives : les épigrammes de Callimaque, par exemple, apparaissent à la fois dans l'édition Beckby de l'*Anthologie Grecque* et dans l'édition des œuvres de Callimaque de Pfeiffer.

32 L'exemple du TLG apparaît très instructif : d'un côté, il montre que des critères ecdotiques élémentaires (le choix de reproduire tout court dans la base de données les meilleures éditions) peuvent ne pas être nécessairement justes ; de l'autre, il enseigne que l'extraction des informations correctes d'un *corpus* textuel exige des outils beaucoup plus sophistiqués que la simple recherche en plein texte²⁴.

Library of Latin Texts

33 Des critères ecdotiques tout à fait différents ont été adoptés par la Library of Latin Text (LLT-A et LLT-B), née à la moitié des années 80, dans le cadre des activités du Cetedoc – Centre de Traitement Electronique des documents (maintenant CTLO – Centre Traditio Litterarum Occidentium) – de l'Université catholique de Louvain, sous la direction du professeur Paul Tombeur. La base de données contient un choix d'œuvres latines non seulement classiques, mais aussi archaïques, du Moyen Âge, de l'Humanisme, de la Renaissance, ainsi que modernes, qui vont de Livius Andronicus jusqu'aux documents du Concile Vatican II (1962-1965).

34 La méthodologie avec laquelle la base de données a été réalisée semble inspirée par des intentions très précises : non plus la simple reproduction de copies spéculaires des éditions imprimées, mais plutôt un véritable travail d'édition philologique et numérique, dans le respect des spécificités de l'informatique et de la philologie, avec l'objectif de réaliser un produit original où les textes soient soumis à un sévère contrôle philologique avant d'être publiés²⁵. Selon ce qu'on lit dans les premières pages de la préface, les textes compris dans la base de données ont été soumis à un rigide examen critique, qui comporte plusieurs phases.

35 1. La Library of Latin Texts veut représenter l'état actuel des études. Les textes, authentiques ou pas, d'attribution incertaine ou bien apocryphes, ont été considérés en tant que tels, en relation aux connaissances philologiques actuelles, non pas à celles des éditeurs respectifs. Il peut donc arriver, qu'un texte qui, dans la Patrologie de Migne, est attribué à un certain auteur, apparaisse comme d'attribution incertaine ou apocryphe à l'intérieur du LLT qui, pourtant, dans ce cas, reproduit précisément l'édition de Migne.

36 2. Les œuvres ont été examinées dans leur ensemble, qui comprend aussi le titre, l'*incipit* et l'*explicit*. L'objectif est celui de donner une datation textuelle certaine à chaque vocable : par

exemple, il n'est pas sans importance de savoir que le mot *apologeticum*, titre d'une œuvre de Tertullien, est postérieure à Tertullien, ou bien que le mot *Consolatio* n'apparaît jamais dans l'œuvre de Boèce, l'auteur du *De Consolatione Philosophiae*.

- 37 3. Dans la Library of Latin Texts, l'unité minimale est la forma. Par forma, on ne veut pas indiquer un ensemble de caractères, séparé par des blancs, comme c'est le cas dans d'autres bases de données, mais une unité lexicale. En d'autres termes, LLT distingue les enclitiques : *virumque*, par conséquent, est traité comme s'il était composé de deux *formæ* distinctes : *virum* et, justement, *que*. Ce procédé, qui dans la linguistique computationnelle est connu sous le nom de « tokenisation »²⁶, dans le cas de la langue latine peut être complexe et risqué.
- 38 4. L'intervention ecdotique ne se limite pas à la tokenisation – qui pourrait être considérée comme un compromis nécessaire pour le traitement des données ; au contraire, elle entre dans les plis du texte et finit par créer des conflits évidents avec l'activité de l'éditeur critique proprement dit. En effet, dans plusieurs cas, le comité éditorial de la Library of Latin Texts déclare avoir corrigé des erreurs, rassemblées sous une rubrique spécifique, appelée *Corrigenda*. L'opération, en soi discutable, a été conduite dans le respect de la philologie, en consultant, quand il était possible, les éditeurs des textes en format papier. Il est toutefois évident qu'une telle attitude comporte une marge élevée d'arbitraire : le statut de l'éditeur numérique, de simple reproducteur, à l'intérieur d'un nouvel outil, de textes déjà existants devient concurrentiel avec celui du philologue-éditeur, qui a pourtant réalisé l'exemplaire d'où le texte numérique dérive.
- 39 5. Des séquences de *formæ* ont été regroupées en unités plus complexes dites *sententiae* : si, dans la Bible, une *sententia* correspond *grosso modo* à un verset, pour d'autres œuvres en prose la *sententia* correspond à une unité textuelle ayant un sens complet. Dans le cas des œuvres poétiques, les *sententiae* regroupent, généralement, une série de vers. Dans les premières éditions de la Latin Library du Cetedoc, il était impossible de retrouver un passage en utilisant des critères habituels (par exemple livre, poème, vers, dans les cas des *Carmina* d'Horace), mais l'on devait procéder en indiquant le numéro de la *sententia*. Ce problème a été par la suite corrigé.
- 40 En ce qui concerne les auteurs et les titres, la Library of Latin Texts présente une forme normalisée, pour éviter que des œuvres ayant un contenu identique n'apparaissent classées sous des titres différents : par exemple, une œuvre qui présente un commentaire *in Mathæum* sera classée avec d'autres œuvres *in Matthæum*; tandis qu'un titre pour lequel l'éditeur de la version originale imprimée avait privilégié la forme *expositio* deviendra *expositio*. De la même manière, les éditeurs numériques ont essayé d'éviter la répétition du sigle « Anonyme » : quand il a été possible, l'œuvre a été inventoriée soit selon le *corpus* auquel elle appartient (par ex. *Scriptores ordinis Grandimontensis*), soit selon l'auteur auquel elle avait été autrefois attribuée (par ex. Pseudo-).
- 41 Une autre innovation substantielle est le soin avec lequel on a évité les doublons : la Library of Latin Texts évite de reproduire deux fois le même texte publié dans des volumes différents. En outre – s'il y a des différentes versions d'un même texte (comme dans le cas des *Discours* de Léon I^{er} le Grand) – on trouve un espace dédié, défini comme *Background to the text*, où l'utilisateur est clairement informé du problème, afin d'éviter qu'une recherche lexicale avec des finalités statistiques ne soit gravement faussée.
- 42 En ce qui concerne les *Formæ*, comme on l'a dit, la Library of Latin Texts prend soin de distinguer les unités textuelles autonomes, qu'elle traite plus ou moins comme des articles d'un dictionnaire. L'opération est loin d'être simple et pas seulement d'un point de vue informatique : par exemple, dans le cas d'un mot comme *suave*, on ne peut pas savoir *a priori* s'il s'agit du neutre de l'adjectif *suavis* ou bien d'un mot composé de l'adjectif possessif *sua* et de l'enclitique *ve*. Après des années de tentatives, les efforts du comité éditorial de la LLT ont été récompensés par des résultats excellents : alors qu'une simple recherche lexicale ayant pour objet *quisque* (*quique*, *cuique* etc.) créait, il y a quelque temps, de sérieux problèmes parce que la tokenisation était imprécise, les nombreux et récents essais que j'ai personnellement effectués sur la Library of Latin Texts démontrent qu'il n'y a pas de marge significative d'erreur.

- 43 L'exemple offert par la Library of Latin Texts est sans aucun doute intéressant, mais présente des interférences dangereuses entre l'éditeur proprement dit et l'éditeur numérique qui prend des responsabilités ecdotiques dont la portée est, à mon avis, supérieure à ses compétences effectives.

Inscriptions of Aphrodisia Project (InsAph)

- 44 Au milieu des années 90, de nombreux spécialistes d'épigraphie grecque et latine ont commencé à considérer le *web* comme un endroit idéal pour publier les résultats de leurs recherches, à l'intérieur de véritables éditions critiques numériques, réalisées tant pour le *web* que pour l'impression. Les raisons de ce choix, peu compréhensibles en apparence, vu les problèmes théoriques et techniques qu'il comporte, deviennent évidentes lorsqu'on prend en compte les rapides considérations qui suivent²⁷. Dès leur première parution, les publications académiques consacrées à l'épigraphie grecque et latine ont dû faire face à de problèmes d'ordre pratique considérables. Pour être publiée d'une manière efficace, chaque inscription doit être accompagnée non seulement d'une édition (diplomatique et critique) et d'un commentaire (épigraphique, linguistique, historique et prosopographique), mais aussi d'une ou plusieurs photographies ; et, dans le cas d'inscriptions très abîmées, la reproduction de l'apographe peut être d'une grande utilité. L'insertion des illustrations à l'intérieur des livres papier se révèle coûteuse ; et souvent, comme le démontrent les grands *corpus* consacrés à l'épigraphie gréco-latine (*in primis* le *Corpus Inscriptionum Latinarum* et les *Inscriptiones Graecae*), on a préféré publier les inscriptions sans aucun type d'image. Malgré ces renoncements, le coût total des *corpus* épigraphiques demeure très élevé, ce qui freine la diffusion de la connaissance. En outre – et c'est le problème le plus sérieux – en contradiction flagrante avec la nature de la discipline – ouverte à des mises à jours continues déterminées par la découverte de nouveaux matériels épigraphiques ou par la révision de documents déjà publiés – les grands *corpus* imprimés contenant des milliers d'inscriptions ne sont pas modifiables, et par conséquent obligent ceux qui souhaitent étudier à fond une épigraphe ou un groupe d'inscriptions à consulter une bibliographie toujours croissante, dispersée dans de nombreuses monographies et revues académiques, souvent très difficiles à trouver.
- 45 Comme le démontrent des projets récemment menés à bout, l'utilisation du *web* permet de résoudre ces problèmes, en offrant à l'épigraphiste – en qualité d'éditeur comme de chercheur – des outils d'une excellente qualité. Un très bon exemple est offert par la comparaison entre une publication académique publiée sur papier, et sa deuxième édition, publiée exclusivement sur le *web*²⁸.
- 46 Les nouveautés offertes par l'édition numérique par rapport à celle en format papier sont évidentes et significatives. Outre toutes les mises à jour nécessaires, des matériaux importants ont été ajoutés : des inscriptions précédemment non publiées, toute la documentation photographique demeurée, jusqu'alors, inédite pour des raisons de coût ; et, surtout, les journaux de J. Gandy Deering qui, pendant l'hiver de 1812-1813, avait reproduit le texte de nombreuses épigraphes d'*Aphrodisia* codicum instar, comme le disent les philologues, parce qu'au début du XIX^e siècle, les inscriptions étaient beaucoup plus lisibles qu'aujourd'hui, deux siècles plus tard.
- 47 Du point de vue technique et ecdotique, on enregistre des nouveautés très importantes, favorisées par l'époque relativement récente de la publication (le projet en ligne a été lancé en 2005). Analysons-les brièvement.
- 48 Le TLG contenait des textes codés en Beta Code – un système qui garantit une grande longévité mais limité aux 128 caractères ASCII ; la Library of Latin Texts a été au contraire réalisée avec un codage propriétaire très fonctionnel, mais utilisable uniquement à travers un *software* et un *hardware* spécifiques et qui, donc, exigera des efforts considérables pour être entretenu dans le temps. Le projet des inscriptions d'*Aphrodisia* consiste, en revanche, en une véritable édition numérique de textes épigraphiques, réalisée *ex-novo*, au moyen de systèmes avancés d'encodage textuel. Il s'agit, en particulier, d'Epidoc, un langage de balisage textuel compatible avec le système TEI – Text Encoding Initiative²⁹. TEI est un système de codage des textes, basé sur le métalangage XML³⁰, qui permet de marquer, à travers l'insertion

de convenables « balises », des portions de texte ou des vocables, en leur attribuant une description :

Dante, *Vita Nova*, Cod. Magl. VI 143, 4r, col. a:
 [6] *Et in quel puncto lo spirito anima
 le lo qual dimora nella camera nella
 quale tutti li spiriti sentia<ti>ui portando
 le loro perceptioni si comincio a mara
 uigliar molto et parlando spetialmente
 alli spiriti del uiso disse queste parole
 Apparui iam beatitudo vestra.*

Balises XML de: "tutti li spiriti sentiⁱ"
tutti li spiriti
^{corr="sensitivi"}
senti^{type="cuneo"}/>
^{hand="b" type="int"}
place="supralinear">ti</sup></sup>
^{reg="v"}^u i^{corr}

49 Le système TEI permet de produire des éditions critiques numériques de niveau universitaire, susceptibles d'être enrichies par la présence d'apparats critiques, de versions différentes du texte (par exemple, dans le cas des épigraphes d'*Aphrodisia*, on peut consulter l'édition diplomatique et critique du texte de chaque épigraphe), et même de variantes graphiques. [<http://insaph.kcl.ac.uk/iaph2007/iAph010201.html>]. En résumé, toute particularité textuelle que l'éditeur critique considère digne de remarque peut être codée. Si, par exemple, dans une édition critique numérique toutes les variantes appartenant à un manuscrit déterminé ont été correctement marquées, il nous sera possible de les rappeler rapidement, en rendant plus facile leur analyse d'ensemble et en obtenant un texte numérique qui reproduise, dans la mesure du possible, le faciès textuel d'un témoin déterminé³¹. L'adoption du système TEI et de ses dérivés comporte, toutefois, quelques désavantages évidents qui en ont limité la diffusion. En effet, le procédé de balisage du texte doit être, en grande partie, effectué manuellement : cela contribue à augmenter la précision du travail mais entraîne aussi une considérable dépense de temps. C'est pourquoi il a été, jusqu'à présent, appliqué avec succès dans des projets importants, mais limités dans leurs objectifs et dans leur étendue.

50 Dans l'ensemble, les bases de données représentent une solution théoriquement très utile pour l'étude du texte : les différentes déclinaisons possibles du modèle présentées dans cet exposé sont toutes très intéressantes, mais elles exigent une connaissance approfondie des critères ecdotiques, souvent absente chez l'utilisateur.

Corpus numérique sémantique

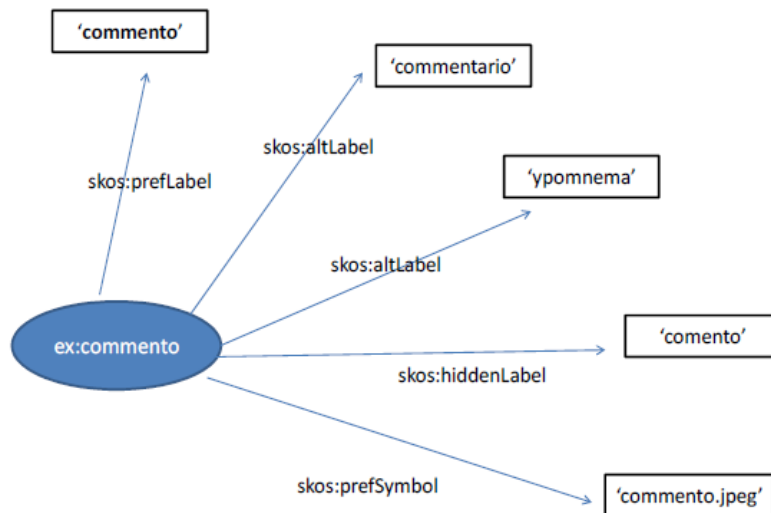
51 Introduit par Tim Berners-Lee en 2001, le *web* sémantique a été dès le début conçu non comme quelque chose de différent du *web* existant, mais plutôt comme une extension de celui-ci, pour permettre d'associer les données à une signification bien définie, dans le but d'augmenter la qualité de l'interaction entre les hommes et les ordinateurs³².

52 Afin de rendre plus compréhensible et applicable le concept de *web* sémantique, Tim Berners-Lee, dans un article publié en 2000, en avait dessiné la structure³³ en la représentant comme une architecture à plusieurs niveaux, hiérarchiquement ordonnés du bas (niveaux plus proches du langage de la machine) vers le haut (niveaux plus proches du raisonnement humain) ; chacun de ces niveaux était caractérisé par son langage spécifique ou par une stratégie spécifique d'organisation de l'information.

53 Parmi les nombreux standards élaborés dans le domaine du *web* sémantique, SKOS (*Simple Knowledge Organization System*) est l'un des plus intéressants³⁴. SKOS est un espace de travail conçu dans le but de développer des spécifications et des standards pour l'usage des systèmes d'organisation de la connaissance à l'intérieur du *web* sémantique. Plus précisément, la fonction remplie par SKOS est celle de connecter les systèmes traditionnels d'organisation de la connaissance des institutions culturelles (par exemple, les *thesaurus* et les systèmes de classification de bibliothèques, musées, archives), à travers les nouvelles structures conçues

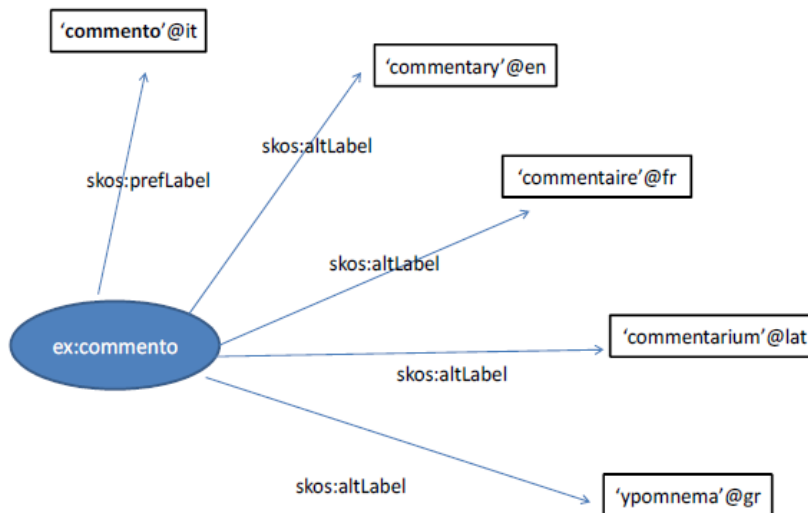
pour le *web* sémantique, telles que les ontologies, les cartes topiques et les projets *open directory*. Né en 2003 en tant que projet *open source* dans le cadre des activités du SWADE – *Semantic Web Advanced Development for Europe* – SKOS a été jugé digne d'attention par le W3C, qui s'est chargé de le développer. En 2005, on a élaboré la première esquisse du document contenant les spécifications de SKOS, mais c'est seulement en août 2009 qu'il a été inclus dans les recommandations du W3C³⁵.

- 54 Du point de vue technique et structurel SKOS, basé sur la syntaxe XML, consiste en une série de classes et de propriétés RDF³⁶, à travers lesquelles il est en mesure de créer un *framework* (structure logicielle) à l'intérieur duquel on peut établir des liens entre des concepts afférents à des domaines cognitifs différents. Actuellement sa fonction peut être double : d'une part il se prête à transférer dans le *web* sémantique les systèmes les plus traditionnels d'organisation de la connaissance ; de l'autre, il permet de construire, à partir de zéro, de simples schémas de concepts. Pour nos objectifs, il sera suffisant de décrire le fonctionnement du SKOS *Core Vocabulary*, qui constitue le niveau le plus simple.
- 55 Dans le système SKOS le concept est l'unité fondamentale d'un schéma des concepts (*concept schema*) et peut être défini comme n'importe quelle entité susceptible de description et formalisation. La première caractéristique intéressante de SKOS est que n'importe quel concept peut être exprimé par un seul terme préféré (*preferred term*), mais peut en même temps avoir beaucoup de termes alternatifs (*alternative tags*). On verra tout de suite de quelle manière.
- 56 Imaginons qu'on veuille créer à l'aide de SKOS une structure sémantique consacrée aux commentaires, considérés en tant que genre littéraire autonome qui depuis l'Antiquité jusqu'à aujourd'hui a rempli la tâche d'éclairer des textes (pas seulement littéraires, mais aussi de médecine, de droit, et scientifiques au sens large).
- 57 Une fois choisi « *commento* » comme terme préféré, nous allons y associer d'autres termes, dont la signification est similaire ou équivalente, et qui peuvent se trouver reliés à des ressources pertinentes : « *commentario* » (un synonyme, dans la langue italienne), « *ypomnema* » (le terme correspondant dans la langue grecque, translittéré), « *comento* » (le même terme, exprimé par un vocable de l'ancien italien), et encore, par exemple, « *commento.jpg* », qui garde le même terme mais associe au concept une image en format .jpg.
- 58 Le schéma peut être expliqué de la manière suivante. Il existe un *skos:concept* défini par le préfixe *ex:* et exprimé à travers des étiquettes différentes : *skos:prefLabel* signalera que « *commento* » est le terme préféré pour indiquer l'homonyme *skos:concept* ; *skos:altLabel* indiquera, au contraire, d'autres étiquettes, à savoir les termes, alternatifs, au moyen desquels le *skos:concept* peut être identifié ; *skos:hiddenLabel* indiquera, en revanche, une étiquette destinée à rester cachée, utile à rassembler de simples variantes graphiques du terme assumé comme principal (dans notre cas, le terme en ancien italien, c'est-à-dire « *comento* ») ; enfin, *skos:prefSymbol* permet d'associer au concept une image (dans ce cas le fichier *commento.jpg*).

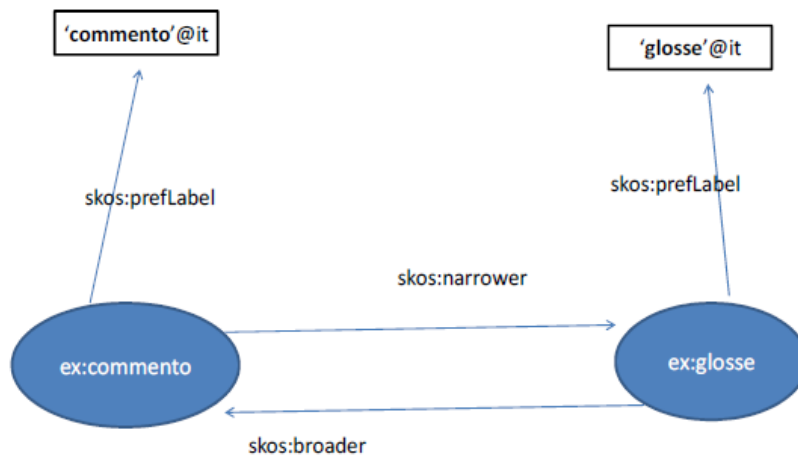


Voici la sérialisation RDF du schéma³⁷ : <rdf:RDF xmlns:rdf=http://www.w3.org/1999/02/22-rdf-syntax-ns#
 xmlns:skos="http://www.w3.org/2004/02/skos/core#">
 <skos:Concept rdf:about="http://www.esempio.it/concetti#commento">
 <skos:prefLabel>commento</skos:prefLabel>
 <skos:altLabel>commentario</skos:altLabel>
 <skos:altLabel>ypomnema</skos:altLabel>
 <skos:hiddenLabel>comento</skos:hiddenLabel>
 <skos:prefSymbol>http://www.symbols.com/commento.jpg</skos:prefSymbol>
 </skos:Concept>
 </rdf:RDF>

- 59 Dans l'optique de la *web* sémantique, SKOS présente une autre caractéristique importante : il peut représenter le même concept dans des langues différentes. L'utilité de cette approche est évidente et transcende le discours relatif au *web* culturel et éditorial : parce que la réalisation d'un véritable réseau sémantique mondial doit nécessairement passer par la création d'un *thesaurus* multilingue partagé, propre à exprimer au moins les concepts fondamentaux à travers une correspondance sans équivoques entre chaque concept et la pluralité des langues.
- 60 En revenant à notre exemple, SKOS nous permet d'agréger dans un unique système sémantique des objets numériques qui soient toujours des commentaires, mais exprimés par des métadonnées en langues différentes de l'italien : *commentary* en anglais; commentaire en français ; *commentarium* en latin et *ypomnema* en grec translittéré, comme on peut voir dans le schéma ci-dessous :



- 61 SKOS n'est pas seulement en mesure d'exprimer des concepts, il peut aussi définir des relations sémantiques entre des concepts différents, à travers trois modalités distinctes : *skos:broader* indique qu'un concept a une signification plus large qu'un autre ; *skos:narrower* décrit la relation contraire, tandis que *skos:related* est plus générique, en se limitant à indiquer qu'entre les deux termes il y a une relation. Si, dans le sillage des exemples précédents, l'on voulait exprimer une relation entre le commentaire et les gloses, on pourrait dire que le commentaire, qu'on doit concevoir comme l'exégèse d'un texte considéré dans tous ses aspects, a une signification plus générale par rapport à la glose (*skos:broader*), qui se limite à l'interprétation des mots obscurs à travers un synonyme. On pourrait de même dire, au contraire, que la glose a une signification plus étroite par rapport au commentaire (*skos:narrower*). On peut décrire cet ensemble de relations à l'aide d'un exemple, affiché dans le schéma suivant :



- 62 Voici sa sérialisation en RDF :

```
<rdf:RDF
  xmlns:rdf=http://www.w3.org/1999/02/22-rdf-syntax-ns#
  xmlns:skos="http://www.w3.org/2004/02/skos/core#">
  <skos:Concept rdf:about="http://www.esempio.it/concetti#glosse">
    <skos:prefLabel>glosse</skos:prefLabel>
    <skos:broader
      rdf:resource="http://www.esempio.it/concetti#commento"/>
```

```

</skos:Concept>
<skos:Concept
rdf:about="http://www.esempio.it/concetti#commento">
<skos:prefLabel>commento</skos:prefLabel>
<skos:narrower
rdf:resource="http://www.esempio.it/concetti#glosse"/>
</skos:Concept>
</rdf:RDF>

```

- 63 Après l'inclusion de SKOS parmi les recommandations officielles du W3C en août 2009, de nombreux projets qui l'utilisent ont été lancés³⁸. L'un d'eux se révèle particulièrement intéressant ; il s'agit de *An integrated view to medieval illuminated manuscripts* d'Antoine Isaac, qui se propose de constituer, à travers SKOS, un lien sémantique entre deux bases de données en ligne, *Mandragore* et *Medieval Illuminated Manuscripts*³⁹. Les deux bases de données en ligne, même si elles ont été conçues avec un objectif similaire – le désir commun était de créer une collection numérique de manuscrits enluminés conservés dans ces deux bibliothèques nationales – ont été réalisées en employant des schémas de métadonnées et des standards de classification différents pour la description du contenu des images : *Mandragore* se sert d'un vocabulaire spécifique, tandis que *Medieval Illuminated Manuscripts* utilise le standard *Iconclass*. L'objectif d'*An integrated view to medieval illuminated manuscripts* est donc d'intégrer les deux systèmes sans les modifier, en rendant possible à l'utilisateur une recherche sur les deux bases de données quel que soit le vocabulaire utilisé.
- 64 Du point de vue de la recherche dans des textes multilingues, le *corpus* sémantique offre des possibilités supérieures à n'importe quel autre outil numérique parce qu'il permet de connecter entre eux des vocables et des concepts identifiés à travers leur signification, quelle que soit la langue dans laquelle ils sont écrits.

Aspects théoriques et techniques : hypothèses de projet

- 65 Dans la conception d'un *corpus* comparatiste des traductions des classiques, il faut considérer tous les aspects théoriques et techniques. Tout d'abord, cependant, il est nécessaire d'identifier précisément les objectifs du nouvel outil numérique et de son contenu. Lors d'une recherche en sciences humaines, il nous arrive presque naturellement de redéfinir les objectifs en relation à ce qui ressort d'une analyse plus approfondie de l'objet d'étude ; mais dans le cas d'un *corpus* numérisé, cette démarche ne semble pas appropriée, car une redéfinition du projet en cours de route pourrait avoir des répercussions graves sur les plans économique et technique. Personnellement, je crois que les objectifs du *corpus* pourraient être les suivants.
- 66 1. Le *corpus* pourrait contenir intégralement les textes et les images des textes, c'est à dire les images des œuvres numérisées à une résolution suffisante pour être lues sur le *web*, mais pas pour être imprimées⁴⁰. Parmi les textes à inclure il faudra, bien sûr, considérer toutes les traductions examinées ; mais, après une étude historique minutieuse, on pourrait évaluer la possibilité d'inclure les textes originaux qui ont inspiré ces traductions⁴¹. Le *corpus* pourrait héberger un moteur de recherche capable d'extraire à la fois les métadonnées des livres (auteur, titre, imprimeur, date, peut-être la page de titre), les textes (préfaces, épîtres dédicatoires, etc.), la traduction proprement dite et toutes les notes, à travers une recherche de texte simple.
- 67 2. Le *corpus* pourrait être conçu comme un produit fini, et être donc disponible en ligne comme le résultat d'une recherche achevée ; selon une perspective différente, le *corpus* pourrait, au contraire, être envisagé comme évolutif, c'est-à-dire ouvert à des enrichissements ultérieurs ; mais il pourrait se présenter aussi comme une plateforme ouverte, sur laquelle les chercheurs accrédités pourraient produire et partager leurs recherches dans une section réservée.
- 68 3. Le *corpus* pourrait utiliser une structure sémantique complexe, capable de créer un réseau d'expressions et des mots reliés les uns aux autres.
- 69 Une fois définis les objectifs, nous irons maintenant les examiner de plus près pour comprendre toutes leurs implications théoriques et techniques.
- 70 1.1 Les images des éditions : d'un point de vue technique, numériser des images, y associer les métadonnées nécessaires (métadonnées structurelles, métadonnées de préservation, métadonnées descriptives etc.), et les présenter avec une interface facile à utiliser et agréable

à regarder, est une opération complexe et coûteuse mais, fondamentalement, testée et sans risque. En outre, ce procédé ne présente pas de problèmes de type ecdotique, comme nous l'avons vu en évaluant l'exemple de la BeicDL. Cette solution, qui est la plus utilisée pour la réalisation des *corpus* et des bibliothèques numériques, offre l'avantage de reproduire la forme d'un livre; l'inconvénient est que le texte reproduit en format image n'est pas sensible aux moteurs de recherche et les systèmes OCR qui peuvent être appliqués automatiquement ne produisent pas de bons résultats.

71 1.2. Les textes des éditions : reproduire dans un *corpus* numérisé une série de textes à partir de l'édition imprimée, pose des problèmes ecdotiques considérables, qui ont des implications très importantes dans l'utilisation du *corpus* – c'est ce qu'on a montré à travers les exemples du TLG, de la Library of Latin Texts et de l'Inscriptions of Aphrodisia Project.

72 1.3. Droit d'auteur : abordons le cas le plus simple, le livre que l'on souhaite inclure dans le *corpus* est déjà numérisé et disponible en réseau (Google Books, Internet Archive, Gallica, etc). Si l'on pense pouvoir le télécharger et le diffuser tel qu'il est, on fait preuve d'une très grande ingénuité. J'évoque l'exemple d'un type de licence connue et répandue comme Creative Commons : il en existe 12 versions de base, chacune avec des variations complexes. Obtenir la permission de publier les images numérisées d'un livre appartenant à une bibliothèque est souvent très difficile, car il faut d'abord résoudre un certain nombre de questions de nature juridique qui nécessitent des compétences spécifiques. C'est pourquoi, à mon avis, l'équipe devrait comprendre, dès le départ, un expert en droit, qui soit en mesure d'évaluer chaque cas et de faire des propositions plausibles aux bibliothèques, si elles sont impliquées dans le projet.

73 1.4. Gestion et conservation : les questions techniques relatives à la bonne conservation des objets numériques et de l'infrastructure dans son ensemble, et les coûts connexes sont une partie essentielle du projet qui sera prise en compte lors du planning des activités. Je rappelle ici que, de façon générale, les projets numériques en sciences humaines, même ceux bien construits, ne semblent pas prendre correctement en considération les coûts d'entretien et de gestion. Alors qu'un livre, une fois publié, ne nécessite plus aucune action (sauf si l'on veut en faire une deuxième édition), un projet numérique a besoin de soins constants. Si l'on travaille dans le domaine des standards et avec des logiciels *open source*, les coûts, limités, concerneront principalement la gestion des applications et leurs mises à jour, mais si l'on a décidé de choisir un produit commercial, il faut savoir que l'infrastructure technique a un cycle de vie de 5 à 7 ans, avec des licences annuelles, suivie d'un nouveau produit partiellement compatible avec le précédent, dont les coûts ne sont pas prévisibles lors de la préparation du budget.

74 2.1. Une fois la recherche terminée, le *corpus* de textes est mis en réseau, complété dans toutes ses parties. D'un point de vue théorique et technique, c'est la solution la plus simple, qui reste toutefois problématique à d'autres égards : d'un côté, les temps de recherches en sciences humaines étant longs, un retard en termes d'années pour la création d'un produit numérique peut avoir des répercussions très graves sur le plan des activités ; de l'autre côté, la réalisation d'un produit fermé et non extensible limite gravement l'utilité de l'outil.

75 2.2. Le *corpus* est un produit évolutif, qui peut être enrichi de nouveaux textes. Du point de vue technique, cette solution est complexe, car elle nous oblige à prévoir le cycle de vie des applications et la typologie des documents qu'il faudra insérer. Du point de vue théorique, il est conseillé d'opter pour des standards ouverts (XML pour le codage ; le protocole OAI-PMH pour le partage des métadonnées ; TIFF pour les images) : ces choix permettent une conservation plus facile des matériaux et permettent de passer aisément d'un système *hardware* à l'autre. Il faut également prévoir un groupe de travail qui s'occupe des aspects techniques et juridiques des produits numériques, même après leur mise en réseau.

76 2.3. Le *corpus* est une plateforme ouverte. Il ne s'agit pas seulement, dans ce cas, de constituer une base de données de textes qui dans le temps peut être augmentée, mais aussi de réaliser un lieu virtuel où les spécialistes peuvent travailler, partager des documents et communiquer. Une fois ouverte au public, la plateforme présenterait seulement des livres et des études numérisés (articles ou monographies) déjà prêts. Des projets comme Interedition (<http://www.interedition.eu/>) nous invitent à aller dans cette direction, bien que ce choix ne semble

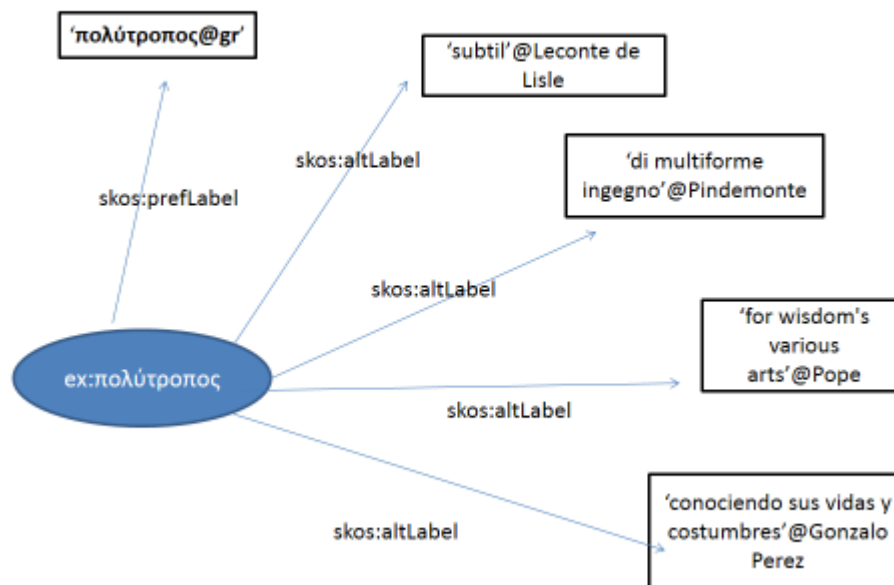
pas rencontrer la faveur des chercheurs italiens. C'est la solution la plus difficile en termes de conception et de coût, mais aussi celle pourvue du plus haut degré d'innovation et la mieux indiquée pour être présentée à l'intérieur de projets européens ou locaux.

3. Réseau sémantique : imaginons qu'un chercheur veuille savoir comment on traduisait en anglais au XVIII^e siècle le terme « πολύτροπος », ou comment on rendait une expression formulaire en espagnol, en latin ou en français, dans les traductions d'une période donnée. C'est un genre d'information précieux, qui permet d'extraire des données sans avoir recours à la lecture complète des textes. Les résultats ne peuvent être obtenus par l'intermédiaire d'un moteur de recherche textuel: en revanche, quelques standards du *web* sémantique – en particulier, SKOS – pourraient se révéler très efficaces⁴².

Comme nous l'avons vu, SKOS (*Simple Knowledge Organization System*) est un domaine de travail du W3C (*World Wide Web Consortium*) pour l'élaboration des normes qui rendent possible la connexion entre les différents systèmes d'organisation de la connaissance utilisés dans les bibliothèques, les musées, les archives (il s'agit des *thesaurus* et des systèmes de classification), et les nouvelles structures conçues pour le *web*.

Comme on l'a déjà dit, dans le système SKOS le concept (*concept*) est l'unité de base de tout système de concepts (concept schéma) et peut être définie comme n'importe quelle entité susceptible de description et formalisation. Une caractéristique intéressante de SKOS est que tout concept peut être exprimé par un terme unique préféré (*preferred term*), mais de nombreux termes alternatifs (*alternative tags*) sont disponibles.

Voici un exemple, en utilisant « πολύτροπος » :



L'exemple permet de voir comment le même terme (πολύτροπος) peut être associé aux traductions correspondantes dans des langues différentes. Voici la sérialisation en XML RDF-SKOS de l'exemple :

```
<rdf:RDF
  xmlns:rdf=http://www.w3.org/1999/02/22-rdf-syntax-ns#
  xmlns:skos="http://www.w3.org/2004/02/skos/core#">
  <skos:Concept rdf:about="http://www.Corpuscomparatiste.fr/concepts#πολύτροπος">
    <skos:prefLabel xml:lang="grc">πολύτροπος</skos:prefLabel>
    <skos:altLabel xml:lang="fre"
      xml:author="Leconte_de_Lisle">subtil</skos:altLabel>
    <skos:altLabel xml:lang="ita" xml:author="Pindemonte">di multiforme
      ingegno</skos:altLabel>
    <skos:altLabel xml:lang="eng" xml:author="Pope">for wisdom's
      various arts</skos:altLabel>
```



```
<skos:altLabel xml:lang="spa"
xml:author="Gonzalo_Perez">conociendo sus vidas y
costumbres</skos:altLabel>
</skos:Concept>
</rdf:RDF>
```

- 82 Une solution de ce type est la plus intéressante au niveau du projet et c'est celle qui offre les résultats les plus utiles pour un chercheur. Basée sur les standards, elle assure la durabilité et l'indépendance du *hardware* et du *software*. Seule la difficulté de sa conception pose des problèmes : elle doit être spécifique pour chaque projet, mais suffisamment précise pour pouvoir être utilisée par d'autres groupes de recherche.

Bibliographie

Bibliographie des études critiques

BURNARD, Lou, SPERBERG MCQUEEN, Michael, *Il manuale TEI Lite. Introduzione alla codifica elettronica dei testi letterari*, edizione italiana a cura di F. Ciotti, Milano, Edizioni Sylvestre Bonnard, 2005.

CERQUIGLINI, Bernard, *Éloge de la variante. Histoire critique de la philologie*, Paris, Seuil, 1989.

CHIESA, Paolo, « Sul controllo filologico delle edizioni critiche digitali », *Filologia mediolatina*, 17, 2010, p. 325-346.

CONSONNI, Chiara, DEANA, Danilo, *L'infrastruttura e il sistema di BeicDL*, in *La Biblioteca europea di Milano. Vicende e traguardi di un progetto*, a cura di A. Padoa Schioppa, Milano, Skira, 2014, p. 153-163.

Library of latin textes – Series A, User's Guide, s. l., Brepols, 2009.

ORLANDI, Tito, *Informatica umanistica – Riflessioni storiche e metodologiche con due esempi*, in *Studi di codifica e trattamento automatico dei testi*, a cura di G. Gigliozzi, Bulzoni, Roma, 1987, p. 1-38.

PERILLI, Lorenzo, *Filologia computazionale*, Roma, Accademia Nazionale dei Lincei, 1995.

TISSONI, Francesco, « Testi latini on line ad accesso libero : una prima valutazione (1 settembre 2004) », *Acme*, 57, 2004, p. 43-79.

TISSONI, Francesco, *Lineamenti di editoria multimediale*, Milano, Unicopli, 2009.

TOMASI, Francesca, *Metodologie informatiche e discipline umanistiche*, Roma, Carocci, 2008.

Ressources Web

Aphrodisias in Late Antiquity : <http://insaph.kcl.ac.uk/index.html>

CTLO – Centre Traditio Litterarum Occidentalium (ex. Cetedoc – Centre de Traitement Electronique des documents) : <http://www.corpuschristianorum.org/centres/turnhout.html>

Epidoc : <http://epidoc.sourceforge.net/>

Ibycus System : <http://www.tlg.uci.edu/about/ibycus.php>

Index Thomisticus : <http://www.corpusthomicum.org/it/index.age>

Interedition : <http://www.interedition.eu/>

LL – Library of Latin Texts : <http://apps.brepols.net/BrepolsPortal/default.aspx>

PHI – Packard Humanities Institute : <http://www.packhum.org/phi/>

TEI – Text Encoding Initiative : <http://www.tei-c.org/index.xml>

TLG – Thesaurus Linguae Græcæ : <http://www.tlg.uci.edu/>

XML – Extensible Markup Language : <http://www.w3.org/XML/>

Notes

1 Le terme « concept » est ici employé dans son sens commun ; successivement, dans la section consacrée à SKOS, il sera employé dans un sens technique. Par « concept » on indique ici, par exemple, une épithète ou une formule homérique qui, par la nature même du langage, est rendue dans les langues modernes à travers plusieurs mots, ce qui rend problématique une recherche purement lexicale.

2 Sur le concept de bibliothèque numérique et sur ses premières déclinaisons voir A. Salarelli, A. M. Tammaro, *La biblioteca digitale*, Milano, Editrice Bibliografica, 2006.

3 Voir, à ce propos, C. Consonni, D. Deana, *L'infrastruttura e il sistema di BeicDL*, in *La Biblioteca europea di Milano. Vicende e traguardi di un progetto*, a cura di A. Padoa Schioppa, Milano, Skira, 2014, p. 153-163. La raison du choix de cet exemple est simple : comme je collabore à la réalisation de la bibliothèque numérique depuis 2009, j'ai pu vérifier personnellement ce que je présente.

4 Le format des images dont on a décidé d'assurer la conservation à long terme, par exemple, est unique (TIFF, Tagged Image File Format) et conforme à une spécification standard. Il en va de même pour les métadonnées qui accompagnent les images, insérées à l'intérieur d'un fichier XML conforme au schéma METS (Metadata Encoding and Transmission Standard), dont les différentes sections contiennent, à leur tour, des métadonnées conformes à d'autres schémas : MARC XML (le schéma XML basé sur le standard MARC, Machine Readable Cataloguing), MIX (metadata for Images in XML Standard) et PREMIS (Preservation Metadata: Implementation Strategies). Il s'agit de langages, schémas et standards entretenus par le Word Wide Web Consortium et par la Library of Congress, qui sont soumis à une évolution constante et qui sont adoptés par les plus importantes bibliothèques numériques du monde.

5 C'est le cas en particulier de la BEIC mais l'on peut retrouver ces mêmes caractéristiques dans les autres bibliothèques numériques réalisées dans le respect des standards.

6 Voir M. PÉREZ CERVERA, *Discovery tools: uno sguardo ai criteri di selezione, impatto e percezioni* : <http://www.ub.edu/blokdebid/es/node/495#sthash.RdSnGlv9.dpuf>

7 Il faut par ailleurs remarquer qu'on ne doit pas se limiter à préserver dans le temps les documents numériques (tels que les livres ou d'autre typologies de documents), il faut aussi en préserver les conditions d'utilisation.

8 La BeicDL utilise Secure Image et Imago Libris : voir C. Consonni, D. Deana, *op. cit.*, p. 154.

9 Voir : <http://www.exlibrisgroup.com/category/DigiToolOverview>. Il existe des alternatives *open source* valables, comme par exemple CONTENTdm de la Digital Library: <https://www.oclc.org/contentdm.en.html>. D'un point de vue opératif et de gestion, l'*open source* exige des interventions de personnalisation et d'entretien qui, à moyen et long terme, ne rendent pas ces applications substantiellement différentes, quant aux coûts et à la qualité, des systèmes propriétaires. La seule condition est que ces derniers, même propriétaires, soient basés sur les standards.

10 Voir : <http://www.niso.org/publications/press/> : Understanding Metadata.

11 Voir : <https://www.loc.gov/standards/mets/METSOOverview.v2.html>

12 Voir : <http://www.koha.org/>

13 *Optical character recognition* : il s'agit d'un système de reconnaissance automatique des caractères qui permet de convertir des images contenant du texte en texte numérisé. Cette technologie ne permet actuellement une lecture efficace ni de caractères non latins ni d'écritures latines particulières.

14 BeicDL a consacré beaucoup d'attention à la mise à point d'un catalogue convivial pour l'utilisateur en se dotant d'un outil de dernière génération, *Primo* de *Ex Libris* : <http://www.exlibrisgroup.com/category/PrimoOverview>. Il existe d'autres outils de découverte *open source* également efficaces, tels que VuFind : <http://vufind-org.github.io/vufind/>.

15 Voir <https://www.handle.net/>.

16 Au sujet de OAI-PMH, voir C. Martignoni, S. Merlini, A. Stella, F. Tissoni, F. Venturi, « Come informatizzare il Novecento letterario. Un caso esemplare: la poesia di Andrea Zanzotto », *La modernità letteraria*, 3, 2010 et F. Tissoni, *Lineamenti di editoria multimediale*, Milano, Unicopli, 2009, p. 148-155.

17 Voir <http://unicode.org/>

18 Le TLG a été réalisé avec un codage textuel spécifique, dénommé *Beta Code*. Sur le codage du grec, de 1972 à aujourd'hui, voir F. Tissoni, *Lineamenti di editoria multimediale*, éd cit., p. 64-69.

19 Professeur Theodore F. Brunner : <https://www.tlg.uci.edu/about/ted.brunner.php>. Après plus de quarante ans, à travers un développement constant, le projet du TLG a été pleinement réalisé ; et après plusieurs migrations sur de différents supports (en dernier, le CD-ROM version E), le TLG est disponible en ligne à l'adresse <<http://www.tlg.uci.edu/>> et consultable par abonnement.

20 Une remarquable exception est représentée par l'étude de L. Perilli, *Filologia computazionale*, Roma, Accademia Nazionale dei Lincei, 1995.

21 Variantes graphiques : par exemple, les mêmes termes qu'un éditeur peut avoir imprimés avec le *iota* adscrit ou souscrit.

22 C'est le cas, fréquent, des fragments qui peuvent apparaître, en tant que citations, à l'intérieur des éditions des auteurs qui les rapportent et, en même temps, dans des recueils spécifiques. Dans plusieurs cas, le fragment est rapporté deux fois et contient des variantes, souvent introduites par l'éditeur moderne.

23 Un exemple : le médecin Galien (*De antidotis*: vol. 14, 38, Kühn) cite un long fragment poétique, qu'il attribue au poète et médecin Andromaque, vécu au I^{er} siècle ap. J.-C. ; dans le TLG, le fragment

d'Andromaque est donc compris à l'intérieur de l'œuvre de Galien. Le même fragment, cependant, apparaît dans le TLG aussi sous le nom de son auteur, Andromaque, mais le texte est tiré d'une édition différente : *Die Griechische Dichterfragmente der Römischen Kaiserzeit* d'Ernst Heitsch (Göttingen 1963). Il est intéressant de comparer le vers 102 du bref petit poème dans les deux versions. Dans le texte d'Andromaque cité par Galien, présent dans le TLG, on lit « εὔδιπρος », tandis que dans l'édition de Heitsch du même fragment, on lit « εὔδιπος ». Les deux entrées sont très rares : chacune compte deux occurrences dans toute la littérature grecque ; toutefois, l'un de deux vocables qu'on lit dans le passage d'Andromaque ne peut pas, évidemment, exister. Il s'agit d'une variante textuelle qui, par la faute de son enregistrement malheureux à l'intérieur du TLG, peut devenir une partie de la langue grecque. Il est évident que si les textes indexés ne sont pas fiables, le résultat d'une analyse fondée sur eux, risque, à son tour, de ne pas l'être.

24 On reconnaît des critères ecodotiques semblables à ceux du TLG dans le PHI (un cd-rom contenant des éditions critiques de tous les textes latins jusqu'au II^{ème} siècle ap. J.-C.), réalisé – ce n'est pas par hasard – par le Packard Humanities Institute <<http://www.packhum.org/phi/>>, fondé par David Packard en 1987. Même dans ce cas, la base de données contenait seulement des éditions numériques spéculaires d'un exemplaire imprimé. Si, dans le cas des textes grecs, ce procédé peut sembler discutable mais pas totalement sans fondement, dans le cas des textes latins il présente des inconvénients encore plus graves, parce que les mêmes mots, dans des éditions différents de plusieurs auteurs, peuvent présenter les lettres *i* ou *j* (par exemple, *ius* ou bien *jus*); contenir *u* ou *v* (*uale* ou bien *vale*) et, moins fréquemment, *c* ou *k* (*Calendæ* ou bien *Kalendæ*) pour nous en tenir à ces seuls exemples. La quantité de ces alternatives graphiques, tout à fait insignifiantes du point de vue ecodotique (elles pourraient même être liquidées, dans certains cas, en tant qu'habitudes éditoriales), est tellement élevée qu'elle oblige l'utilisateur à répéter plusieurs fois la même recherche ; et même dans ce cas la fiabilité du résultat reste incertaine. Après plusieurs versions, qui ne présentaient pas de substantiels changements, le projet a été considéré comme conclu et le cd n'a plus été mis à jour après la version 5.3 livrée en 1991.

25 Ces intentions ont été éclairées pour la première fois dans la « Préface » à la deuxième édition du Cetedoc, parue en 1986, « Paul Tombeur moderante », et ont été par la suite répétées jusqu'à aujourd'hui (Library of Latin Texts Series A, User's Guide, 2009).

26 Voir F. Tomasi, *Metodologie informatica e discipline umanistiche*, Roma, Carocci, 2008, p. 191-211.

27 J'ai déjà exposé ces réflexions dans F. Tissoni, *Lineamenti di editoria multimediale*, éd. cit., p. 129-132.

28 C. Roueché, *Aphrodisias in Late Antiquity : The Late Roman and Byzantine Inscriptions*, London, Society for the Promotion of Roman Studies, 1989 et <http://insaph.kcl.ac.uk/iaph2007/>

29 F. Tissoni, *Lineamenti di editoria multimediale*, éd. cit., p. 110-123.

30 *Ibid.*, p. 90-102.

31 En ce qui concerne Epidoc (voir *ibid.*, p. 132-145) on se limitera à dire qu'il est un sous-ensemble de TEI développé par Thomas Elliott pour l'épigraphie grecque et latine (adapté récemment à la papyrologie), qui permet de reproduire parfaitement tous les signes diacritiques introduits par la convention de Leyde en rendant possible une correspondance parfaite entre l'édition papier et celle numérique.

32 « *The Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, enabling computers and people to work in cooperation* » : T. Berners-Lee, J. Handler, O. Lassila, « The Semantic Web », *Scientific American*, 17, 2001.

33 T. Berners-Lee, « Semantic Web-XML 2000 », disponible à l'adresse : <http://www.w3.org/2000/Talks/1206-xml2k-tbl/slide10-0.html>

34 F. Tissoni, *L'editoria multimediale nel nuovo web*, Milano, Unicopli, 2010, p. 89-96. Voir aussi le plus récent T. Baker, S. Bechhofer, A. Isaac, A. Miles, G. Schreiber, E. Summers, « Key choices in the design of Simple Knowledge Organization System (SKOS) », *Web Semantics: Science, Services and Agents on the World Wide Web*, 20, 2013, p. 35-49.

35 <http://www.w3.org/TR/2009/REC-skos-reference-20090818/>

36 RDF Resource Description Framework : voir <http://www.w3.org/RDF/> et F. Tissoni, *L'editoria multimediale nel nuovo web*, éd. cit., p. 60-67.

37 Je vous invite à observer la déclaration des espaces de noms RDF e SKOS à travers le préfixe xmlns : il faut tenir compte du fait que les url <<http://esempio.it>> et <<http://www.symbols.com>> ont été insérées à titre d'exemple et doivent donc être considérées comme fictives.

38 Voir : <http://www.w3.org/TR/skos-ucr/>.

39 <http://mandragore.bnf.fr/> et <http://www.kb.nl/manuscripts/>. Il s'agit des deux principaux projets européens de numérisation de manuscrits enluminés : le premier est français, sous la direction de la Bibliothèque Nationale de France ; le second est dirigé par la Koninklijke Bibliotheek, la Bibliothèque

Nationale Néerlandaise. Une description du projet et de son état d'avancement est disponible ici : http://www.cs.vu.nl/STITCH/BNF_KB_demo.html

40 Il s'agit de faciliter la concession des droits de publication de la part des bibliothèques concernées.

41 Pour clarifier le concept, on va considérer un cas célèbre. On sait que Leopardi a traduit la *Batrachomyomachia* d'Homère et on connaît la richesse de la bibliothèque de son père Monaldo : un examen de la traduction permettra d'identifier avec une relative facilité l'édition grecque utilisée par Leopardi, qui pourrait être introduite dans le *corpus*.

42 Voir M. van Gendt, A. Isaac, L. van der Meij, St. Schlobach, *Semantic Web Techniques for Multiple Views on Heterogeneous Collections : A Case Study*, in J. G. C. Thanos, M. Felisa Verdejo, R. C. Carrasco (Eds.), *Research and Advanced Technology for Digital Libraries*, 10th European Conference, ECDL 2006, Alicante, Spain, September 17-22, Berlin-Heidelberg, Springer, 2006, p. 426-437.

Pour citer cet article

Référence électronique

Francesco Tissoni, « Pour un corpus numérique comparatiste des traductions d'Homère », *Corpus Eve* [En ligne], Homère en Europe à la Renaissance. Traductions et réécritures, mis en ligne le 31 décembre 2015, consulté le 02 janvier 2016. URL : <http://eve.revues.org/1273>

À propos de l'auteur

Francesco Tissoni

Francesco Tissoni (1967), spécialiste de Littérature Grecque, enseigne "Editoria Multimediale" à l'Université de Milan. Il dirige des projets novateurs dans le domaine des humanités numériques, notamment auprès de la Biblioteca Europea d'Informazione e Cultura (www.beic.it).

Droits d'auteur

© Tous droits réservés

Résumé

Après avoir rappelé les principes qui ont inspiré les collections de livres numériques telle la Beic Digital Library, les bases de données de textes tel le Thesaurus Linguae Græcæ et la Library of Latin texts, nous pouvons réfléchir sur la notion de "corpus numérique sémantique" en analysant ce qui existe déjà et ce que les chercheurs pourront élaborer à partir des exigences spécifiques d'un corpus plurilingue de traductions d'Homère.

Entrées d'index

Mots-clés : corpus numérique sémantique, bibliothèque numérique, bases de données des textes

Domaines linguistiques : domaine italien

Index chronologique : XXe siècle, XXIe siècle, XVIe siècle