

Pour une annotation sémantique des textes: le projet symogih.org et la Text encoding initiative

Francesco Beretta

▶ **To cite this version:**

Francesco Beretta. Pour une annotation sémantique des textes: le projet symogih.org et la Text encoding initiative. Bruniana e Campanelliana, Ricerche filosofiche e materiali storico - testuali, 2016, XXII (2), 10.19272/201604102005 . halshs-01505635

HAL Id: halshs-01505635
https://shs.hal.science/halshs-01505635

Submitted on 8 Jun 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - ShareAlike 4.0 International License

Pour une annotation sémantique des textes: le projet *symogih.org* et la *Text encoding initiative*

par

Francesco Beretta

(francesco.beretta@ish-lyon.cnrs.fr)

Texte publié dans *Bruniana & Campanelliana*, XXII, 2, 2016, pp.453-465 :

prière de se référer au texte publié pour la citation

(doi 10.19272/201604102005)

Les textes se trouvent au cœur de la recherche en histoire: qu'il s'agisse de sources, d'éditions et de monographies, ou de travaux de synthèse, l'historien les lit, identifie les concepts et les connaissances qu'ils contiennent, et qui sont significatifs au point de vue de sa problématique, puis prend des notes et produit ainsi de nouveaux textes. Il s'agit d'un processus d'enrichissement sémantique opéré par le spécialiste au cours de sa lecture. Toutefois, ce travail d'annotation, même saisi sur un support numérique, est généralement déconnecté du texte de départ et n'est pas structuré au point de vue informatique: il n'est donc pas exploitable par les logiciels, alors qu'on pourrait le mettre en valeur grâce à un processus permettant d'interroger les textes et d'opérer sur la couche sémantique ajoutée par le chercheur. L'historien serait ainsi équipé avec de nouveaux outils et méthodes afin de renforcer son emprise sur la transcription numérique des documents et, ensuite, ces textes enrichis sémantiquement pourraient être facilement mis à la disposition d'autres chercheurs et du public sur internet.

Annotation sémantique et éditions en ligne

Le développement des méthodes et technologies numériques appliqués à la recherche en sciences humaines, et en particulier à l'histoire, permet désormais de mettre en valeur le travail d'enrichissement sémantique des textes.[1] Ces méthodes ont été appliquées dans plusieurs projets

1 Pour une présentation des technologies sémantiques, parmi de nombreuses autres, voir *Handbook of semantic web technologies. Foundation and technologies*, edited by J. Domingue, D. Fensel, J. A. Hendler, Berlin/Heidelberg, Springer, 2011. Pour une application au domaine de l'histoire: A. Meroño-Peñuel et al., *Semantic Technologies for Historical Research: A Survey*, « Semantic Web – Interoperability, Usability, Applicability », VI, 2015, pp. 539-564 (http://www.semantic-web-journal.net/content/semantic-technologies-historical-research-survey) (tous les sites web cités ont été consultés en juin 2016).

d'édition en ligne de documents de l'époque moderne, relevant en particulier de l'histoire intellectuelle. À titre d'exemple, mentionnons l'édition des œuvres et de la correspondance de Galileo Galilei réalisée par le *Museo Galileo* de Florence, dans le cadre du projet *Galileo//thek@*,[2] qui permet d'explorer le lexique des documents en lien avec les textes édités,[3] tout comme l'*Archivio dei filosofi del Rinascimento*, l'un des projets réalisé dans le cadre de l'*Istituto del Lessico Intellettuale Europeo e Storia delle Idee,*[4] qui met à disposition du public l'annotation des concepts présents dans les œuvres – entre autres philosophes – de Tommaso Campanella.[5] Mentionnons également le projet d'édition annotée des correspondances savantes des Pays-Bas, *Circulation of Knowledge and Learned Practices in the 17th-century Dutch Republic*, qui propose une navigation ergonomique mettant en relation les textes et les objets qu'ils mentionnent.[6] Signalons enfin la revue en ligne *RIDE. A review journal for digital editions and resources*[7] qui est dédiée à l'évaluation et à la discussion des projets d'édition en ligne, et qui met l'accent sur la qualité de la méthode adoptée et l'ergonomie de l'accès aux ressources publiées.

En dépit de la multiplication de ces ressources en ligne, la méthode d'annotation sémantique évoquée ci-dessus reste souvent inaccessible aux chercheurs en histoire dans leur travail quotidien, et ce en dépit de l'existence de plates-formes qui mettent à disposition non seulement des ressources textuelles mais encore des outils, formations et supports méthodologiques, par exemple dans le cadre de la plate-forme des *Éditions en ligne de l'École des chartes,*[8] d u *Centre d'Études Supérieures de la Renaissance* à Tours[9] ou du projet *Biblissima*.[10] Cette situation dépend non seulement de la création récente de ces outils, et de la nécessité de compter au moins deux ou trois générations afin qu'une innovation méthodologique importante se répande plus largement dans une communauté disciplinaire, mais encore de l'absence de structures de recherche plus proches des historiens mettant à leur disposition ces compétences méthodologiques, voire un environnement numérique permettant de les accompagner dans leur travail quotidien d'annotation des textes.

Nous présenterons ici une expérience qui a permis de mettre en place une méthode et une infrastructure dédiée au travail d'annotation sémantique des textes en lien avec leur édition en ligne mais aussi avec le travail de recherche des historiens: le *Portail XML*[11] réalisé par le *Pôle histoire numérique* du *Laboratoire de Recherche Historique Rhône-Alpes* (LARHRA – UMR5190).[12] Cette plate-forme permet de produire et de publier des textes enrichis par une annotation sémantique

2 http://www.museogalileo.it/esplora/portalegalileo.html.
3 http://galileoteca.museogalileo.it/GTConsult/?lang=it .
4 http://www.iliesi.cnr.it/ .
5 http://www.iliesi.cnr.it/ATC/lessico.php .
6 http://ckcc.huygens.knaw.nl/epistolarium/ .
7 http://ride.i-d-e.de/ .
8 http://elec.enc.sorbonne.fr/ .
9 http://cesr.univ-tours.fr/ .
10 http://www.biblissima-condorcet.fr/ .
11 http://xml-portal.symogih.org .
12 http://larhra.ish-lyon.cnrs.fr/pole-histoire-numerique .

effectuée à partir de données issues de la base de données collaborative du projet *symogih.org* (Système modulaire de gestion de l'information historique).[13] La finalité principale de l'environnement numérique mis en place est de servir de support à la recherche en histoire, en facilitant l'accès aux pratiques informatiques disciplinaires comme la production collaborative des données, les analyses statistiques, spatiales ou de réseaux sociaux, l'encodage de textes numériques en lien avec les données structurées qui en sont extraites.

Dans ce dernier domaine, trois parmi les projets en cours ont donné lieu à une mise en ligne d'une portion des documents en cours d'encodage, en utilisant un premier habillage expérimental qui vise à étudier les technologies et modalités d'affichage permettant de faciliter la lecture et l'exploration des textes. Un premier projet, réalisé dans le cadre de la thèse de doctorat de Rosemonde Letricot, concerne une édition critique numérique des *Mémoires* de Léonard Michon, notable lyonnais du 18e siècle.[14] Un deuxième projet, *Society Religion Science. Digital intellectual history*, se propose de mettre à disposition du public les ressources textuelles concernant l'histoire intellectuelle produites au cours du travail d'analyse des textes et de production de données des chercheurs qui participent au projet.[15] Il s'agit de mettre à la disposition du public non des documents inédits mais des textes enrichis sémantiquement par la lecture de l'historien qui les exploite tout d'abord pour ses propres travaux scientifiques et, ensuite, rend accessible sur internet le fruit de son travail. Le troisième projet, en cours de réalisation par Christine Chadier sous la direction d'Yves Krumenacker, propose une *Édition numérique des Actes des églises réformées de Bourgogne au XVIIe siècle.*[16]

Dans la suite de cette contribution, nous exposerons les principes essentiels de la méthode d'annotation sémantique proposée dans le cadre de la plate-forme du *Pôle histoire numérique*. Cette méthode, qui s'inscrit dans le contexte d'un intérêt croissant pour l'encodage sémantique des textes,[17] a été présentée lors de la conférence annuelle de la *Text encoding initiative* (TEI)[18] tenue à Lyon en 2015.[19] Elle met en dialogue l'approche générique de production de données mise en œuvre au sein du projet *symogih.org* avec le standard de représentation des textes numériques développé

13 http://symogih.org/ .

14 Rosemonde Letricot, édition critique numérique des Mémoires de Léonard Michon, notable lyonnais, doctorat d'histoire moderne sous la direction de Bernard Hours, LARHRA UMR 5190 / Université Lyon 3, financé par le dispositif ARC de la Région Rhône-Alpes http://journal-michon.symogih.org .

15 http://srs.symogih.org/ .

16 http://synodes-protestants.symogih.org/ .

17 Voir, entre autres, Ø. EIDE, *Ontologies, Data Modeling, and TEI,* « Journal of the Text encoding initiative », VIII, 2014-2015, http://jtei.revues.org/1191 and A. JORDANOUS, A. STANLEY, C. TUPMAN, *Contemporary transformation of ancient documents for recording and retrieving maximum information: when one form of markup is not enough*, « Proceedings of Balisage: The Markup Conference 2012 », VIII, 2012, http://www.balisage.net/Proceedings/vol8/html/Jordanous01/BalisageVol8-Jordanous01.html .

18 Site web du consortium: http://www.tei-c.org .

19 F. BERETTA, *The symogih.org project and TEI: encoding structured historical data in XML texts,* « Text Encoding Initiative Conference and Members' Meeting 2015. Connect, Animate, Innovate », Lyon, 2015, http://tei2015.huma-num.fr/fr/ . Cf. https://halshs.archives-ouvertes.fr/halshs-01251915 .

par la TEI. Il s'agira ici de présenter une méthode générique d'annotation sémantique, susceptible d'être adaptée à tout type de texte en lien avec la recherche historique, et d'en illustrer les enjeux.

La gestion collaborative de données géo-historiques

Mettre en place une annotation numérique de textes en lien avec la recherche en histoire implique de pouvoir se référer à un système de gestion de l'information qui fournisse des outils permettant de modéliser le monde qu'il s'agit d'étudier, d'identifier les objets présents dans ce monde, de décrire les relations que ces objets entretiennent entre eux et de les situer dans l'espace et dans le temps. Un tel système de connaissances, structuré par un modèle, est appelé en informatique une ontologie. Nous décrirons ici le processus qui a conduit à la mise en place de l'ontologie du projet *symogih.org,* né en 2008 de la volonté de quelques historiens du LARHRA de mutualiser les données structurées produites au cours de leur recherche afin de permettre leur réutilisation par d'autres collègues. Cette démarche s'inscrit dans la logique de la 'curation des données' [20] entendue au sens d'un enrichissement et d'une amélioration constantes des données afin de garantir leur qualité, leur accessibilité et leur préservation. À titre d'exemple, les données produites au cours du projet SIPPAF,[21] financé pendant trois ans par l'Agence nationale française de la recherche et ayant mis en place un système d'information consacré aux patrons français (XIX^{e}-XX^{e} siècle), continuent à être enrichies et utilisées, notamment dans le cadre du projet SIPROJURIS qui a constitué un système d'information consacré aux professeurs de droit en France de 1804 à 1950.[22]

Un nombre croissant d'historiens et de projets internes et externes au LARHRA (actuellement plus d'une soixantaine d'utilisateurs et une quinzaine de projets) utilisent cet environnement numérique de recherche afin de produire et de mutualiser leurs données. La plate-forme comprend également un système de gestion des données spatiales, accessible depuis le site de partage de ressources GEO-LARHRA,[23] ainsi qu'un point d'accès SPARQL qui permet d'interroger directement une portion restreinte des données mise à disposition forme au format RDF.[24] Les données sont mises à disposition du public sous licence *Creative Commons Attribution-ShareAlike 4.0 International* comme l'ensemble des ressources du projet. De plus, les données récoltées au cours de ces projets peuvent être réutilisées par des doctorants et des historiens afin d'élargir les bases de leur recherche.

La mise en place d'une plate-forme de production et de curation collaborative de données géo-historiques imposait d'adopter dès le début une modélisation générique, susceptible de s'adapter à

20 Voir l'entrée *data curation* de l'édition anglaise de Wikipedia (https://en.wikipedia.org/wiki/Data_curation) et la bibliographie qui y est mentionnée.
21 http://www.patronsdefrance.fr/ .
22 http://siprojuris.symogih.org/ .
23 http://geo-larhra.ish-lyon.cnrs.fr/ .
24 http://symogih.org/?q=rdf-publication .

tout type d'information quelles que soient les thématiques de recherche ou la période étudiée, sans devoir restructurer la base pour chaque nouveau projet.[25] Plus récemment, la publication des données au format RDF et la poursuite de l'interopérabilité avec d'autres producteurs de données utilisant les technologies du web sémantique, ont amené à réécrire le modèle générique sous forme d'ontologie et à opérer son alignement avec des référentiels répandus dans le monde de la conservation des biens patrimoniaux, tels CIDOC-CRM et FRBR.[26]

Deux principes fondamentaux ont guidé l'opération de modélisation de la vision du monde propre au projet *symogih.org*. Il s'agit, d'une part, de la séparation entre la production des données et la problématique de recherche qui guide leur collecte. Certes, toute production de données trouve son origine dans un questionnement: toutefois, les connaissances stockées informatiquement doivent être modélisées de la manière la plus objective possible, afin de permettre leur réutilisation pour de nouvelles recherches. D'autre part, il est nécessaire de procéder à une atomisation, c'est-à-dire à une décomposition des connaissances en éléments correspondants à des propositions simples et autonomes. À titre d'exemple, nous utiliserons la proposition « Galileo Galilei enseigne les mathématiques à l'Université de Padoue entre 1592 et 1610 » qui exprime une connaissance élémentaire.

Au centre de la représentation simplifiée de l'ontologie dans la figure 1. se trouvent les deux classes principales: la classe *Object* et la classe *KnowledgeUnit*. La première regroupe tous les 'objets' possédant une identité propre et qui subsistent comme dans la durée, qu'il s'agisse d'êtres concrets (tels une personne, une maison, un manuscrit) ou abstraits (tel un concept, une entité bibliographique, une profession). Dans la proposition retenue comme exemple, on reconnaît les objets 'Galilée', la discipline des 'mathématiques', 'l'Université de Padoue' ainsi que, de manière indirecte, le lieu 'Padoue'. Chaque objet est identifié par un *Uniform Resource Identifier* (URI) stable et déréférençable sur le site du projet,[27] et par une notice qui exprime succinctement ses caractéristiques essentielles afin que d'autres chercheurs comprennent aisément de quel objet il s'agit. Dans la fig. 1, les objets sont regroupés en dix sous-classes de la classe *Object* (par ex. *Actor, Collective Actor, Abstract Object*, etc.) qui sont construites de la manière la plus objective et abstraite possible afin d'être adaptées à tous les contextes de recherche.

La seconde classe regroupe les 'unités de connaissance' (classe *KnowledgeUnit*, cf. fig. 1) définies en tant qu'assertions de l'historien qui décrivent une relation subsistant entre objets située dans le temps et, éventuellement, dans l'espace. Ces assertions sont atomisées et conçues de la

25 F. BERETTA, P. VERNUS, *Le projet SyMoGIH et la modélisation de l'information: une opération scientifique au service de l'histoire*, « Les Carnets du LARHRA », I, 2012, pp. 81-107 (http://halshs.archives-ouvertes.fr/halshs-00677658).

26 F. BERETTA, *L'interopérabilité des données historiques et la question du modèle: l'ontologie du projet SyMoGIH*, in *Quels enjeux numériques pour les médiations scientifique et culturelle*, sous la dir. de J.-L. Minel, Paris, 2016 (sous presse).

27 Cf. par exemple l'URI de Galileo Galilei, http://symogih.org/resource/Actr161 .

manière la plus objective possible afin de permettre leur réutilisation dans d'autres contextes. La proposition qui fait état de l'enseignement de Galilée à Padoue à un moment donné est un exemple d'information atomisée ou unité de connaissance qui met en relation, avec un sens bien précis – l'enseignement – et pendant une période donnée, une personne, une institution et une discipline. Une instance de la classe *KnowledgeUnitType* est définie pour chaque type d'information qu'on souhaite stocker: elle explicite le sens des données produites et permet de comprendre l'articulation des objets qui entrent en relation au sein de la connaissance, la participation de chacun d'entre eux étant définie par un rôle précis (classes *Role* et *RoleType*).[28.]

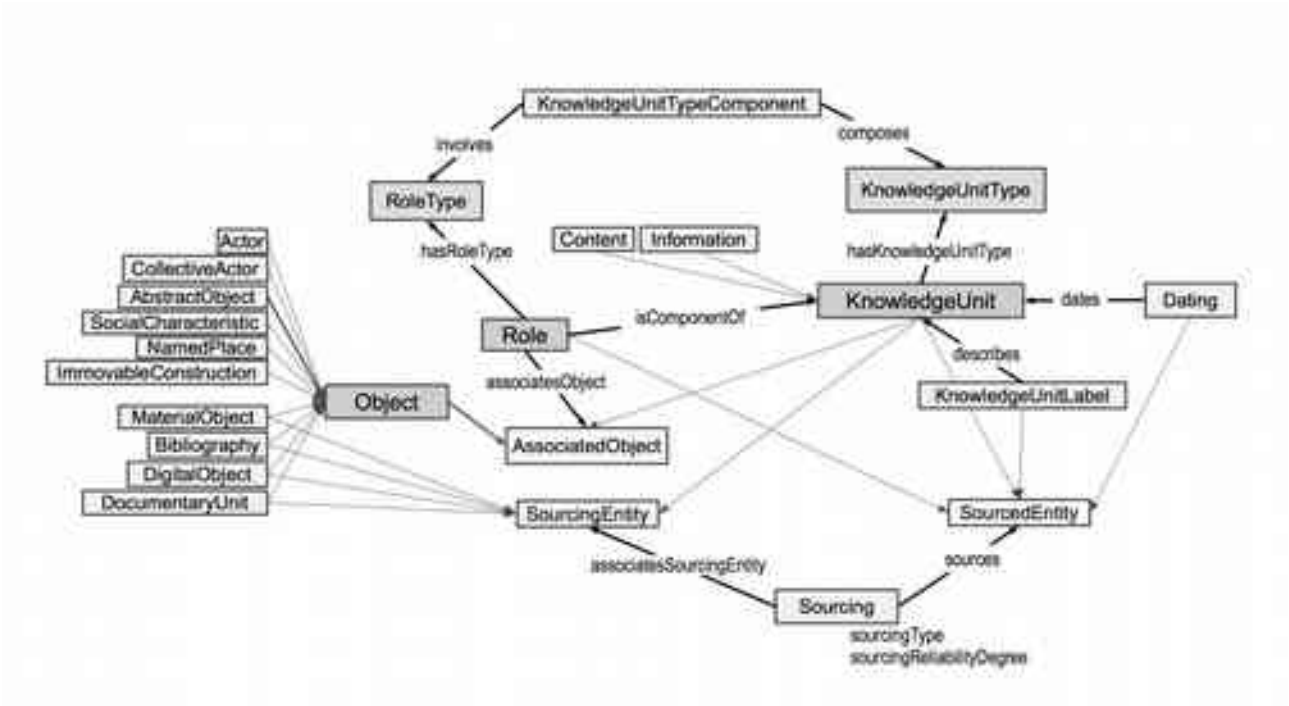


Figure 1. L'ontologie du projet *symogih.org* - version 0.2.1

L'ensemble des instances des types d'unités de connaissance est publié sur le site du projet. Ce vocabulaire structuré décrivant le monde historique s'enrichit progressivement, de manière collaborative, afin de modéliser de nouveaux types d'informations au fur et à mesure de l'élargissement du projet et des souhaits des participants: le système d'information est donc ouvert et évolutif. Les unités de connaissance, qui expriment les assertions des historiens concernant les caractéristiques des objets à un moment donné du temps, ou les relations qui subsistent entre objets, sont elle-mêmes munies d'un identifiant sous forme d'URI[29] et peuvent ainsi être mobilisées dans la

28 Cf. le type d'information "Enseignement", http://symogih.org/resource/TyIn97 . Sur la même page on a également accès aux données publiques produites en utilisant cette instance du modèle.

29 http://www.symogih.org/resource/Info94542 est l'URI de l'unité de connaissance proposée comme exemple,

production de nouvelles informations au même titre que les objets. En d'autre termes, une connaissance peut être utilisée dans une relation de causalité afin d'expliquer les origines de telle autre connaissance, par exemple une nomination ou un voyage.

En termes d'interopérabilité, l'ontologie du projet *symogih.org* présente la même structure cognitive que la *Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE),*[30] une ontologie de haut-niveau conçue comme moyen d'étudier les structures essentielles du langage naturel en tant que compréhension de la réalité.[31] En particulier, la catégorie des *endurants*, entités qui subsistent avec la même essence dans le temps, tels les objets physiques, les concepts ou les humains, équivaut à la classe *Objects* de l'ontologie *symogih.org*, tandis que les *perdurants* de DOLCE, entités qui se développent dans le temps tout en se modifiant d'un instant à l'autre, tels les événements ou les processus, correspondent aux unités de connaissance: ils expriment une relation subsistant entre objets à un moment donné du temps, qu'il soit ponctuel ou étendu. La même perspective cognitive se retrouve dans le modèle du CIDOC-CRM,[32] créé pour permettre l'interopérabilité des données produites dans le domaine de la conservation des biens culturels: les classes *Persistent Item* et *Temporal Entity* correspondent respectivement aux objets et, avec quelques nuances, aux unités de connaissance de *symogih.org*.

Le fait d'avoir adopté cette structure cognitive permet non seulement l'interopérabilité avec les données produites selon le modèle du CIDOC-CRM mais encore la réécriture de la plupart des données issues d'autres modèles sous forme de relations entre objets, situées dans l'espace et dans le temps, en mettant ainsi en relation les connaissances produites par les historiens au sein de l'environnement numérique du projet *symogih.org* avec les données disponibles sur internet.[33]

DES DOCUMENTS AUX TEXTES NUMÉRIQUES ANNOTÉS

Une fois mise en place une ontologie qui permet de modéliser le monde historique, les objets qu'il contient et leurs relations réciproques, on peut aborder l'encodage sémantique des textes. La première étape consiste dans la transcription des documents vers un format numérique. Dans le contexte du projet *symogih.org*, visant la production collaborative et la réutilisation des données, il a paru judicieux d'adopter la transcription et le balisage des textes en XML (*Extensible Markup Language*) en suivant les recommandations de la *Text encoding initiative* (TEI). Largement utilisées dans les milieux bibliothéconomique et académique depuis 1994, ces *Guidelines* permettent, entre

concernant l'enseignement de Galilée à Padoue.

30 Voir le résumé des résultats du projet: http://cordis.europa.eu/result/rcn/41438_en.html et la présentation dans http://en.wikipedia.org/wiki/Upper_ontology#DOLCE_and_DnS .

31 C. MASOLO, S. BORGO, A. GANGEMI, N. GUARINO, A. OLTRAMARI, *WonderWeb Deliverable D18 Ontology Library (final)*, Trento, Laboratory For Applied Ontology, 2003, p, 13 (téléchargeable en version PDF depuis le site http://wonderweb.man.ac.uk/deliverables.shtml).

32 Publié pour la première fois sous forme complète en 1999, CIDOC-CRM est devenu en 2006 la norme ISO21127: www.cidoc-crm.org/official_release_cidoc.html .

33 BERETTA, *L'interopérabilité des données historiques, cit.* (note 26).

autres, de décrire la structure des textes numériques, de gérer l'apparat critique et l'établissement du texte, ainsi que son annotation sémantique.[34] Comme elles s'organisent en une vingtaine de modules, et comprennent plus de cinq cent éléments permettant d'encoder le texte, appelés 'balises', il est indispensable pour chaque projet utilisant ces recommandations de définir un schéma d'encodage spécifique et de choisir les modules et les balises qui seront effectivement utilisées. Par conséquent, un guide d'encodage a été mis en place qui permet d'effectuer le lien entre les deux sémantiques: celle de la TEI et celle du projet *symogih.org*. Le guide est accessible publiquement sous forme de wiki dans le manuel des utilisateurs du projet.[35]

Le schéma d'encodage retenu vise l'interopérabilité au sein de la plate-forme et possède donc une dimension générique: les modules TEI utilisés sont réduits au strict minimum et une partie des métadonnées des documents numériques est géré directement dans l'ontologie du projet. Sur cette base on produit des textes numériques qui peuvent résulter soit de la transcription directe de documents, soit de la récupération de textes numériques déjà existants, telles les reproductions d'ouvrages imprimés disponibles sur internet. Les documents sont mis en forme selon le schéma d'encodage préconisé. Tel est le cas de l'annotation en cours de réalisation de la correspondance de Galilée, s'inscrivant dans le contexte du projet *Society Religion Science* présenté précédemment. Les textes de la correspondance issus de l'Édition nationale des œuvres de Galilée ont été mis à disposition du public par le projet *Liber Liber* sous la forme de documents au format RTF, distribués sous licence *Creative commons.*[36] Ils ont été transformés en documents XML et éclatés lettre par lettre afin d'être réutilisés et enrichis grâce à une annotation sémantique qui se réfère aux objet et aux connaissances produites au sein l'ontologie du projet *symogih.org*.

La première étape de mise en forme des textes prévoit un encodage minimaliste permettant de distinguer des subdivisions, des paragraphes, des listes, etc. Dans la perspective de la TEI, mais aussi dans la tradition des éditions critiques de documents, un texte est toujours un *objet construit*: il n'est pas naturellement donné, il résulte d'une lecture particulière du document qui peut être davantage orientée vers une reproduction la plus fidèle possible de l'original, comme dans les éditions diplomatiques, ou vers un accès facilité au texte, orienté vers la lisibilité. L'historien peut donc décider, en fonction de la finalité de son projet, d'enrichir la mise en forme de base retenue pour tous les projets en ajoutant, par exemple, les modules de la TEI qui permettent de traiter la reconstitution critique du texte.

Après cette étape de structuration et de mise en forme, on procède à l'annotation sémantique qui s'articule en deux parties, en accord avec la structure de l'ontologie présentée ci-dessus. Dans une première étape, on identifie les entités nommées, chaînes de caractères qui nomment ou se

34 http://www.tei-c.org/Guidelines/ .
35 https://groupes.renater.fr/wiki/symogih/symogih_manuel/edition_de_textes_en_xml-tei .
36 http://www.liberliber.it/online/autori/autori-g/galileo-galilei/ .

réfèrent aux objets définis dans l'ontologie. Pour ce faire on utilise les balises « name » ou « rs » de l'espace de noms TEI et on associe l'objet correspondant de l'ontologie *symogih.org* grâce à l'attribut *@ref* qui contiendra l'URI de l'objet, créé préalablement dans la plateforme. On aura par exemple ces deux formes d'encodage qui permettent d'identifier la personne de Galilée dans le texte:

```
<name ref="Actr161">Galileo Galilei</name>
<rs ref="Actr161">Filosofo e <rs type="commonNoun" ref="SoCh1468">Matematico</rs> Primario del Serenissimo Duca di Toscana</rs>
```

Dans le premier cas, Galilée est mentionné explicitement, et son nom peut directement être associé à son identifiant sous forme d'URI, *http://www.symogih.org/resource/Actr161*, abrégé dans l'encodage. Dans le deuxième cas, la lecture par le chercheur permet d'identifier le mathématicien du grand-duc de Toscane même s'il n'est pas mentionné nommément: l'encodage sémantique permettra de retrouver toutes les occurrences de la personne de Galilée dans les textes et, dans le deuxième cas – qui ajoute un deuxième niveau d'encodage en annotant également la qualité de 'mathématicien' (SoCh1468) –, on pourra également analyser les concepts qui ont été utilisés dans les documents afin de qualifier la personne en question.[37]

Pour accompagner ce processus d'annotation sémantique, le projet *symogih.org* propose d'utiliser la logiciel de textométrie TXM.[38] L'analyse textométrique permet d'explorer le lexique des textes et d'en mettre en évidence les spécificités grâce à des outils tels les index lexicaux, les concordances et les cooccurrences. La figure 2 montre le lexique lié au concept de « mathématiques » dans le corpus de la correspondance de Galilée de la fin du 16e siècle jusqu'à 1620.

37 Pour plus de détails concernant la pratique d'encodage, le lecteur pourra se référer au manuel d'encodage cité ci-dessus à la note 35.

38 http://textometrie.ens-lyon.fr/ .

Matematico	83	Matematica	7	matematicam	1
matematiche	44	Mathematica	4	matematice	1
matematico	44	mathematicarum	4	matematicha	1
Mathematico	39	mathematicorum	4	matematicho	1
matematici	31	mathematum	4	Matematicho	1
matematica	27	mathematicae	3	Matematici	1
mathematici	26	mathematicam	3	mathemata	1
mathematico	21	matemateci	2	Mathemata	1
Mathem	16	mathem	2	Mathemathices	1
mathematiche	15	mathematicamente	2	mathemathiche	1
mathematica	14	mathematicas	2	Mathematicae	1
Matem	11	mathematice	2	Mathematicarum	1
mathematicis	11	mathematicos	2	Mathematices	1
Mathematicus	11	Mathematicum	2	Mathematici	1
Matematiche	9	Mathematum	2	mathematicum	1
Mathematiche	9	matemathici	1	mathematicus	1

Figure 2. Lexique lié au concept de ‘mathématiques’

Le corpus étant en deux langues, latin et italien, il n’est pas possible d’avoir directement recours à la lemmatisation, c’est-à-dire au regroupement de toutes les formes d’un mot à une forme canonique appelée lemme. Cette exploration grâce au logiciel TXM permet de relever les différentes utilisation du concept, sous forme de substantif ou d’adjectif, et déclinées en termes de discipline ou de qualité propre aux spécialistes des mathématiques. Ce n’est que la lecture du document par l’historien qui permettra de choisir d’encoder la forme abrégé ‘Matem.’ de cette façon:

```
<rs type="commonNoun" ref="SoCh1468">Matem.</rs>
```

en identifiant dans ce cas, en fonction du contexte, le caractère social « mathématicien ».

L’enrichissement sémantique produit par cette annotation permettra de regrouper les mentions des mêmes objets dans les textes et d’en explorer le contenu grâce à des requêtes utilisant le langage xQuery puis d’en visualiser les résultats en utilisant un logiciel approprié. On pourra, par exemple, extraire toutes les cooccurrences de concepts à l’intérieur de structures des textes tels les paragraphes (balise « p »), les phrases (balise « s ») ou des segments construits par le chercheur (balise « seg »). Ces mêmes portions et subdivisions du texte peuvent être annotées avec l’attribut *@ana* (analyse) afin d’en indiquer la ou les thématiques principales tout en pointant vers le concept

défini dans l'ontologie: un paragraphe annoté, par exemple, sous la forme « ana="AbOb213" » contiendra un texte concernant les mathématiques. En réinjectant de manière appropriée cette annotation dans le logiciel TXM, il sera possible d'étudier le vocabulaire spécifique des textes concernant les mathématiques par rapport à d'autres portions du corpus concernant d'autres sujets.

Dans une deuxième étape de l'annotation sémantique, il s'agit de découper dans le texte les unités de connaissance qui mettent en relation les objets ou qui en expriment les caractéristiques. Pour ce faire, on utilise les balises évoquées précédemment qui permettent d'identifier les phrases ou les segments. Ces portions de texte, de préférence de longueur réduite afin de prendre en compte le principe d'atomisation exposé ci-dessus, contiennent des objets, sous forme d'entités nommées identifiées par le chercheur, que les textes mettent en relation au même titre que les unités de connaissance produites au sein de l'ontologie. Le sens de cette relation peut être explicité en créant une donnée dans l'ontologie et en associant l'identifiant de la donnée à la portion de texte qui vient d'être délimitée grâce à l'attribut *@ana*.

```
<s ana="Info94542"><name ref="Actr161">Galileo Galilei</name> enseigne les <rs
type="commonNoun" ref="AbOb213">mathématiques</rs> à l'<name ref="CoAc54">Université de
Padoue</name> <date from="1592" to="1610">entre 1592 et 1610</date></s>
```

Dans ce cas, le texte balisé et la donnée de l'ontologie représentent les deux faces d'une même unité de connaissance identifiée par le chercheur dans le texte.

Une autre approche possible consiste à encoder directement l'unité de connaissance dans la portion de texte identifiée, sous cette forme:

```
<s ana="TyIn97"><name ref="Actr161" ana="TyRo12">Galileo Galilei</name> enseigne les <rs
type="commonNoun" ref="AbOb213" ana="TyRo131">mathématiques</rs> à l'<name ref="CoAc54"
ana="TyRo21">Université de Padoue</name> <date from="1592" to="1610">entre 1592 et
1610</date></s>
```

On indique ainsi la typologie de la connaissance encodée et on explicite celles des rôles respectifs qui reviennent à chacune des entités nommées mises en relation par la connaissance: ce procédé est réalisé en utilisant les identifiants URI des types d'unités de connaissance et des types de rôles définis dans l'ontologie.[39] Sur cette base, il est possible de développer des méthodes d'annotation semi-automatiques de textes et d'extraction de connaissances qui permettront d'enrichir l'ontologie en produisant – grâce à des requêtes appropriées en langage xQuery – un volume important de données.[40] Afin d'éviter de trop charger le texte numérique, on peut adopter une méthode d'encodage dite 'débarquée' ou 'externalisée' (*stand-off*): dans ce cas, on identifie chaque balise insérée dans le texte grâce à un identifiant unique auquel vont se référer les annotations: celles-ci ne seront pas stockées dans le texte annoté lui-même mais dans un autre texte numérique ou dans un

39 Cf. la note 29 ci-dessus.

40 F. BERETTA F., *Exploration d'un corpus de notices biographiques: identification d'entités nommées, extraction de connaissances historiques et visualisation avec la méthode du projet* symogih.org, Journées « Big Data Mining and Visualization »: Focus sur les Humanités Numériques dans le Big Data, Lyon, France, juin 2015, https://halshs.archives-ouvertes.fr/halshs-01166424 .

format RDF.[41]

Les textes encodés en XML/TEI sont déposés sur un serveur afin de permettre leur exploration en lien avec l'ontologie stockée dans la base de données relationnelle ou publiée au format RDF. Il est ainsi possible de préparer des requêtes préconfigurées que les chercheurs utiliseront dans l'espace de la plate-forme dédié à leur corpus. Ils pourront ainsi suivre en temps réel l'avancement de leur encodage et avoir une vue d'ensemble sur leur manière d'utiliser les balises, sur les effectifs et la typologie des objets et des unités de connaissance annotées dans les textes, sur les relations qui subsistent dans les textes entre ces objets, sur le vocabulaire utilisé, etc. Au terme de ce processus, les textes enrichis pourront être publiés, si souhaité, sur le *Portail XML* présenté ci-dessus, afin de rendre accessible l'annotation sémantique à d'autres lecteurs.

CONCLUSION

Les expériences effectuées avec la méthode d'encodage sémantique des textes proposée au sein du projet *symogih.org*, de même que la mise en place de la plate-forme d'analyse et de publication des documents annotés, conduisent à une triple constat.

Premièrement, le travail effectué sur les textes au sein de cet environnement numérique entraîne une lecture plus précise des documents. D'une part, il est nécessaire de réfléchir à la structuration du texte dans son ensemble, ce qui oblige à s'interroger sur la finalité de l'établissement du texte et de l'encodage, et donc à expliciter les critères de production et de mise en forme du texte numérique en relation avec le document original. D'autre part, l'annotation sémantique impose de définir avec précision, dans l'ontologie, les concepts qu'on souhaite encoder en lien avec la problématique de sa propre recherche. Il sera en même temps possible d'explorer de manière fine le vocabulaire utilisé par les documents au sujet de ces concepts. Cette démarche méthodologique conduit donc à une analyse plus approfondie des sources étudiées.

Deuxièmement, l'annotation sémantique permet d'opérer plus facilement sur les textes numériques à l'aide de logiciels d'exploration lexicale, d'annotation semi-automatique d'entités nommées, de visualisation et de fouille de données. L'environnement mis en place permet une emprise efficace sur des corpus numériques qui peuvent être volumineux: à partir de cet encodage, une extraction semi-automatique de connaissances permettra d'injecter de nouvelles données dans l'ontologie, en renforçant le lien qui la relie aux documents originaux.

Troisièmement, la mise en ligne des textes dans un environnement de travail dédié, ainsi que le fait d'avoir adopté des formats génériques permettent de réaliser aisément des éditions annotées de textes numériques, en mettant ainsi à la disposition d'autres chercheurs et du public le résultat de

41 Voir, entre autres, P. BAŃSKI, *Why TEI stand-off annotation doesn't quite work and why you might want to use it nevertheless*, « Proceedings of Balisage: The Markup Conference 2010», V, 2010 (http://www.balisage.net/Proceedings/vol5/html/Banski01/BalisageVol5-Banski01.html) and H. A. CAYLESS, *Rebooting TEI Pointers*, « Journal of the Text encoding initiative », VI, 2013 (https://jtei.revues.org/907).

l’enrichissement sémantique des documents opéré par l’historien. Dans un habillage moderne et dynamique, il est ainsi possible de restituer à la collectivité, de manière facilement accessible, l’important travail d’érudition et de réflexion qui accompagne l’édition annotée des documents.