Le portail XML du projet symogih.org : un projet d'édition numérique collaborative de sources et d'informations historiques

Francesco Beretta, Rosemonde Letricot

Chapitre 7

« Le portail XML du projet symogih.org : un projet d'édition numérique collaborative de sources et d'informations historiques »

Francesco Beretta[1], Rosemonde Letricot[1]

[1] Pôle histoire numérique, Laboratoire de Recherche Historique Rhône-Alpes (LARHRA – UMR5190)

1. Introduction

Les éditions numériques de documents dans le champ des sciences humaines se sont fortement développées ces dix dernières années sous l'impulsion des communautés scientifiques[1] . Qu'elles soient à des fins de médiation patrimoniale ou dans le but d'expérimenter de nouveaux parcours d'analyse, ces initiatives ont permis à la fois une mise à disposition massive de corpus informatisés mais aussi le développement de nouvelles pratiques numériques tant au niveau de la mise en forme de contenus que de l'architecture d'information qui les porte. Les fonctionnalités offertes au lecteur en matière de visualisation de données, d'interaction et de recherche demandent également une réflexion sur les procédés

[1] Nous pensons particulièrement aux plates-formes mises en ligne par des institutions comme l'École nationale des chartes (http://elec.enc.sorbonne.fr/), l'Institut de Recherche et d'Histoire des Textes (http://www.cn-telma.fr/), l'École normale supérieure (liste des éditions numériques sur le site http://ahn.ens-lyon.fr).

techniques qui seront développés et dont dépendra la réception des textes numériques[2].

Le but de ce chapitre est de proposer un retour d'expérience sur la mise en place d'une plate-forme d'édition de sources en lien avec la recherche historique : le « Portail XML »[3] déployé en 2015 par l'équipe du Pôle histoire numérique (PHN) du Laboratoire de Recherche Historique Rhône-Alpes (LARHRA – UMR5190). Cette plate-forme permet de produire et de publier des textes enrichis en unissant une annotation sémantique qui utilise la formalisation XML/TEI d'une part et les données issues de la base de données collaborative du projet symogih.org[4] d'autre part. Cet environnement numérique au service de la recherche permet le stockage et la mutualisation de données géo-historiques et facilite le recours aux pratiques computationnelles comme les analyses statistiques, spatiales ou des réseaux sociaux.

Si l'effort des chercheurs, des ingénieurs et des doctorants qui ont créé et qui portent le projet symogih.org s'est concentré, dans une première phase, sur la production de données structurées à partir de l'étude des sources historiques, la question s'est ensuite posée de comment établir ou, en fait, rétablir le lien subsistant entre les données tirées des documents et les textes eux-mêmes, afin de profiter de la richesse de l'annotation sémantique en lien avec le vocabulaire et la structure des sources. Ce chapitre portera sur la manière dont sont réinvesties ces données structurées au regard de deux projets d'éditorialisation collaborative de textes. Le premier concerne une édition critique numérique des *Mémoires* de Léonard Michon[5] relatant la vie politique, culturelle, religieuse et sociale des élites bourgeoises lyonnaises du début du XVIII[e] siècle. Le second projet est consacré à une annotation sémantique de la correspondance de Galileo Galilei et s'inscrit dans un projet plus

---

2 Afin d'évaluer la qualité des nouvelles éditions numériques, un périodique en ligne a été créé en 2014: RIDE, A review journal for digital editions and resources, http://ride.i-d-e.de/ . Concernant l'évaluation des publications de données de la recherche, voir également le périodique en ligne *Scientific data*, http://www.nature.com/sdata/ .

3 http://xml-portal.symogih.org .

4 http://symogih.org/ .

5 Édition menée dans le cadre d'une recherche doctorale : 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.

large de mise à disposition de ressources concernant l'histoire intellectuelle à l'époque moderne[6].

Notre propos consistera à présenter la méthodologie générale adoptée pour mettre en place le nouveau portail XML du projet symogih.org et à réfléchir sur l'évolution des processus d'éditorialisation des documents et d'annotation des textes numériques. Nous aborderons les principes généraux de son architecture d'information qui mêle une pratique d'éditions de textes et une démarche collaborative de curation de données géo-historiques. Puis nous soulèverons quelques points de discussion à partir des premiers résultats de la plate-forme concernant le phénomène d'éclatement que produit le numérique sur le document. Passant d'un état unitaire matériel à un ensemble de contenus numériques, calculés et recomposés par nos terminaux de consultation [BAC 04], se pose alors la question des procédés d'éditorialisation et de la cohérence des espaces de publication mis en œuvre par rapport à la narrativité de l'objet initial.

2. Le projet *symogih.org* et l'interopérabilité des données géo-historiques

2.1. La gestion collaborative de données géo-historiques

Le projet « Système modulaire de gestion de l'information historique » (SyMoGIH), désormais appelé projet symogih.org du nom du site web qui publie l'ontologie et une partie des données[7], est né en 2008 de la volonté de quelques historiens du LARHRA de mutualiser les données structurées produites au cours de leur recherche afin de les mettre à disposition et de permettre leur réutilisation par d'autres collègues. Cette démarche s'inscrit dans la logique de la « curation des données »[8] dans le sens d'un enrichissement et d'une amélioration constantes des

---

[6] http://srs.symogih.org/ . Francesco Beretta, chargé de recherche au CNRS, est l'un des principaux porteurs du projet symogih.org et travaille sur l'histoire intellectuelle en Europe à l'époque moderne et contemporaine ainsi que sur la transformation digitale de la méthode historique. Cf. ses publications récentes sur HalSHS: https://halshs.archives-ouvertes.fr/search/index/?qa[auth_t][]=Francesco+Beretta&sort=producedDate_tdate+desc .

[7] http://www.symogih.org .

[8] Voir l'entrée “data curation” de l'édition anglaise de Wikipedia (https://en.wikipedia.org/wiki/Data_curation, consulté le 15 février 2016) et la bibliographie qui y est mentionnée.

données afin de garantir dans la durée leur qualité, leur accessibilité et leur préservation.

Cette expérience a eu du succès dans la mesure où 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 l'environnement numérique de recherche afin de produire et de mutualiser leurs données. À titre d'exemple, les données d'un projet tel SIPPAF[9], consacré à la mise en place d'un système d'information consacré aux patrons français (XIX^e^-XX^e^ siècle) et financé dans une première phase par l'Agence nationale de la recherche, 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[10]. Les données des deux projets se recoupent partiellement ce qui entraîne une amélioration progressive et mutuelle, en dépit de la fin de la période de financement du premier projet. Ces données ont été produites selon le modèle générique qui sera présenté ci-dessous et sont hébergées dans le même entrepôt : le travail de curation peut donc être effectué aisément dans une même interface et les données mises-à-jour seront ensuite publiées automatiquement sur les sites respectifs.

Au sein du projet symogih.org est également disponible un lot de données géo-historiques, accessible depuis le site de partage de ressources GEO-LARHRA[11], ainsi qu'un point d'accès SPARQL, limité à une portion restreinte des données disponibles dans la plate-forme[12]. En effet, l'ensemble du silo de données n'est pas entièrement accessible dans la mesure où il faut prendre en compte le lien étroit subsistant entre la production des données et les recherches en cours des participants au projet. Au moment de la création d'une donnée, il est donc laissé à son producteur de décider du niveau de communicabilité.

Il existe ainsi trois niveaux d'ouverture cumulables successivement : l'un ouvrant les données à l'ensemble des utilisateurs enregistrés qui participent au projet ; un deuxième autorisant leur publication en lecture sur la plate-forme symogih.org et sur tout site web dédié à un projet spécifique qui souhaite rendre visible telle information ; un troisième procédant à leur publication au format du *Resource*

[9] Cf. le site du projet: http://www.patronsdefrance.fr/ .

[10] http://siprojuris.symogih.org/ .

[11] http://geo-larhra.ish-lyon.cnrs.fr/ .

[12] Cf. cette page: http://symogih.org/?q=rdf-publication .

Description Framework (RDF) à travers un point d'accès SPARQL qui permet d'interroger directement les données dont la structure est explicitée par une ontologie documentée publiquement (voir ci-dessous). Les données sont mises à disposition sous licence *Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International* comme l'ensemble des ressources du projet. Afin de garantir la stabilité du système d'information, les données, une fois publiées, ne peuvent pas être dé-publiées mais seulement modifiées et améliorées.

L'ouverture d'un *SPARQL endpoint* donnant accès à une partie des données produites par les chercheurs permet de les rendre directement interopérables avec l'ensemble des données disponibles sur le web, publiées au format RDF. S'ouvrent ainsi des perspectives inédites pour la recherche en histoire : grâce au requêtage simultané sur plusieurs entrepôts de données ou à l'application de technologies de raisonnement sémantique et de fouille, il est possible d'élargir et d'enrichir progressivement le lot de données soumis à analyse dans le cadre d'un questionnement de recherche en vue de produire de nouvelles connaissances [BER 15c].

2.2 Du modèle relationnel générique à une ontologie interopérable

La mise en place d'une plate-forme de production et de curation collaborative de données historiques, fondée sur une base de données relationnelle utilisant la technologie PostgreSQL, imposait dès le début d'adopter une modélisation générique, susceptible de s'adapter à tout type d'information quelles que soient les thématiques de recherche ou la période étudiée, sans devoir restructurer la base pour chaque nouveau projet [BER 12]. 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, nous ont amenés 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, tel CIDOC-CRM et FRBR [BER 15b].

Deux principes fondamentaux ont guidé l'opération de modélisation. 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 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. Chaque objet est identifié par un *Uniform Resource Identifier* (URI) stable et déréférençable sur le site du projet[13], et par une notice qui exprime succinctement ses caractéristiques essentielles. Les objets sont regroupés en onze sous-classes de la classe *Object* (par ex. *Actor, Collective Actor, Abstract Object,* etc.) qui sont construites de la manière la plus objective possible afin d'être adaptées à tous les contextes de recherche.

La seconde regroupe les unités de connaissance (classe *KnowledgeUnit*, cf. figure 1.) qui sont définies en tant qu'assertions de l'historien et qui décrivent une relation atomisée entre objets, tout en situant cette relation dans le temps et, éventuellement, dans l'espace. 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. 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 impliqués dans la connaissance, la participation de chacun d'entre eux étant définie par un rôle précis (classes *Role* et *RoleType*). Dans le cas d'une information de type « enseignement », on aura ainsi la personne qui enseigne, la discipline enseignée, l'institution d'enseignement, etc.[14].

---

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

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

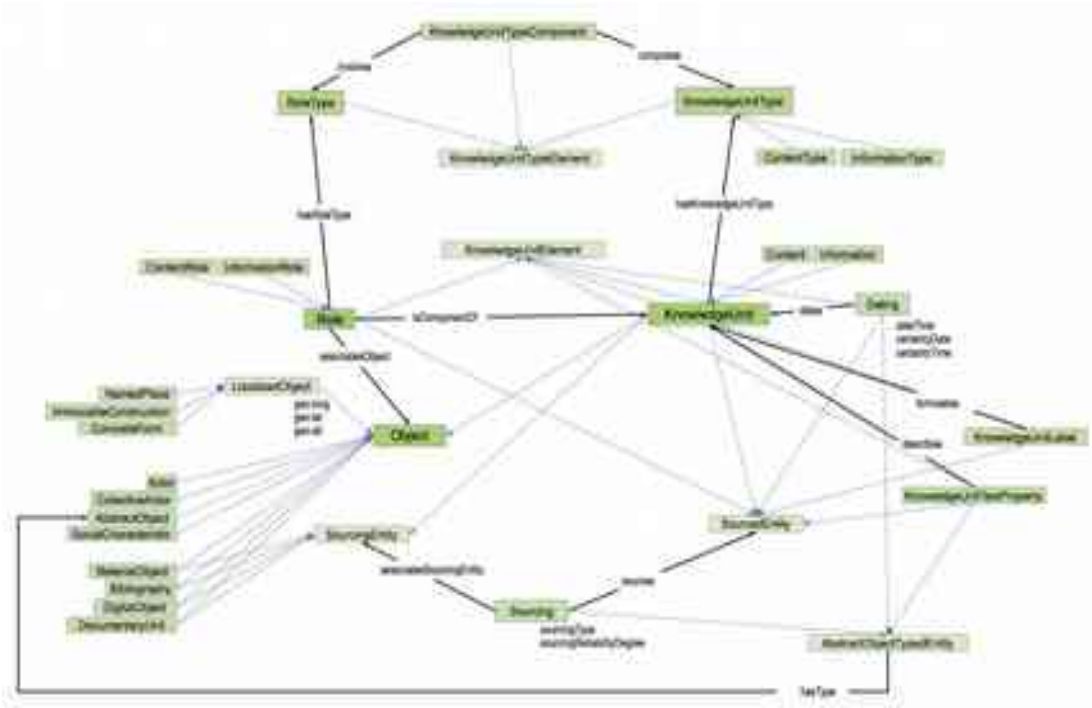

Figure 1. L'ontologie du projet symogih.org - version 0.2.1

L'ensemble des instances des types d'unités de connaissance est publié sur le site du projet. Elles s'enrichissent progressivement 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 entre objets, sont elle-mêmes munies d'un URI[15] et peuvent ainsi être mobilisées dans la production de nouvelles informations au même titre que les objets.

En termes d'interopérabilité, l'ontologie du projet symogih.org présente la même structure que la *Descriptive Ontology for Linguistic and Cognitive Engineering*

[15] http://www.symogih.org/resource/Info94542 est l'URI de l'unité de connaissance proposée comme exemple, concernant l'enseignement de Galilée à Padoue.

(DOLCE)[16], une ontologie de haut-niveau conçue comme moyen d'étudier les structures essentielles du langage naturel en tant que compréhension de la réalité [MAS 03, p.13]. 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*, tandis que les « perdurants » de DOLCE, entités qui se développent dans le temps tout en se modifiant d'un instant à l'autre, tel les événements ou les processus, correspondent aux unités de connaissance de l'ontologie symogih.org : ils expriment une relation subsistant entre objets à un moment précis du temps, qu'il soit ponctuel ou étendu. La même perspective cognitive se retrouve dans le modèle du CIDOC-CRM, créé pour permettre l'interopérabilité des données produites dans le domaine de la conservation des biens culturels, dont les classes *Persistent Item* et *Temporal Entity* correspondent respectivement aux objets et, avec quelques nuances, aux unités de connaissance de symogih.org[17].

Le fait d'avoir adopté cette structure cognitive dans l'ontologie symogih.org 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 [BER 16]. Il est donc possible d'enrichir et de mettre en dialogue les connaissances produites par les historiens au sein de la plate-forme du projet avec les données disponibles sur le web, en particulier en les réutilisant dans le contexte d'éditions numériques de textes[18].

3. Les procédés d'éditorialisation

L'approche de modélisation des connaissances historiques qui vient d'être présentée permet aux historiens d'extraire des sources, et en particulier des textes lus et analysés, les données structurées qu'ils vont stocker dans la plate-forme collaborative. Si on dispose d'une transcription numérique de ces mêmes textes, il est possible, voire souhaitable aux fins de la recherche, de rétablir le lien entre la donnée structurée et le passage du texte à partir duquel elle a été tirée. L'annotation

[16] 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 .

[17] 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.

[18] D'autres projets partagent cette approche: à titre d'exemple cf. [JOR 12].

des textes peut être effectuée sans l'aide d'une infrastructure numérique en ligne : le chercheur peut assembler lui-même ses données et produire des requêtes mêlant ces différentes technologies. Toutefois, dans une perspective de partage et de médiation des données scientifiques, il paraît judicieux de mettre à disposition de la communauté disciplinaire et du public le texte de la source étudiée, enrichi de l'apparat critique constitué par le travail de recherche qui va en faciliter la compréhension et l'appropriation. À cette fin, une nouvelle plate-forme de stockage et d'annotation de documents XML a été mise en place au sein du projet symogih.org. Elle a aussi pour but de faciliter le travail du chercheur en mettant à sa disposition un environnement numérique en ligne dont la maintenance est assurée par l'équipe du Pôle histoire numérique. Cette démarche s'effectue en parallèle de la capitalisation de données structurées proposée par le projet symogih.org afin de permettre une intégration directe dans le texte transcrit des données recueillies lors de la consultation des documents d'archives et bibliographiques. Il s'agit d'inscrire dans la méthodologie globale du projet et dans l'architecture de son système d'information les pratiques de transcription et d'annotation des textes propres au travail d'historien.

3.1 Architecture de la plate-forme et annotation des textes

La plate-forme repose sur une architecture de l'information qui exécute une « boucle de transformation des données en informations puis des informations en connaissances » [BRO 13]. La numérisation et l'annotation du document décrivant tant sa structure que son contenu, en identifiant par exemple les entités nommées et leurs relations sémantiques, permettent de créer de la donnée textuelle enrichie. Celle-ci prend sens à l'intérieur d'un système d'information basé sur une ontologie qui explicite les propriétés, la hiérarchie et les inter-dépendances entre les différentes informations extraites du texte numérique sous forme de données structurées. Enfin, ces informations réintégrées dans le contexte d'une édition numérique ou d'un questionnement scientifique participent à la production de nouvelles connaissances. Le schéma de la figure 2 permet de comprendre les rouages de ce processus de création de connaissances historiques dans le contexte de la plate-forme XML du projet symogih.org et montre comment s'organisent et s'interconnectent les grands ensembles techniques, technologiques et méthodologiques.

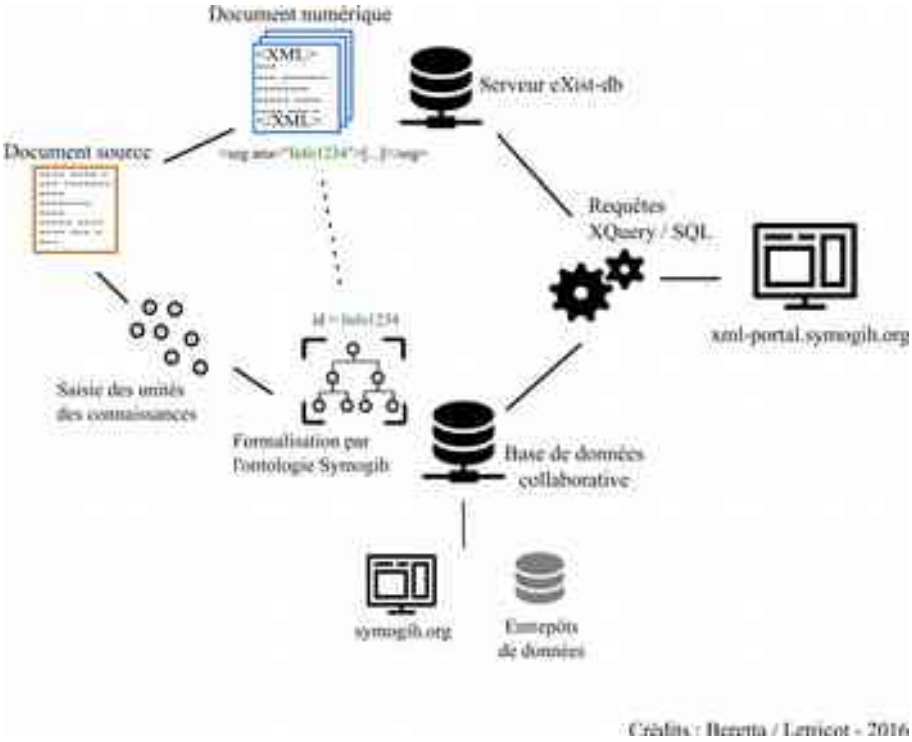

Figure 2. *Schéma des procédés d'éditorialisation du projet symogih.org*

La lecture du document original, qu'il soit sous format analogique ou numérique, amène à deux processus parallèles : d'une part, l'identification d'objets et l'extraction de connaissances qui mettent en relation ces objets selon le modèle du projet symogih.org ; d'autre part, la transcription et le balisage du texte en XML, en

utilisant les recommandations de la *Text encoding initiative* (TEI)[19]. Comme les *Guidelines* de la TEI s'organisent en une vingtaine de modules consacrés à des aspects très divers de l'encodage des textes, chaque projet utilisant ces recommandations doit définir un schéma d'encodage spécifique et choisir les modules et les balises qui seront effectivement utilisées. Dans le contexte de l'approche collaborative propre au projet symogih.org, un guide générique d'encodage a été mis en place permettant d'effectuer le lien entre les deux sémantiques : celle de la TEI et celle présidant à la production des données structurées. Le guide est accessible sous forme de wiki dans la partie publique du manuel des utilisateurs du projet symogih.org[20].

Ce schéma d'encodage 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 encodés est gérée sous forme de données structurées dans la plate-forme. Si l'appropriation des règles de la TEI et l'encodage des textes requièrent un investissement non négligeable de la part du chercheur, cette étape l'amène en retour à réfléchir de manière approfondie sur la structure et le contenu de la source. En identifiant certaines parties du texte ou chaînes de caractères, il est en effet possible d'imaginer des extractions de contenus et d'étudier la récurrence de types de balises ou d'attributs.

Après cette étape de structuration et de mise en forme des textes, on procède à leur annotation sémantique[21]. Celle-ci s'articule en deux parties, en accord avec la structure de l'ontologie du projet symogih.org présentée ci-dessus. On identifie d'abord les entités nommées : chaînes de caractères qui nomment ou se réfèrent aux objets définis dans l'ontologie. Pour ce faire on utilise les balises *name* ou *rs* de

[19] Largement utilisé dans les milieux scientifique et académique depuis sa création en 1987, ce projet propose sous forme de *Guidelines* un catalogue de balises XML qui permettent entre autres de décrire la structure des textes numériques, de gérer l'apparat critique et l'établissement du texte dans une édition, ainsi que son annotation sémantique. Site web du consortium http://www.tei-c.org (consulté le 15 février 2016).

[20] https://groupes.renater.fr/wiki/symogih/symogih_manuel/edition_de_textes_en_xml-tei .

[21] Ce processus a été l'objet d'une présentation à la conférence annuelle de la TEI en 2015 [BER 15b] .

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 plate-forme. Ensuite, il s'agit de découper dans le texte les unités de connaissance qui mettent en relation les objets ou qui expriment leurs caractéristiques. Pour ce faire, on utilise soit des éléments structurels, tel l'élément *s* qui délimite une proposition, soit des éléments relevant déjà de l'analyse sémantique, tel l'élément *seg* qui permet de définir n'importe quel segment de texte.

Ces portions de texte, de préférence de longueur réduite afin de prendre en compte le principe d'atomisation exposé ci-dessus, contiennent ainsi des objets, sous forme d'entités nommées identifiées, qu'ils mettent en relation au même titre que les unités de connaissance de l'ontologie. Le sens de cette relation peut-être explicité en créant une donnée structurée dans le système d'information et en associant son URI à la portion de texte qui vient d'être délimitée grâce à l'attribut @ana. Le texte balisé et la donnée de l'ontologie représentent ainsi les deux faces d'une même unité de connaissance, identifiée par le chercheur. Sur cette base il est possible de développer des méthodes d'extraction de connaissances et d'annotation automatisées des textes [BER 15a].

Les textes encodés en XML/TEI sont déposés sur un serveur eXist-db (cf. figure 2) afin de permettre leur exploration en lien avec l'ontologie stockée dans la base de données relationnelle grâce au langage d'interrogation xQuery. Il est ainsi possible de préparer des requêtes préconfigurées que les chercheurs utilisent dans l'espace privé de la plate-forme afin de suivre en temps réel l'avancement de l'encodage, le type de balisage effectué, les effectifs des objets et unités de connaissance associées aux textes, etc. Le processus d'annotation sémantique peut être accompagné par une analyse textométrique. Un projet en collaboration avec l'équipe TXM du laboratoire ICAR, qui développe cette plate-forme de textométrie[22], est en cours afin d'intégrer dans le logiciel TXM une interface d'annotation en lien avec l'ontologie symogih.org, ainsi que virtuellement avec toute autre ontologie[23]. Cette démarche permet d'étudier le vocabulaire et la structure des textes en lien avec les connaissances annotées et extraites par les historiens.

Enfin, les textes enrichis peuvent être publiés sur le Portail XML afin de les rendre accessibles à d'autres lecteurs. Avant de présenter quelques aspects de cette

[22] http://textometrie.ens-lyon.fr/ .

[23] https://groupes.renater.fr/wiki/txm-info/public/annotation/specs_manual_annotation .

démarche, mentionnons brièvement les spécificités de l'encodage et du traitement des données adoptés dans les deux projets retenus ici.

3.2 Les spécificités des projets Michon et Galilée

Pour l'édition critique des *Mémoires* de L. Michon, nous ne souhaitions pas produire une reproduction fidèle des pages du manuscrit mais focaliser nos efforts sur l'identification des principaux espaces d'écriture (corps du texte, marge, titre, paragraphe etc.) et des entités nommées (acteurs individuels, acteurs collectifs, lieu). En effet, l'intérêt de ce texte du point de vue de l'histoire de Lyon est de comprendre le regard que son auteur porte sur ses contemporains. La récurrence et la temporalité des occurrences de noms dans le texte sont donc primordiales pour répondre à ce questionnement. En parallèle de l'encodage XML, nous avons également engagé, à l'aide de la plate-forme symogih.org, un travail de production de données prosopographiques concernant les personnes citées et leur rôle par rapport aux événements relatés.

La structure générale du schéma XML demeure assez simple avec un *teiHeader* contenant les métadonnées de description de l'ouvrage, un élément *front* pour la partie qui contient les titres, un corps de texte contenu dans l'élément *body*. À l'intérieur de cette balise, nous subdivisons le texte pour marquer la progression chronologique en utilisant des éléments *div* rassemblant les contenus pour chaque mois de l'année. Puis nous effectuons un second découpage à l'intérieur de ceux-ci avec autant de balises *seg* que de segments d'information. Nous en déterminons la longueur par la lecture du texte ou en suivant la mise en forme donnée par l'auteur quand celle-ci est suffisamment explicite. C'est au niveau de cet élément *seg* (segment) que nous ajoutons la datation (balise *date*) et que nous procédons à une identification des entités nommées en utilisant les balises *name* ou rs. Les notes marginales sont signifiées par l'élément *note* (attribut : place="margin") auquel nous ajoutons également une datation, car du point de vue des pratiques d'écriture, il est intéressant de savoir à quel moment l'auteur a procédé à une relecture et à une correction de ses volumes.

Du point de vue de la segmentation des fichiers XML et de leur stockage sur le serveur eXist-db, nous avons respecté la chronologie du récit et non l'organisation physique du volume, nous produisons ainsi un fichier XML par année des *Mémoires*. Quand une année s'étend sur deux volumes, nous plaçons la transcription respectivement dans deux balises *text* rassemblés sous un élément *group*. Chaque

balise *text* reçoit en attribut l'identifiant attribué au volume dans la base de données. De cette manière, chaque partie du texte est associée au volume correspondant.

L'annotation sémantique de la correspondance de Galilée s'inscrit dans un projet de recherche portant sur l'histoire des sciences à l'époque moderne. Les textes de la correspondance 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*[24]. Ils ont été transformés en documents XML et éclatés lettre par lettre afin d'être réutilisés et enrichis grâce à une annotation sémantique effectuée en lien avec les données structurées de la plate-forme symogih.org, selon les principes exposés ci-dessus.

L'annotation d'entités nommées et d'unités de connaissance dans le textes des lettres ainsi que dans d'autres textes contemporains, en cours de réalisation, vise à reconstituer la dynamique des échanges et la progression des idées dans les différents réseaux de correspondants. Il s'agit de reconstituer un graphe virtuel d'objets (acteurs, concepts, événement) présents dans les textes et dans l'ontologie, et d'analyser les relations qui subsistent entre eux selon leur évolution dans l'espace et dans le temps. La publication en ligne des textes annotés permet de restituer au public, sous la même licence, le texte des lettres sémantiquement enrichi, et d'expérimenter en même temps les possibilités de visualisation du graphe virtuel des objets dont parlent les textes.

3.3. Quelques fonctionnalités du Portail XML

La partie publique du Portail XML du projet symogih.org a été développée afin présenter au public les premiers résultats de cette approche. Elle a donc une dimension essentiellement expérimentale qui connaîtra encore d'importants développements.

[24] http://www.liberliber.it/online/autori/autori-g/galileo-galilei/ .

Figure 3. *Frise chronologique de la correspondance de Galilée*

Au niveau de la page d'accueil, le menu principal en tête de la page centralise la navigation vers les différents contenus de l'édition. On accède par l'onglet “Parcourir” (ou Browse) aux différents parcours de lecture : accès à l'ensemble des textes et à leurs métadonnées, aux listes des noms de personne ou d'institution, etc. Dans l'image ci-dessus, nous voyons un accès au texte par l'intermédiaire d'une frise chronologique interactive où chaque lettre de la correspondance de Galilée y est placée selon sa date (figure 3). En cliquant sur l'une d'elle, une fenêtre contenant le texte apparaîtra au-dessus et on pourra ainsi naviguer aisément d'une lettre à l'autre tout en suivant le déroulement temporel. Des hyperliens renvoient vers l'espace de lecture présentant l'annotation sémantique et, en attendant la mise en place d'un affichage de l'ontologie à l'intérieur du portail, vers la fiche correspondante du site symogih.org.

Figure 4. *Interface de l'édition numérique des « Mémoires » de L. Michon*

Concernant les *Mémoires* de L. Michon, les développements se sont portés sur l'espace de lecture (figure 4) divisé en deux parties : l'une centrale dédiée à l'affichage du texte et une latérale où sont présentées la liste des liens avec les objets et les unités de connaissance de la base de données. Au passage de la souris sur l'une de ces informations, la partie du texte qui a été encodée et identifiée est surlignée en couleur. La table des matières à gauche, qui affiche les subdivisions mensuelles, permet de naviguer dans le texte selon son déroulement chronologique. À l'heure où nous écrivons ces lignes, l'interactivité globale reste encore limitée mais à terme

nous aimerions mettre en place un parcours de lecture fluide pour passer dans un même environnement du texte aux objets et aux unités de connaissance, pour rebondir enfin sur d'autres portions de textes, regroupées dynamiquement en fonction des intérêts du lecteur.

Figure 5. *'Fiche' d'un acteur*

Enfin, de nouveaux espaces virtuels peuvent être créés grâce à la méthode des application web composites *(mashups)* regroupant sur une seule page des données issues de l'ontologie symogih.org mais aussi provenant du web des données, telle la fiche d'un acteur dans l'édition de la correspondance de Galilée (figure 5).

4. Discussion

Développer une plate-forme d'édition numérique oblige à s'interroger tant sur les modalités de transformation du document que sur les dispositifs de médiation. En effet, le sens même du terme édition et des activités qu'il induit ont profondément été remodelés par l'arrivée du web [DUC 04]. La distinction qui existait, dans l'univers de l'édition papier, entre la mise en forme des contenus et la gestion de la publication, émanant « d'une profession qui, dès l'origine, est partagée entre deux rôles distincts : la fonction éditoriale et la fonction entrepreneuriale. » [GEN 14] se dissout peu à peu. Produire un document numérique soulève en plus de la structuration de son contenu (qu'elle soit pleinement maîtrisée ou totalement fondue dans des dispositifs WYSIWYG) la question de l'accessibilité depuis différents terminaux, l'exploitabilité potentielle, les modalités de partage, la promotion, les droits d'auteur, etc. Même dans son acception la plus large, le terme « édition » reste encore assez faible pour représenter toutes les facettes de la vie numérique d'un document. S'il ne s'agit plus seulement de préparer un texte à devenir lisible sur le web mais à en faire une donnée dans un espace connecté, ouvert et en perpétuelle recomposition, la notion d'éditorialisation telle qu'elle a été définie récemment par Marcello Vitali-Rosati [VIT 16] rend compte de la multiplexité des dynamiques autour de la production et de la publication de contenu. Cette notion s'avère particulièrement éclairante lorsqu'on réfléchit à l'interaction entre travail individuel et collectif d'annotation de textes et de production de données tel qu'il est pratiqué dans la plate-forme XML du projet symogih.org. En effet, les différents chaînons de l'éditorialisation se distribuent entre les différents espaces de collaboration autour du texte, des unités de connaissance et de leur modélisation, des données issues du web.

Toutefois, il en résulte une difficile organisation des contenus qui met en tension l'état originel du document avec sa réalité négociée par les terminaux numériques de consultation. Alors même que l'atomisation des connaissances et la structuration des contenus que nous réalisons avec l'environnement symogih.org pourraient nous permettre de nous abstraire de la matérialité de l'objet, celle-ci reste pour le lecteur, et l'historien en particulier, une source de contextualisation et de légitimation de ce qui est lu. Pour l'édition des *Mémoires* de L. Michon, nous nous sommes heurtés à ce problème par rapport au texte des notes marginales. Pour le moment, elles sont signifiées par des puces de couleur qui au clic les affichent dans des *popups*. La difficulté est que l'auteur des *Mémoires* s'est servi de l'espace libre présent sur le papier pour commenter, parfois assez longuement, le texte principal. De sorte que le mode d'affichage actuel n'est pas satisfaisant, il faudrait ajouter un espace supplémentaire sur la surface déjà réduite de l'écran, ou alors intégrer le contenu de

ces notes à l'intérieur du texte principal sans toutefois négliger l'importance « du sens dans la forme » [PED 06] et de fait rompre le contrat de lecture qui en délimite l'interprétation. Dans le champ des écrits du for privé [BAR 15] dans lequel ces *Mémoires* s'inscrivent, le rapport entre la forme et le contenu est au cœur des grilles d'analyse historiques des pratiques d'écriture, notamment sur leurs dimensions sociales et culturelles mais aussi sur le rapport à l'individu et à l'intimité.

Dans la première mouture du portail, nous avons choisi de mettre en place un affichage en deux parties avec d'un côté le texte et de l'autre une liste de mentions d'événements, de personnages et d'institutions qui, au clic, renverront vers de nouvelles pages contenant des explications plus précises. En ce sens, nous nous éloignons d'un apparat critique en notes de bas de page, comme cela se fait généralement pour les éditions critiques imprimées. Cet espace d'explicitation des contenus différent de l'habituel, peut-être un peu moins riche en termes de lisibilité, nous semble en revanche gagner en interactivité dans la mesure où les utilisateurs seront amenés à parcourir les différentes ressources par rebond ou par thématique plutôt que de procéder à une lecture linéaire du document. Reste à savoir comment le public se réappropriera *in fine* ce texte recomposé, car selon Bruno Bachimont « Le problème […] renvoie à celui de la lisibilité, et comment on passe de la matérialité de la présentation à la dynamique de l'interprétation » [BAC 07].

Les prochains développements seront axés sur une exploitation de l'hypertextualité du document, par l'ajout d'applications (un moteur de recherche, facettes thématiques) et d'objets graphiques interactifs comme un accès par graphes, une géolocalisation des événements sur un fond de carte, etc. En ce sens, l'édition électronique dans le cadre de projets scientifiques s'inscrit dans la démarche de redéfinition du document numérique tendant de plus en plus à la médiatisation.

Les choix éditoriaux qui préfigurent les modalités d'identification et d'interaction avec l'information interviennent très tôt dans la temporalité de ces projets, au moment de la création des fichiers et de la structuration des contenus [DAC 10]. Nous en avons fait l'expérience au moment de choisir la granularité du traitement numérique : par exemple, une division des segments d'information trop étendue a été remplacée par un découpage plus atomisé du texte, tout en imbriquant des segments dans d'autres segments. Cet exemple montre aussi que, s'il « n'est pas souhaitable de laisser l'informatique aux mains des seuls informaticiens : ce n'est pas au shérif de faire la loi » [MOR 07], il ne faut pourtant pas réduire le ressort des technologies numériques à la seule activité de développement informatique, ni supposer qu'il n'y ait pas percolation entre techniques, technologies et humanités. Le chercheur en

humanités numériques doit s'impliquer dans toute la chaîne de traitement des données et même s'il ne peut pas maîtriser toutes les technologies en comprendre les enjeux en lien avec sa problématique.

5. Conclusion

Les projets d'édition numérique autour des *Mémoires* de L. Michon et de la correspondance de Galilée présentés ici montrent que l'annotation sémantique effectuée lors de la lecture et de l'analyse des textes permet de leur ajouter une couche de connaissances qui en enrichit l'interprétation et en facilite l'accès. De plus, le fait d'utiliser les données issues d'une plate-forme collaborative comme le projet symogih.org, et de surcroît construites dans une logique d'interopérabilité avec d'autres données disponibles sur le web, inscrit l'opération d'édition et d'annotation dans une logique d'éditorialisation qui tire profit des ressources existantes et ouvre des perspectives entièrement nouvelles. Enfin, la démarche elle-même d'annotation peut devenir collaborative et la publication en ligne des documents annotés met à la disposition d'autres chercheurs et du public des matériaux textuels enrichis qui, grâce aux nouvelles technologies, s'ouvrent à des parcours de lecture renouvelés, dynamiques et interactifs.

6. Bibliographie

[BAC 04] BACHIMONT B., CROZAT S., « Instrumentation numérique des documents : pour une séparation fonds/forme », *Revue I3 – Information Interaction Intelligence,* 4 (1), 2004 (http://archivesic.ccsd.cnrs.fr/sic_00001017).

[BAC 0 7] BACHIMONT B., « Nouvelles tendances applicatives : de l'indexation à l'éditorialisation », dans *L'indexation multimédia,* Paris, Hermès, 2007 (http://cours.ebsi.umontreal.ca/sci6116/Ressources_files/ Bachimont FormatHerme %CC%80s.pdf).

[BAR 15] BARDET J.-P., RUGGIU F.-J. (dir.), *Les Écrits du for privé en France : de la fin du Moyen Âge à 1914*, Collection : Orientation et méthodes, Comité des travaux historiques et scientifiques (CTHS), 2015.

[BER 1 2] BERETTA F., VERNUS P., « Le projet SyMoGIH et la modélisation de l'information : une opération scientifique au service de l'histoire », *Les Carnets du LARHRA,* n° 1, 2012, p. 81-107 (http://halshs.archives-ouvertes.fr/halshs-00677658).

[BER 15a] 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 .

[BER 15b] BERETTA F., « 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, France, octobre 2015, https://halshs.archives-ouvertes.fr/halshs-01251915 .

[BER 15c] BERETTA F., « Recherche historique et interopérabilité des données : le projet symogih.org, plateforme collaborative de gestion de l'information historique », *Mégadonnées et interopérabilité dans les Humanités Numériques*, Lille, France, décembre 2015, https://halshs.archives-ouvertes.fr/halshs-01253226 .

[BER 16] BERETTA F., « L'interopérabilité des données historiques et la question du modèle: l'ontologie du projet SyMoGIH », dans J.-L. MINEL (dir.), *Quels enjeux numériques pour les médiations scientifique et culturelle*, à paraître en 2016 aux PUF.

[BRO 13] BROUDOUX C., CHARTRON G., CHAUDIRON S., « L'architecture de l'information : quelle réalité conceptuelle ?» . *Études de communication*, 2013, pp.13-30. http://archivesic.ccsd.fr/sic_00998367

[DAC 10] DACOS M. , MOUNIER P. , *L'édition électronique*, Collection Repères, La Découverte, 2010.

[DUC 04] DUCOURTRIEUX C., « L'édition électronique en quête de définition(s) » . *Le Médiéviste et l'ordinateur*, n°43, 2004. http://lemo.irht.cnrs.fr/43/43-02.htm

[GEN 14] GENÊT P., POIRIER P., « La fonction éditoriale et ses défis » dans E. SINATRA, M. VITALI-ROSATI (éd.) Pratiques de l'édition numérique, Collection « Parcours numériques », Les Presses de l'Université de Montréal, Montréal, 2014, p.15-29.

[JOR 12] JORDANOUS A., STANLEY A., TUPMAN C., « 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*. Balisage Series on Markup Technologies, vol. 8 (2012). http://www.balisage.net/Proceedings/vol8/html/Jordanous01/BalisageVol8-Jordanous01.html

[MAS 03] MASOLO[MAS 03] C., BORGO S., GANGEMI A., GUARINO N., OLTRAMARI A., *WonderWeb Deliverable D18 Ontology Library (final)*, Trento, Laboratory For Applied Ontology, 2003, téléchargeable en version PDF depuis le site http://wonderweb.man.ac.uk/deliverables.shtml .

[MOR 07] MORAND B., « Le logiciel, sujet et objet de la norme ». *Droit et société,* 1/2007 (n°65), p.41-51, www.cairn.info/revue-droit-et-societe-2007-1-page-41.htm

[PED 06] PÉDAUQUE R., T., *Le document à la lumière du numérique,* Caen : C&F Editions, 2006.

[VIT 16] VITALI-ROSATI M., « What is editorialization ? ». *Sens Public*, 2016/1, www.sens-public.org/article1059.html