Encyclopédies médiévales en milieu numérique : les nouveaux enjeux de SourcEncyMe pour le traitement des auctoritates

Isabelle Draelants, Emmanuelle Kuhry

Encyclopédies médiévales en milieu numérique

Les nouveaux enjeux de SourcEncyMe pour le traitement des auctoritates

Isabelle DRAELANTS(1), Emmanuelle KUHRY(2)

(1) (2) CNRS

Fondé sur un patrimoine textuel, SourcEncyMe cherche à rassembler un corpus de textes vérifiés et à en enrichir le contenu scientifique au fil des ans. Le retour d'expérience proposé ici sur ce programme, dédié à l'étude des sources des encyclopédies médiévales, vise à montrer quelles solutions ont été mises en œuvre dans ce cas pour affronter les questions posées plus généralement par les humanités numériques dans le domaine de l'histoire des textes ou des sciences de l'érudition.

Certains constats douloureux sont en effet autant de défis professionnels et sociétaux à surmonter dans des projets mêlant érudition et informatisation des données. Les humanités numériques ne sont plus un domaine récent ; en conséquence, le Web est, aussi, un cimetière à projets arrêtés dans leur lancée. L'obsolescence des outils et technologies mis en œuvre en est une des causes, l'arrêt des financements obtenus par dépôt de projet en est une autre. Ce temps bref des financements et la vitesse des changements technologiques vont à l'encontre de la pérennisation requise par la recherche, et surtout de la nature même de l'érudition, qui nécessite expertise et temps long. L'investissement considérable en travail, et en expérience accumulée, se heurte au vieillissement rapide des outils, et le chercheur en humanités doit sortir de son domaine d'expertise pour devenir à la fois entrepreneur et ouvrier de l'édition électronique. Nombre de chercheurs en sciences de l'érudition ont le sentiment de perdre leur âme, car ils disposent de moins de temps pour approfondir leur discipline, alors qu'il est devenu inimaginable de ne pas recourir aux outils numériques ; sans parler de la transformation du vocabulaire de nos disciplines, qui doivent emprunter au jargon informatique pour rester audibles : les contenus érudits de toute nature perdent leur nom pour être étiquetés en « données » ou « métadonnées ». Dans le même temps, les

exigences ont augmenté : il ne s'agit plus seulement de créer chacun son outil pour mettre en ligne des résultats vérifiés, mais d'utiliser des langages d'encodage (ex. XML) et de transformation (ex. XSLT), des vocabulaires qui sont autant de normes définies par des communautés particulières (ex. TEI) et d'enrichir des référentiels toujours plus nombreux. À cela s'ajoute la pression récente de la « dataisation » (big data) qui incite à privilégier la quantité des données à leur qualité critique. Il faut désormais aussi offrir en open source le codage utilisé et la description des données dans un plan de gestion de données (Data management plan – DMP), et définir des licences attachées au projet. Cette exigence juridique attire à bon escient l'attention sur la question importante des droits et pose le problème du coût des projets de recherche publique livrés en *open source*. Ce statut entraîne en effet le risque d'une exploitation sans convention ni collaboration de tout corpus financé par l'argent public de la recherche française et mis en accès libre sous licence[1]. Enfin, il faut tendre vers une interopérabilité croissante entre les réalisations existantes en humanités numériques.

Après un rappel rapide des objectifs scientifiques du projet, et une description des outils créés initialement, sont présentés ici quelques problèmes spécifiques relatifs au corpus des encyclopédies médiévales latines, c'est-à-dire les questions liées à la structure des œuvres, difficile à rendre par une arborescence XML univoque, à la pseudépigraphie (l'attribution médiévale des citations à des *auctoritates* / « autorités »), à l'identification des citations, et aux annotations sur la tradition textuelle. La deuxième partie de cette contribution s'attache à expliquer la transition et la transformation que nous avons dû opérer sur ces outils initiaux, pour leur permettre de subsister en s'adaptant à l'évolution des humanités numériques. On montre quelles difficultés techniques ont également dû être résolues, pour garantir le passage d'un outil sous forme de bases de données reliées en PhP-MySQL, à une base de données XML native unique (sous BaseX) et le passage d'une interface d'administration collaborative, à un outil de balisage « chercheur-friendly » en XML-TEI.

1 Historique, objectifs, partenaires et collaborations de SourcEncyMe [2]

Le projet SourcEncyMe – *SOURCes des ENCYclopédies MÉdiévales* – a pour objectif d'identifier les innombrables sources de l'encyclopédisme médiéval, de rendre compte de la profonde intertextualité médiévale et de la richesse de l'acte de « compilatio », une notion qui se rapproche davantage de celle de *compilation* en anglais, qui signifie « com-position », que du mot « compilation » en français, qui peut avoir un sens péjoratif.

[1] Le corpus SourcEncyMe a fait l'objet d'une exploitation pour extraction automatique de sources par Cuvelier, de Valeriola et Engelbeen 2020.

[2] L'historique du projet, ainsi que diverses présentations, en français et en anglais sont lisibles sur le « Carnet Hypothèses » de l'Atelier Vincent de Beauvais https://ateliervdb.hypotheses.org/99, consulté le 20 août 2021. Une rubrique du site SourcEncyMe retrace également les principales étapes : http://sourcencyme.irht.cnrs.fr/historique. On trouvera diverses présentations archivées sur le Carnet de recherches de l'Atelier Vincent de Beauvais : https://ateliervdb.hypotheses.org/files/2016/03/Sourcencyme-presentation-courte-mars-2010.pdf et https://ateliervdb.hypotheses.org/files/2020/11/SourCencyMe-A-Tool-to-investigate-the-Sources-of-Medieval-Philosophy-April-2014.pdf. Le Plan de gestion de données (DMP) a été mis en ligne sur « Opidor » : https://dmp.opidor.fr/

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

SourcEncyMe est envisagé comme un outil de référence pour connaître la bibliothèque savante des encyclopédistes médiévaux. Il permet d'observer les techniques de compilation médiévales, au moment où l'effort d'assimilation des connaissances antiques et arabes fut le plus important dans l'histoire occidentale médiévale, au XIII^{e} siècle. SourcEncyMe rassemble les encyclopédies médiévales latines et a pour but d'étudier leurs sources, leurs *auctoritates*, c'est-à-dire tous les auteurs et les œuvres utilisées par les encyclopédistes, en particulier au « siècle d'or » des encyclopédies médiévales. Tissées de 75 à 95 % de citations de textes contemporains ou parfois antérieurs de plusieurs siècles, ces encyclopédies représentent un objet emblématique de l'intertextualité des œuvres médiévales. L'enjeu de leur étude est de donner accès à un large héritage littéraire, scientifique, philosophique, théologique, en dévoilant la stratigraphie des informations textuelles accumulées. Le projet met plus particulièrement l'accent sur la philosophie naturelle, c'est-à-dire sur la science de la nature, aussi appelée à l'époque « physique ».

La citation est une constante dans le monde savant médiéval : les auteurs médiévaux renvoyaient à leurs aînés, comme nous avons l'obligation de la référence à la bibliographie utile. A l'époque, la balance penchait cependant du côté du respect de la parole et du texte que de la critique de la pensée. La célèbre phrase attribuée à Bernard de Chartres par Jean de Salisbury et représentée sur un vitrail de la cathédrale de Chartres est emblématique de cet *habitus* de l'intellectuel médiéval : « nous sommes comme des nains portés sur les épaules de géants. Nous voyons ainsi plus de choses que les anciens et de plus éloignées, non par la pénétration de notre propre vue ou parce que notre taille est plus haute, mais parce qu'ils nous soulèvent et nous haussent de toute leur hauteur gigantesque. » (Bernard de Chartres, *via* Jean de Salisbury, *Metalogicon*, III, 4, éd. Hall et Keats-Rohan 1991). Dans cet esprit, l'objectif est de faire de SourcEncyMe un outil pour l'érudition qui vise à la fois à donner accès à un corpus d'encyclopédies médiévales qui avaient elles-mêmes pour but de totaliser le savoir, tout en identifiant les sources de ce savoir. Ce projet d'envergure nécessite d'avancer par étapes pour créer et accroître la taille du corpus et mettre en œuvre une érudition spécialisée.

Les bases de données qui ont formé le corpus, ainsi que la plateforme collaborative de travail permettant d'enrichir les données scientifiques, ont été créés en 2010 au sein de l'*Atelier Vincent de Beauvais* à Nancy[3], avec l'aide du laboratoire CNRS ATILF. Le site SourcEncyMe.irht.cnrs.fr a été ouvert au public le 24 février 2016 à l'Institut de recherche et d'histoire des textes (IRHT, CNRS). Depuis deux ans, grâce

[3]Le petit *Atelier Vincent de Beauvais* se consacrait depuis la fin des années 1970 à l'étude de la copieuse encyclopédie de Vincent de Beauvais (*c.* 3 millions de mots). M. Paulmier-Foucart et M.-Chr. Duchenne avaient commencé à transcrire le texte du *Speculum maius* à partir des manuscrits dans les années 1980 et à le mettre en ligne avec l'aide de l'ATILF (laboratoire CNRS spécialisé dans le traitement automatique de la langue française). À l'arrivée d'I. Draelants en 2002 comme chargée de recherche CNRS, l'*Atelier* a élargi ses travaux à l'étude des encyclopédies naturalistes médiévales en général. Fin 2013, début 2014, sa mutation à l'IRHT a entraîné le transfert des bases de données et des activités de l'*Atelier Vincent de Beauvais,* ainsi que du projet SourcEncyMe, à l'IRHT. Cette phase de transition entre deux laboratoires, et ce transfert matériel et technique, a représenté un coup d'arrêt, car cela s'est traduit par un manque de financement et un manque de personnel. L'ouverture au public du site sourcencyme.irht.cnrs.fr a eu lieu après diverses adaptations techniques réalisées avec l'aide du pôle numérique de l'IRHT (Orléans) et quelques mois de travail financés par l'Equipex BIBLISSIMA. Depuis septembre 2019, l'IRHT a déménagé sur le Campus Condorcet au Nord de Paris (Plaine Saint-Denis).

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

au financement du LabEx HASTEC[4], SourcEncyMe connaît une nouvelle mutation technique en vue du passage d'un ensemble de BDD, techniquement obsolète, jumelé à une plateforme (interface) d'administration collaborative, à un corpus intégré en XML-TEI avec une seule BDD en XML reliée à la création d'une interface de balisage. Celle-ci est destinée à remplacer, auprès des contributeurs au projet, la plateforme collaborative.

SourcEncyMe doit aujourd'hui conjoindre croissance des contenus spécialisés et évolution des outils techniques. Après une dizaine d'années d'existence[5], l'outillage nécessite une complète rénovation, en cours, et des moyens humains pour introduire des contenus de la qualité attendue par les publications d'érudition traditionnelles.

Il faut souligner que l'objectif de SourcEncyMe n'est pas d'offrir un corpus, mais de permettre des identifications sur les sources des encyclopédies médiévales qui forment ce corpus. Le corpus SourcEncyMe avait pour contenus principaux en 2016 le *Speculum maius* de Vincent de Beauvais, à savoir les *Speculum historiale, naturale* et *doctrinale* (achevés vers 1260) transcrits à partir d'un manuscrit (pour l'*historiale*) et de l'édition de Douai de 1624 (pour les *Specula naturale* et *doctrinale*) ; les diverses versions des prologues des œuvres de Vincent de Beauvais transcrites à partir des manuscrits ; l'*Historia naturalis* de Juan Gil de Zamora (*c.* 1280) d'après l'édition critique de García et Ballester 1994 ; et l'édition critique du livre IV sur les poissons de l'*Hortus sanitatis*, une encyclopédie pharmacologique du XV^e^ siècle (Jacquemard, Gauvin et Lucas-Avenel, 2013), grâce à la collaboration avec le projet Ichtya des latinistes du CERLAM à Caen[6]. Aujourd'hui, le partenariat continue avec le pôle numérique de la MRSH de Caen pour un partage de technologie, notamment en ce qui concerne l'outil d'indexation et le plug-in PluCo.

Entre 2008 et 2017 se sont succédé à l'*Atelier Vincent de Beauvais* à Nancy, au *Centre de médiévistique Jean-Schneider*, puis à l'IRHT à Paris des doctorants et post-doctorants dont les résultats de recherche ont contribué au programme ou seront intégrés dans le corpus : Eduard Frunzeanu (Montréal) avec un projet sur le *Speculum naturale* et une contribution à l'édition du livre VIII, *De mundo,* du *De proprietatibus rerum* de Barthélemy l'Anglais[7] ; Iolanda Ventura (Florence) avec une recherche sur la botanique encyclopédique et la coordination de l'édition critique du *De proprietatibus rerum* de Barthélemy l'Anglais (voir Ventura, 2007) ; Riccardo Saccenti (Bologne) a préparé l'étude des sources et prévoit l'édition des livres I et VIII sur la morale du *Compendium philosophiae* pour leur intégration au corpus ; Sébastien

[4]HASTEC : Laboratoire d'Excellence *Histoire et anthropologie des savoirs, des techniques et des croyances.* (https://labexhastec-psl.ephe.fr/). Le projet SourcEncyMe s'inscrit en particulier dans les axes 4 « Doctrines et techniques intellectuelles et spirituelles : philosophie, science et religion », et 6 « Technologies numériques et transformations des connaissances ».

[5]Le projet a noué diverses collaborations et a bénéficié de financements successifs depuis celui de l'Agence nationale de la recherche en 2008 (ANR) jusqu'au soutien actuel du Laboratoire d'Excellence HASTEC.

[6]Projet Ichtya : https://www.craham.cnrs.fr/projet/programmes/ichtya/. Le CERLAM formé de philologues a été intégré ensuite dans le CRAHAM. Au pôle « Document numérique », créé par Catherine Jacquemard, avec le concours essentiel de Pierre-Y. Buard (ingénieur), un projet sur la littérature ichtyologique latine a démarré peu après SourcEncyMe, avec pour corpus de départ le livre sur les poissons de l'*Hortus sanitatis*. Le livre ichtyologique de cette compilation pharmacologique du XV^e^ siècle exploitant largement le *Speculum maius* de Vincent de Beauvais a été intégré dans le corpus SourcEncyMe.

[7]Voir *Draelants, Isabelle et Eduard Frunzeanu, avec la collab. de Iolanda Ventura. À paraître en 2022.*

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

Moureau (UCL Louvain) a étudié les sources alchimiques du *Speculum maius* ; Tomas Zahora (Fordham, E.U. – Univ. Monash, Australie) a travaillé sur le *Speculum morale* apocryphe mis sous le nom de Vincent de Beauvais, qui sera intégré au corpus, et sur l'automatisation de l'identification des sources ; Mattia Cipriani (Florence) a préparé une nouvelle édition critique de la version auctoriale du *Liber de natura rerum* de Thomas de Cantimpré dont la version préparatoire est en ligne. Les doctorants qui ont participé ou participent au projet sont Emmanuelle Kuhry (Nancy), qui mène l'édition critique des livres de philosophie naturelle du *Compendium philosophiae* et des recherches sur l'utilisation encyclopédique de la *Glossa anglicana* ; Irene Villarroel Fernández (Univ. Complutense, Madrid) avec une thèse sur le florilège moral tiré des livres IV et V du *Speculum doctrinale* de Vincent de Beauvais ; Beatrice Amelotti (Università degli Studi, Pavie) avec une thèse sur le livre IX de l'encyclopédie de Giovanni da San Giminiano ; Elisa Lonati (EPHE, Scuola Normale de Pise) qui termine une thèse sur le *Chronicon* d'Hélinand de Froidmont, et Ombeline Fichant (EPHE) qui étudie l'*Opusculum de naturis animalium* (c. 1210) inédit dont le texte sera mis en ligne.

2 Les outils et les données du site avant la mutation technique

Pour comprendre l'expérience spécifique de SourcEncyMe, il importe de décrire rapidement les outils et les données tels qu'ils existaient avant la mutation en cours.

Le XIII^{e} siècle est emblématique du désir médiéval de bâtir sur l'héritage textuel du passé. A cette époque, la quantité d'informations textuelles a considérablement augmenté, grâce aux traductions du XII^{e} siècle, en particulier dans le domaine de l'histoire naturelle (Aristote, Avicenne et leurs commentateurs, médecins et astronomes antiques et arabes, etc.). Dans un contexte historique et culturel qui voit se multiplier les outils de travail, florilèges, encyclopédies et tables des matières, l'acte de « compilation » apparaît comme une noble entreprise de reconstruction du savoir à partir de l'ensemble des briques anciennes, auxquelles les auteurs-compilateurs ajoutent la science de leur temps et le ciment de leur propre assemblage.

Pour une utilisation aisée par le lecteur médiéval, les informations proposées par les encyclopédistes de cette époque se présentent soit sous forme de citations, parfois abrégées, classées par chapitres thématiques, soit dans des catalogues alphabétiques (en particulier pour les *naturalia* comme les plantes ou les animaux). Ces citations sont généralement référencées par un « marqueur de source » que les savants médiévaux. Il s'agit du nom de l'« autorité », que les savants médiévaux présentent souvent sous la forme des noms de l'auteur et de l'œuvre cités : par exemple, *Aristoteles in libro metheororum, Constantinus in Pantegni, in libro Canonis Avicenne.* Ceci nous confronte immédiatement au fossé qui peut exister entre une référence médiévale et son identification actuelle. Ces raisons ont justifié le choix d'adopter, pour SourcEncyMe, une structure fondée sur ce découpage en « unités de citations » précédées par un « marqueur de source ». La structure du site de consultation est donc organisée en arborescence, depuis le découpage en livres et en chapitres, puis en subdivisions adoptées par le compilateur médiéval, jusqu'à « l'unité de base » que constitue la citation

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

précédée de son « marqueur ». Cette dernière peut elle-même être divisée en « segments de citation » au cours de son identification par le chercheur (voir ci-dessous).

SourcEncyMe rassemble les éléments suivants :

1.) un corpus annoté de textes encyclopédiques latins balisés en XML-TEI (c. 5 millions de mots : 4.400.286 mots en comptant les citations seules, 4.739.474 avec les annotations et identifications posées, décompte au 9 juin 2020). On y trouve, à cette date, les textes suivants :

- *Speculum maius* de Vincent de Beauvais : *naturale, doctrinale, historiale* ;
- Prologue au *Speculum maius,* en quatre versions ;
- Prologues aux « petites œuvres » de Vincent de Beauvais ;
- *Liber de natura rerum* de Thomas de Cantimpré, en deux versions, auctoriale en 20 livres et remaniée en 13 livres (appelée « Thomas III. »)
- *Historia naturalis* de Juan Gil de Zamora
- *Hortus sanitatis,* livre III
- *Tractatus de naturis animalium* d'Engelbert d'Admont
- *Compilatio de libris naturalibus Aristotelis et aliorum quorundam philosophorum* (*Compendium philosophiae*), Livre III

2.) un ensemble de méta-données critiques :

a) des mémentos, c'est-à-dire des fiches bio-bibliographiques relatives aux œuvres et auteurs-sources (*auctoritates*), cités implicitement ou nommément ;

b) pour chacune des autorités citées, nous avons créé un « nom canonique », c'est-à-dire un nom latin standardisé unique, pour les œuvres comme pour les auteurs, pour éviter par exemple de confondre deux vies de saints aux noms similaires, ou deux œuvres du même nom écrites par deux auteurs différents (comme le *De anima* d'Augustin et celui d'Aristote ou celui d'Avicenne, ou le *De naturis rerum* d'Alexandre Nequam, à ne pas confondre avec celui d'Isidore ou de Bède le Vénérable, ou encore avec le *Liber de natura rerum* de Thomas de Cantimpré) ;

c) des identifications des segments de citations ;

d) des annotations sur les intermédiaires de transmission (versions du texte, traductions, sources intermédiaires du compilateur, etc.) ;

e) une interface collaborative de travail en ligne, en cours de remplacement par une interface de balisage (à l'usage des collaborateurs). Elle permettait jusqu'ici de compléter les mémentos, d'intégrer les identifications de sources, d'annoter par des commentaires sur la tradition textuelle.

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

C'est sur ce dernier point que la modification technique en cours est la plus considérable, comme le montrent l'explication et les schémas ci-dessous. Il importe auparavant de donner quelques mots d'explication sur le site actuel sourcencyme.irht.cnrs.fr, les encyclopédies consultables et en cours de traitement, celles dont l'intégration est prévue, et les modes de consultation et de recherche.

Le corpus continue de rassembler de nouvelles encyclopédies ; nous avons ainsi ajouté en juin 2021 le *Tractatus de naturis animalium* d'Engelbert d'Admont dans l'édition critique de Max Schmitz (2007) et l'édition critique ainsi que l'identification (à 90 %) des sources du livre III du *De plantis* du *Compendium philosophiae* (*Compilatio de libris naturalibus Aristotelis et aliorum philosophorum*), par E. Kuhry. Le texte de la *Philosophia* de Daniel de Morley (*Liber de superioribus et inferioribus)* de l'édition de Gregor Maurach de 1971 est en cours de traitement et sera intégré en 2021, ainsi que le *De natura rerum* d'Alexandre Nequam dans l'édition de Thomas Wright de 1863. Une collaboration avec l'Université de Monash en Australie nous permet de disposer du texte, doté d'un balisage élémentaire, du *Speculum morale,* qui constituait la quatrième partie, apocryphe, du *Speculum maius* dans l'édition de Douai de 1624. Nous prévoyons d'intégrer progressivement les textes suivants : le *Pro conservanda sanitate* de Vital de Four, du début du XIV^{e} siècle, dans l'édition de 1531 ; le *Proemium* et le livre musical du *Liber quattuor distinctionum,* intégré dans la première partie du *Liber introductorius* de Michel Scot (avant 1230), dans l'édition de Christian Meyer, 2009 ; le *Liber de moralitatibus* de Marcus d'Orvieto, dans l'édition de Gerhard Etzkorn de 2005.

D'autres éditions critiques sont en cours, elles viendront prendre place dans le corpus. C'est le cas de l'édition de la version « vulgate », c'est-à-dire celle qui a circulé au XIII^{e} siècle, du *Liber de natura rerum* de Thomas de Cantimpré, par Mattia Cipriani, qui devrait être achevée en 2022 ; de l'édition du *De mundo,* à savoir le livre VIII du *De proprietatibus rerum* de Barthélemy l'Anglais (Draelants, Frunzeanu, 2022). En fonction des droits que laissera la maison d'édition Brepols, les livres I-IV (Long et Roling, 2007) et XVII (Ventura, 2007) pourront aussi être mis en ligne. Le huitième livre, sur la morale, du *Compendium philosophiae*, en cours d'édition par Riccardo Saccenti, est également promis à la diffusion *via* SourcEncyMe. Nous espérons également pouvoir donner accès au *De floribus rerum naturalium* d'Arnold de Saxe (I. Draelants) et aux livres VIII-XII (en cours d'édition par E. Lonati, doctorante) et XLV-XLIX du *Chronicon* d'Hélinand de Froidmont, ainsi qu'à l'*Opusculum de naturis animalium* (c. 1210 ?) qu'étudie Ombeline Fichant, doctorante. L'accès à des transcriptions de divers recueils de propriétés qu'Emmanuelle Kuhry a mis au jour et dont elle a montré l'importance comme précurseurs des encyclopédies et pour leur processus de rédaction[8], apporte aussi une plus-value considérable et inédite que pourra offrir le projet SourcEncyMe. Ainsi, les recueils appelés d'après leurs incipits *Angelus purus natura,* et *Abies arbor alta,* ainsi que « l'Anonyme de Clairvaux » pourront être mis en ligne.

Le site du projet donne accès à la fois aux éditions ainsi enregistrées et à l'identification progressive de leurs sources. Il permet toutes sortes de requêtes sur les textes et

[8] Voir Kuhry 2018 [avec une liste des témoins pour *Angelus purus natura*].

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

leurs *auctoritates*[9]. Outre l'historique du projet et les mentions légales et de propriété, le menu offre un *vademecum* important, car il constitue le mode d'emploi destiné aux utilisateurs et aux collaborateurs du projet pour les informer de la *forma mentis* dans laquelle l'organisation et le traitement des données textuelles ont été conçus. La « connexion » permettait jusqu'ici aux collaborateurs d'accéder à la plateforme collaborative pour ajouter identifications et annotations (voir ci-dessous la section sur l'ancienne plateforme collaborative), mais cette option est appelée à disparaître au profit d'un encodage direct en XML-TEI dans un outil de balisage. La page d'accueil permet comme dans n'importe quel moteur de recherche une recherche « brute » sur tout mot latin, avec la possibilité d'utiliser les opérateurs booléens. Les troncatures sont possibles en utilisant l'astérisque. Les réponses portent sur les occurrences dans l'espace d'une citation ; elles peuvent être triées d'après les encyclopédies enregistrées (dans la colonne à gauche). Elles indiquent l'œuvre et la place de l'unité de citation dans l'œuvre, en termes de livre, de chapitre, de subdivision éventuelles, ainsi que le marqueur de citation (référence médiévale) quand il existe. La même recherche peut être menée *via* la commande « Accès aux textes », ou « Recherche corpus ». Pour une consultation plus poussée, l'« accès aux textes », une fois sélectionné l'onglet « consulter », permet de choisir une encyclopédie en particulier. Un « clic » sur les réponses donne accès, sur la gauche de l'écran, à une vue de l'encyclopédie consultée, sous forme d'arborescence conçue pour être déroulée (figure 1 ci-dessous, pour le *Liber de natura rerum* de Thomas de Cantimpré) : en cliquant sur les points de branchement (triangles), on peut entrer dans la structure de l'œuvre, du livre au chapitre, du chapitre à la subdivision, et de la subdivision à l'unité de citation, indexée par son marqueur de référence sur lequel il suffit ensuite de cliquer.

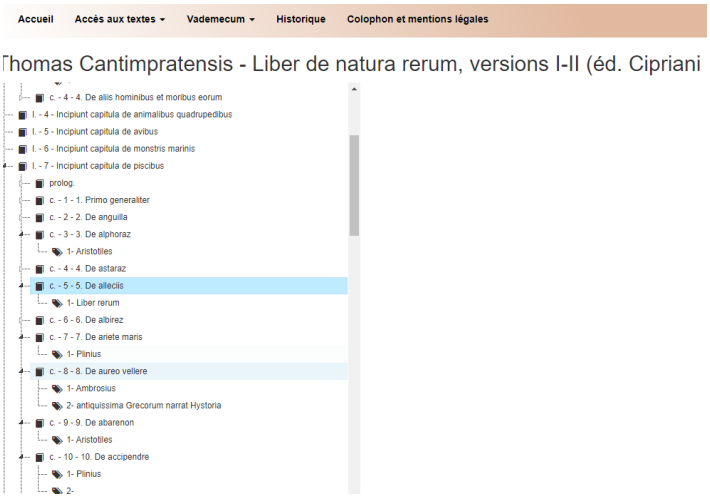

Fig. 1 : Arborescence des textes du corpus[10]

[9] http://sourcencyme.irht.cnrs.fr/

[10] Toutes les images de cette contribution sont produites par les autrices.

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

La recherche de sources en particulier permet de savoir si ces œuvres sont représentées dans le corpus encyclopédique enregistré. Par exemple (figure 2 ci-dessous), une recherche sur « De anima » permet de lister les sources portant ce nom et utilisées dans les différentes encyclopédies. Il est ensuite possible de choisir celle qu'on désire parmi les noms canoniques présentés. En cliquant sur « mémento œuvre », on accède à un titre qui donne l'attribution à son auteur réel et aux « auteurs allégués » auxquels l'œuvre a pu être attribuée par les encyclopédistes. L'accès à la fiche-mémento (fiche bio-bibliographique) est possible par ce biais. Une des possibilités envisagées désormais est, plutôt que de compléter les mémentos par des informations bio-bibliographiques qui sont susceptibles d'être accessibles assez aisément ailleurs, de limiter le contenu du mémento à ce qui est strictement nécessaire pour la navigation entre les œuvres et auteurs-sources (nom canonique, noms attestés, attribution, date, édition de référence ou manuscrit choisi comme référence…), et de renvoyer vers des référentiels existants, à partir du nom canonique de l'auteur réel, supposé ou pseudépigraphe. En effet, ces dernières années, ces divers référentiels d'autorités se sont fortement développés et de plus en plus d'identifiants uniques ou pérennes existent sur le Web. Un partenariat avec l'Equipex BIBLISSIMA + offre de réelles perspectives à cet égard[11].

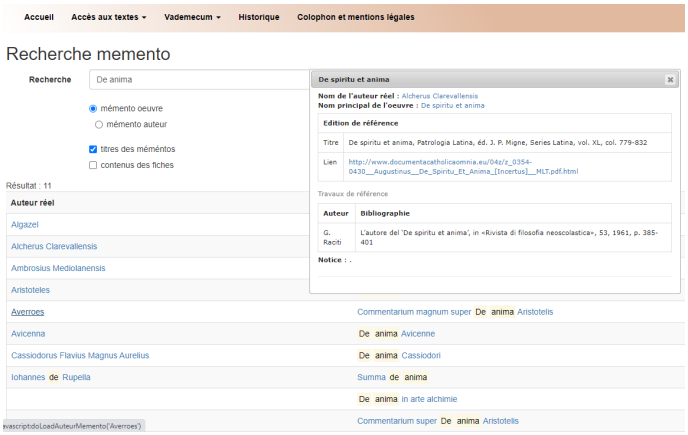

Fig. 2 : La recherche d'une fiche mémento[12]

Le choix d'une « recherche avancée du corpus » permet de croiser les interrogations. Par exemple, saisir *Aristoteles* comme auteur d'une source, c'est-à-dire comme *auctoritas*, permet de choisir une œuvre d'Aristote en particulier, comme le *De animalibus*, pour y faire une recherche de termes précis. Dans l'exemple ci-dessous (figure 3), la requête porte sur toutes les formes de « **animal* » (avec troncature) ou encore de « *prudenti** », dans l'idée de rechercher les occurrences qui pourraient traiter de la compétence animale.

[11] C'est-à-dire la deuxième phase de financement sous forme d'Equipex de la *Bibliotheca bibliothecarum* : https://projet.biblissima.fr/fr/actualites/biblissima-observatoire-cultures-ecrites-argile-a-imprime

[12] La figure montre les champs du mémento sur le *De spiritu et anima* attribué à Alcher de Clairvaux, avec le lien vers l'édition de référence utilisée pour l'identification dans le corpus, et un lien vers de la bibliographie de base, ainsi qu'une note en texte libre, qui peut attirer l'attention sur des particularités historiographiques.

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

Fig. 3 : La recherche avancée du corpus

3 Problèmes de critique textuelle et solutions techniques

On aborde dans cette partie certains problèmes de critique textuelle posés par les encyclopédies médiévales en cœur de cible du projet SourcEncyMe, en particulier les questions de hiérarchie des autorités, de versions multiples d'une même œuvre, et de pseudépigraphie.

Le problème de la hiérarchie des autorités (*auctoritates*) requiert de savoir quelle est la source réellement citée par l'encyclopédiste (ou le texte intermédiaire qui a fourni la citation), par rapport à celle qui est annoncée par le « marqueur de citation » médiéval, quelles sont les éventuelles sources secondaires de la citation (qu'elles soient mentionnées ou non) et quel est l'ordre dans lequel il faut mettre en abîme les divers jalons entre le passage cité et l'encyclopédiste qui le transmet. L'objectif est donc de rétablir la stratigraphie des autorités. Par exemple, le *Tractatus de naturis animalium* d'Engelbert d'Admont, rédigé vers 1300, s'appuie essentiellement sur les anciennes autorités encyclopédiques des *Étymologies* d'Isidore de Séville (VII^{e} siècle) et des *Collectanea rerum memorabilium* de Solin, qui s'inspire lui-même d'Isidore et de Pline l'Ancien (III^{e} siècle). Cependant, Engelbert y ajoute des auteurs classiques et patristiques comme Virgile et Ovide, des extraits de la *Zoologie* d'Aristote, ou des extraits d'auteurs énigmatiques comme l'*Experimentator* (qui n'est pas l'encyclopédie utilisée par Thomas de Cantimpré et éditée par Janine Deus, même s'il y a des points communs avec ce texte) et fait quelques commentaires de son cru pour organiser l'information. Sous un même marqueur de source « *Solinus* », nous pouvons trouver deux citations de Solin abrégées, comme cousues par un commentaire d'Engelbert pour former la trame du texte. Un commentaire personnel de l'auteur encyclopédique suit souvent une citation d'*auctoritas* ; seule une identification précise de la source permettra de découper l'unité de citation en « segments de citation ». Le travail d'identification révèle ainsi toute la complexité de l'intertextualité. Engelbert peut donner pour marqueur d'autorité « Démocrite », alors que la citation provient de Solin (y compris l'allusion à Démocrite) et que Solin a emprunté le texte des mots de Démocrite à Pline. Il arrive ainsi que ce qui suit la citation dans le texte d'Engelbert ne soit plus une citation, mais un résumé de la pensée d'Aristote dans le *De animalibus* traduit par Michel Scot de l'arabe ou par Guillaume de Moerbeke à partir du grec ; l'une et l'autre des traductions étaient déjà disponibles vers 1300, et celle de Moerbeke, bien que meilleure, n'a pas beaucoup circulé. Ainsi, dans la citation suivante :

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

> Sicut Democritus physicus dicit, sola mulier est animal menstruale. Hoc autem secundum Aristotelem ita intelligendum est, quod femine aliorum menstruant etiam, sed non ita sepe neque ma<n>ifeste »

On peut reconnaître Pline, *Naturalis historia,* VII, 13 : *Solum autem animal menstruale mulier est*, cité en réalité par Solin, *Collectanea rerum memorabilium,* c. 1 : *Itaque, ut Democritus physicus ostendit, mulier solum animal menstruale est.* Le passage est suivi par un petit extrait du *De animalibus* d'Aristote, sans que la longueur de l'extrait permette de déterminer la traduction médiévale qui a été utilisée.

Le lien entre le marqueur d'autorité et le « nom canonique » donne à son tour lieu à des problèmes complexes lors du découpage en unités de citations. On se confronte aussi à l'absence de marqueur (plusieurs citations d'auteurs divers sans nouveau marqueur), à des marqueurs multiples (p. ex., Aristote et Ambroise de Milan cités ensemble parce qu'ils livrent une information similaire à propos d'un même animal ; Dioscoride, Evax et Aaron mentionnés ensemble à propos d'une collection minéralogique trouvée sous ces trois noms dans un manuscrit consulté par l'encyclopédiste[13]), à des marqueurs collectifs (p. ex. *philosophi,* ou *physiologi,* ou *experimentatores*), à des marqueurs spécifiques à un compilateur pour une traduction, une œuvre ou une partie d'œuvre non identifiée par la critique (p. ex. *Liber Veneris, Liber eternorum, Esculapius de membris…*), à des marqueurs vagues (*quidam dicunt, poeta*[14]*, propheta*[15]*, apostolus,* etc.), et – en lien avec ce qui vient d'être exposé – à des marqueurs emboîtés. Un exemple fréquent de ce dernier cas se rencontre par exemple chez Vincent de Beauvais, qui cite Thomas de Cantimpré, qui lui-même cite Isidore qui rapporte les dires de Pline. Il se peut que le marqueur de source indiqué par Vincent de Beauvais soit « *Liber de natura rerum* », ou parfois Isidore ou Pline… Selon le cas, on identifiera la source en rapport avec le marqueur de citation indiqué par l'encyclopédiste, mais une annotation permettra d'indiquer quel est l'intermédiaire qui a fourni la citation. D'où l'importance, d'une part, de l'identification correcte du passage en lien avec le marqueur de source, et d'autre part, de l'apparat d'annotations, qui permet d'indiquer quel est l'intermédiaire qui a fourni à l'encyclopédiste la citation.

Le cas de versions multiples d'une même œuvre, citées dans une même encyclopédie sous un même nom ou des noms divers, se rencontre en particulier lors de traductions médiévales multiples d'une même œuvre, du grec au latin ou de l'arabe au latin. Pour chaque version de l'œuvre, il faut créer une identité différente, reliée à un même mémento relatif à l'œuvre originale (telle qu'elle a été créée par l'auteur). Par exemple, une référence médiévale identique ou très semblable peut renvoyer à trois versions différentes du *De celo et mundo* d'Aristote et à une paraphrase d'Avicenne. Il existait en effet, entre 1175 et 1275, quatre traductions différentes de cette œuvre ; en conséquence, une référence médiévale telle que « *Aristoteles in libro de celo et mun-*

[13] Voir ci-dessous à propos de cette source minéralogique.

[14] Par exemple, sous « *poeta* », Engelbert d'Admont peut citer une autorité aussi rare que l'*Epistola ad Fedolum*, poème retrouvé dans un manuscrit de Saint-Gall par l'érudit Goldastus, dont l'édition de 1606 (*Alemannicorum rerum scriptores*) a été reprise par J.-P. Migne.

[15] *Propheta* peut désigner un livre biblique en particulier mis sous le nom d'un prophète. Dans SourcEncyMe, nous avons distingué les livres bibliques comme des œuvres distinctes pour l'identification, soit sous le nom du livre biblique, ou sous le nom du prophète ou de l'auteur auquel il est attribué.

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

do » peut renvoyer au XIII^e^ siècle soit : 1. au *Liber celi et mundi* mis sous le nom d'Avicenne et souvent inclus dans son *De naturalibus,* c'est-à-dire sa *Physique.* Ce résumé-paraphrase arabe du *De coelo* d'Aristote fut probablement effectué dans le milieu de Hunayn ibn Ishâq (le Johannitius des Latins). Cette œuvre est utilisée par Daniel de Morley dans sa *Philosophia* écrite entre 1175 et 1200 ; quant à Arnold de Saxe, il la nomme *Liber celi et mundi secundum veterem translationem* dans son *De floribus rerum naturalium* (vers 1230-1240)[16] ; 2. au *De celo et mundo* d'Aristote traduit de l'arabe vers 1143 par l'Italien Gérard de Crémone d'après Ibn al-Bitrîq (*Liber de celo et mundo secundum novam translationem* chez Arnold de Saxe), une traduction connue comme la *translatio vetus*[17] ; 3. *De celo et mundo* d'Aristote accompagné du commentaire d'Averroès, traduit de l'arabe par Michel Scot et dédié à Étienne de Provins en 1231, une version du texte qu'utilise Barthélemy l'Anglais dans son *De proprietatibus rerum* achevé vers 1247 ; 4. au *De celo et mundo* traduit du grec peu après 1230 par Robert Grosseteste ; 5. à la nouvelle traduction gréco-latine par Guillaume de Moerbeke réalisée entre 1260 et 1270, et appelée par la critique plus récente « *translatio nova* » ; 6. sous la référence *De celo et mundo* peut aussi se cacher le *De processione mundi* ou *De creatione celi et mundi,* appelé aussi *Liber de prima forma et materia* de Dominicus Gundissalinus (Domingo Gonçalvez), un traité théologique et cosmologique écrit par un savant hispanique qui était lui-même traducteur de diverses œuvres philosophiques de l'arabe au latin au milieu du XII^e^ siècle (Édition Fidora, Soto Bruna et Alonso del Real 1999).

Le problème de la pseudépigraphie, c'est-à-dire les cas où une source porte un « marqueur » d'autorité faux aux yeux de la critique historique actuelle, est particulièrement complexe et crucial dans le corpus encyclopédique médiéval. La référence médiévale indique dans ces cas une attribution à une *auctoritas* différente de celle de l'auteur réel ; il faut donc, à l'intérieur du corpus, et dans les mémentos, faire le lien entre les marqueurs médiévaux de source et leur identification moderne. Le problème s'accroît quand un même nom médiéval d'*auctoritas* couvre un amalgame de plusieurs œuvres ou auteurs, ou des collections de textes dont l'histoire de la transmission diffère. Un des cas les plus complexes est celui de la matière minéralogique mise sous le nom du médecin grec du I^er^ siècle Dioscoride. Selon leur méthode de travail ou les bibliothèques qu'ils ont eues à leur disposition, les encyclopédistes du XIII^e^ siècle citent toutes sortes de versions différentes de son ouvrage pharmacologique *De materia medica,* écrit initialement en grec mais dont plusieurs versions latines ont circulé et ont été abrégées ou mêlées à d'autres œuvres sur les plantes ou les pierres. Le livre V était consacré à la minéralogie et aux remèdes tirés de ces matières (pierres et métaux). Selon les cas, les citations peuvent provenir, par exemple, du « Dioscoride lombard » traduit au IV^e^ siècle, ou d'une autre version latine du VI^e^ siècle, en passant par la révision mise sous le nom d'Evax dans la 2^e^ moitié du XI^e^ siècle et mise en ordre alphabétique à l'époque de Constantin l'Africain. Ces diverses versions se trouvent le plus souvent citées sous un même marqueur « Diascorides ». Le contenu révisé du livre V minéralogique de Dioscoride a été mis en conséquence sous les noms d'autres auteurs de lapidaires, comme Damigéron (qui a écrit en grec) et Evax, dont le traité

[16] Cf. Gutmann 1997.

[17] Ed. par I. Opelt dans l'édition critique du Commentaire sur le *De caelo* d'Albert le Grand par Hossfeld 1971. Sur la tradition arabe, voir Endress 1966.

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

des pierres antique, en version latine, a été lui-même conjoint au lapidaire en vers de Marbode de Rennes au XI^e siècle (ce dernier porte souvent en prologue la lettre au « roi Evax »). Au XI^e siècle, le texte très abrégé est repris dans une compilation alphabétique probablement exécutée dans le Sud de l'Italie qui sera très diffusée du XII^e au XIV^e siècle. À la version alphabétique de Dioscoride sont aussi ajoutés des matériaux arabes, dont le *Liber de gradibus* adapté par Constantin l'Africain au XI^e siècle et le *De physicis ligaturis* traduit de l'œuvre de Qustâ ibn Luqâ, ainsi que le lapidaire de Damigéron, et des extraits de la *Practica* du *Pantegni* de Constantin et peut-être du *De dietis* d'Isaac Israeli. Des manuscrits présentent des passages de ce lapidaire composite dans le même ensemble que le lapidaire en vers de Marbode de Rennes (*Liber lapidum*), qui se trouve parfois sous le marqueur « Dioscorides ». En outre, l'herbier *De herbis femininis* s'est mêlé au texte de Dioscoride dès la fin de l'Antiquité, et ce pseudépigraphe est cité, notamment par Vincent de Beauvais, sous le nom d'*Herbarium* car il est aussi mêlé à la tradition de l'Herbier du Pseudo-Apulée.

Ce bref aperçu d'une histoire des textes donne une idée de la richesse de l'héritage intellectuel auquel les encyclopédistes médiévaux donnent accès, et met en lumière la nécessité d'une critique d'attribution rigoureuse qui est l'essence du projet SourcEncyMe. Pour chacune des œuvres « réelles » en cause, il nous faut créer une fiche-mémento, et mentionner dans chacune des fiches toutes les formes de titres attestées dans les encyclopédies ; cela permet, lors de l'identification, de faire le lien vers le mémento de l'auteur réel. Des mémentos seront assignés à chacun des noms canoniques des auteurs allégués, et d'autres mémentos créés pour chacun des noms canoniques des auteurs réels. Pour chaque mémento-auteur, on indiquera tous les noms des œuvres qui lui sont attribuées ; pour chaque mémento-œuvre, on indiquera le nom de son auteur réel ainsi que des auteurs pseudépigraphes sous lequel ou lesquels l'œuvre circule.

Si ces quelques exemples ont permis de rappeler les objectifs scientifiques poursuivis par SourcEncyMe, il convient maintenant d'examiner de quelles façons le support technique et informatique du projet a dû évoluer.

4 Le corpus SourcEncyMe : aspects techniques

4.1 Corpus, métadonnées, structure XML-TEI : spécificités des textes encyclopédiques et du projet SourcEncyMe

À l'origine, le projet portait sur l'encyclopédie tripartite de Vincent de Beauvais, le *Speculum Maius*, initialement géré dans une base de données Filemaker, dans lequel le texte avait été découpé en citations. Au début du projet, les partenaires ont mené une réflexion sur la structure XML-TEI à adopter, ce qui a abouti à l'établissement d'une première DTD (*Document Type Definition*) par les collaborateurs de l'ATILF. Le langage XML-TEI paraissait le plus à même de rendre compte de la complexité de la structure du corpus et des relations d'intertextualité entre les œuvres.

La particularité du projet SourcEncyMe est de proposer une triple couche de métadonnées sur ce corpus :

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

1.) les identifications des sources des citations, qui consistent en des notes posées sur des segments de citation. Ces identifications sont entrées en mode « stand-off embarqué »[18], c'est-à-dire dans le fichier, non pas au sein de la ligne, mais au moyen d'un système de pointeurs qui renvoient à un identifiant unique sous la forme de l'attribut `@target` sur l'élément `note`, qui contient l'identification. La valeur de cet attribut `@target` renvoie vers l'attribut `@xml:id` de l'élément `milestone`. Cet attribut `@xml:id` est un identifiant unique posé sur le premier élément `milestone`. Le couple de `milestone` fonctionne comme des ancres qui délimitent le début et la fin du segment annoté (figure 4).

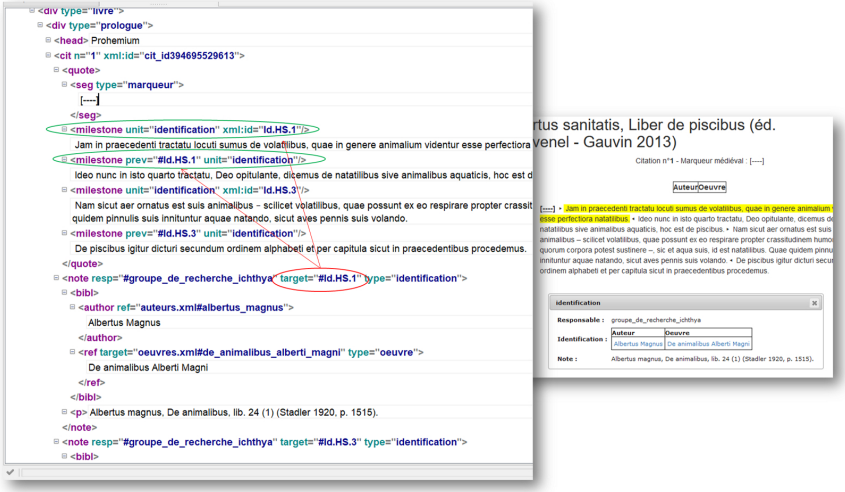

Fig. 4 : Structure des identifications de sources

2.) les annotations sur la transmission sont également contenues dans des notes. À la différence des identifications, ces notes sont ancrées à un point précis de la citation, là encore en mode « stand-off embarqué ». L'attribut `@target` sur l'élément `note` pointe vers une ancre placée dans le texte, sous la forme de l'élément `anchor`. Les deux types de notes, annotations et identifications, sont distingués au moyen d'un attribut `@type` qui permet de définir une typologie (figure 5).

[18] Le mode « stand-off » représente une solution possible aux problèmes posés par des couches d'annotation multiples et potentiellement concurrentes. Lorsque l'annotation est externe au fichier, elle permet aussi de laisser le document source vierge de toute interprétation (ce qui n'est plus le cas lorsqu'on l'encode), bien que la seule transcription puisse être déjà considérée comme un acte d'interprétation. Voir Spadini, Turska, 2019.

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

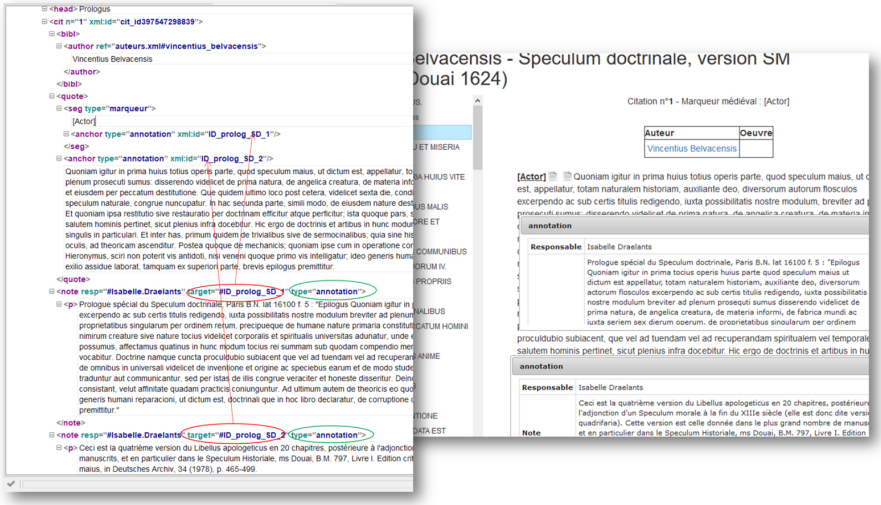

Fig. 5 : Structure des annotations

3.) des fiches bio-bibliographiques, appelées « mémentos auteurs » et « mémentos œuvres », qui fonctionnent comme un thésaurus d'autorités externe, et dont la structure peut être particulièrement développée selon les cas, notamment pour refléter les questions de transmission, à l'aide des informations de type auteur réel, auteur allégué, pseudépigraphie, mais aussi version d'œuvre.

On trouve également dans ce corpus une dimension supplémentaire formée par les liens vers les fiches bio-bibliographiques, au moyen de noms canoniques uniques dont chacun fonctionne comme un identifiant unique pour chaque fiche, vers lequel pointent un certain nombre d'éléments du corpus :

- les noms canoniques liés aux marqueurs de citation. Les marqueurs représentent la mention de l'autorité médiévale (auteur ou œuvre ou les deux) telle qu'elle est invoquée par le compilateur (figure 6)

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

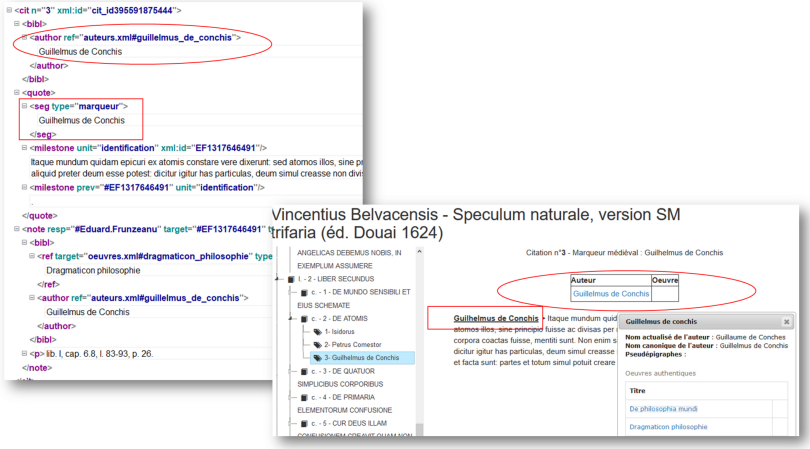

Fig. 6 : Les noms canoniques liés aux marqueurs de citation

- les identifications de sources, qui permettent de préciser, compléter voire corriger l'information de la source donnée par le compilateur, après consultation de la ou des éditions de référence de la source (figure 7).

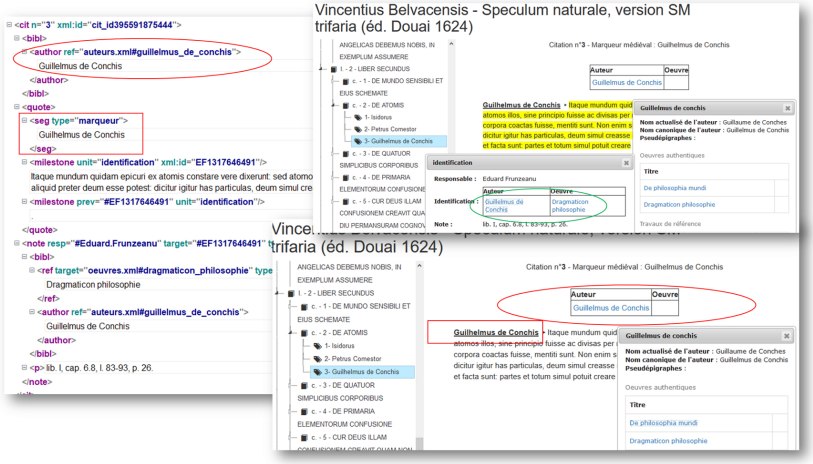

Fig. 7 : Les noms canoniques dans les identifications

- les liens hypertextes à l'intérieur même des mémentos, qui permettent de rebondir d'un mémento à l'autre, en consultant les œuvres authentiques et pseudépigraphiques d'un auteur, ou les auteurs sous le nom desquels il circule (figure 8).

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

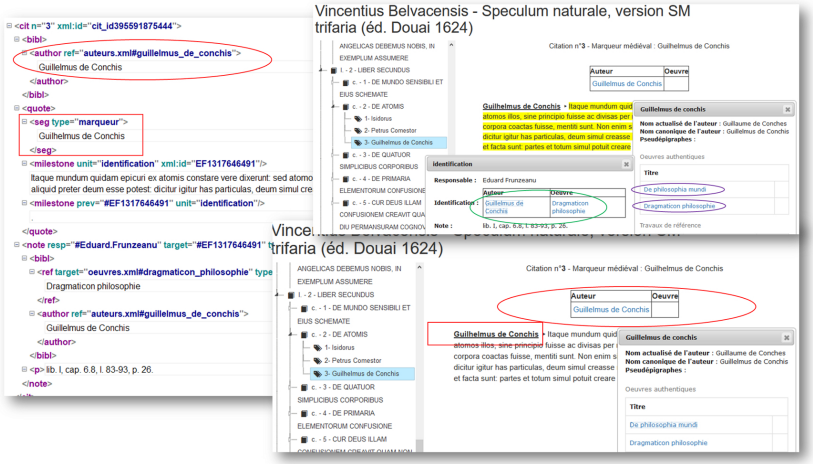

FIG. 8 : Les liens entre mémentos *via* le nom canonique

4.2 Besoins, fonctions et problèmes

Au début du projet, l'idée d'un encodage par les chercheurs eux-mêmes semblait hors de portée. Il était également nécessaire que les textes restent éditables pendant tout le processus, car la tâche d'identification des segments de citation, pensée pour être menée dans une phase du travail plus tardive que l'encodage du texte, pouvait amener à redécouper les citations. Cette contrainte a posé des problèmes majeurs aux équipes informatiques et nécessité la création d'une plateforme collaborative d'édition couplée à une base de données PhP-MySQL présentant, d'un côté, les mémentos stockés en un certain nombre de fiches et, de l'autre, les textes du corpus stockés dans les tables sous forme de fragments de code XML-TEI comprenant les textes du corpus, mais aussi les notes en mode stand-off embarqué pour l'identification et l'annotation.

À partir de la création de la base de données et de la plateforme collaborative, l'enrichissement du corpus se faisait selon les trois phases illustrées par l'image 9 ci-dessous (Fig. 9).

La première possibilité est d'utiliser un éditeur XML, comme par exemple Oxygen en mode « Texte », ce que nous avons fait pour certaines encyclopédies. Cette méthode, qui facilite l'accès au code brut et à la manipulation des données, a néanmoins quelques inconvénients. Cette même facilité d'accès permet de dévier rapidement de la structure, ce qui est tentant lorsqu'on travaille sur des encyclopédies qui n'entrent pas exactement dans le cadre défini au départ pour l'encyclopédie très structurée de Vincent de Beauvais, ou lorsqu'on se retrouve face à des textes présentant un rapport lâche à la source, ou des sources multiples et difficiles à séparer au sein des citations. Cependant, tout écart du schéma prévu initialement pose immanquablement des problèmes au moment de l'exploitation des données encodées. Un autre inconvénient est que cette méthode suppose une intervention manuelle, ce qui peut devenir très lourd lorsqu'il s'agit de réaliser des identifications, d'insérer un nom canonique

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

ou simplement de numéroter des citations, des éléments de structure, ou d'attribuer des identifiants uniques. La tâche d'encodage, qui représente déjà en soi un travail long et contraignant, devient alors fastidieuse et démultiplie les possibilités d'erreurs.

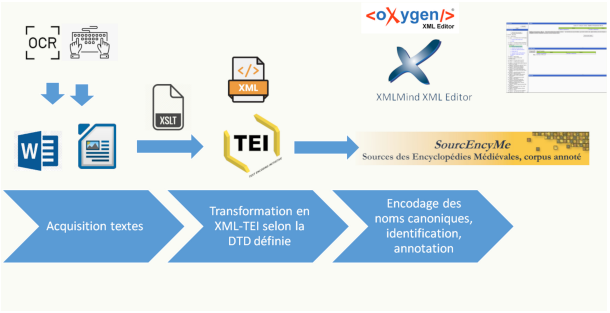

Fig. 9 : Workflow du projet SourcEncyMe

La deuxième solution, celle d'encoder les textes sans passer par un éditeur XML, a été de les enrichir sur la plateforme collaborative créée par L'ATILF. Celle-ci permet :

- d'ajouter des noms canoniques à la citation ou de modifier ceux figurant déjà dans la base (figure 10)

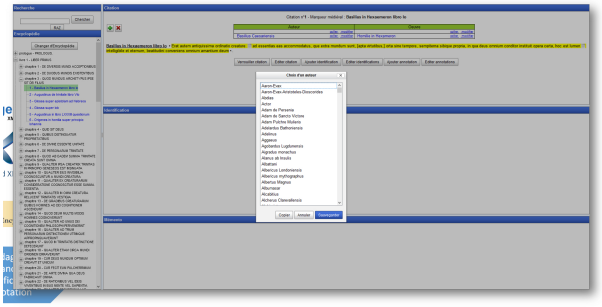

Fig. 10 : La plateforme collaborative : accès aux noms canoniques

- d'identifier des segments de citation en enregistrant les liens vers les noms canoniques auteur et œuvre de la source, ou en les modifiant
- d'ajouter des annotations dans le texte
- éventuellement, d'éditer le texte brut de la citation. Cette possibilité a été ajoutée à la demande des chercheurs afin de pouvoir corriger des coquilles.

La nature complexe du système plateforme collaborative/base de données rend à présent compliquée sa maintenance et sa mise à jour. Aujourd'hui nous trouvons

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

de plus en plus fréquemment des erreurs provoquées par le dédoublement de certains types de données dans la base ou par un dysfonctionnement du procédé d'identification des segments de citation. Ceci nous a amenés à modifier la chaîne de travail et donc les outils utilisés, afin de réduire la distance formelle séparant le chercheur des données, et donc de limiter la possibilité d'introduire des erreurs. Deux changements majeurs ont été opérés :

- la migration de la base de données en PhP-MySQL vers une base de données XML native (BaseX)
- l'abandon de la plateforme collaborative et le passage par un environnement de balisage dans un éditeur XML pour la phase d'encodage des textes

5 Migration de la base de données vers BaseX

La migration de la base de données a été effectuée à l'été 2020 ; elle nous a donné l'occasion de faire converger la structure des textes du corpus et des mémentos de SourcEncyMe et la structure XML-TEI qui sous-tend les éditions électroniques et les thesauri du pôle « Document numérique » de la MRSH de Caen[19]. En effet, l'objectif était d'utiliser les outils développés à Caen, notamment deux réalisations :

- un environnement de balisage en XML-TEI sous XMLMind Editor (XXE) destiné aux compilations médiévales et utilisé pour encoder les textes du projet Ichtya sur la faune aquatique[20]. Il a permis la publication de l'édition électronique du livre sur les poissons de l'*Hortus sanitatis*[21].
- et le plugin PluCo[22], qui s'intègre dans les environnements de balisage et qui est destiné à l'indexation des données par le biais des *thesauri,* et à l'enrichissement des notices d'autorités contenues dans ceux-ci[23].

6 Workflow en construction

Afin de permettre l'enrichissement des textes du corpus une fois transformés en documents XML-TEI, plusieurs outils ont été créés dont un environnement de balisage personnalisé.

La modification de l'environnement de base s'appuie sur une table de concordance entre la structure implémentée dans l'environnement Ichtya pour les compilations et la structure voulue pour le corpus SourcEncyMe, ainsi que sur le manuel du pôle numérique mis en ligne pour l'environnement de balisage des compilations et qui décrit le schéma utilisé.

[19] La migration a été effectuée par un stagiaire en Licence 3 Mathématiques et Informatique Appliquées aux Sciences Humaines et Sociales de Nancy, Boubou Kane. Son travail a été encadré et poursuivi par Henri Seng, ingénieur informaticien au Pôle numérique de l'IRHT.

[20] https://www.unicaen.fr/recherche/mrsh/document_numerique/outils/compilations ; http://www.xmlmind.com/home.html

[21] https://www.unicaen.fr/puc/sources/depiscibus/

[22] https://www.unicaen.fr/recherche/mrsh/document_numerique/outils/pluco

[23] https://www.unicaen.fr/recherche/mrsh/document_numerique/outils/thesauri

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

L'environnement propose de travailler avec plusieurs vues, dont la première permet de composer et structurer son document en ajoutant des divisions et des unités de citation (figure 11).

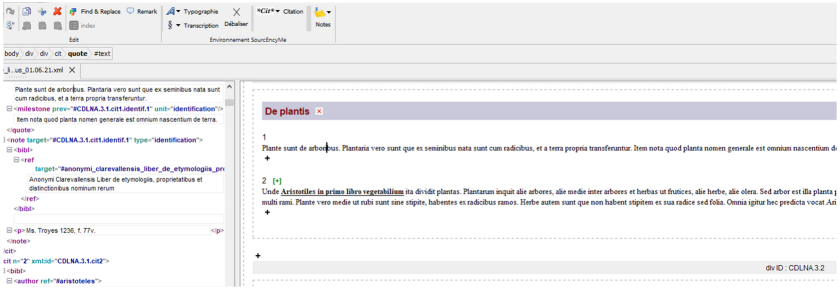

Fig. 11 : L'environnement de balisage sous XXE : vue de transcription

Une seconde vue permet d'ajouter les références aux mémentos auteur et œuvre liées au marqueur médiéval, que l'on va préciser grâce à une liste interactive qui exploite l'ensemble des mémentos (figure 12).

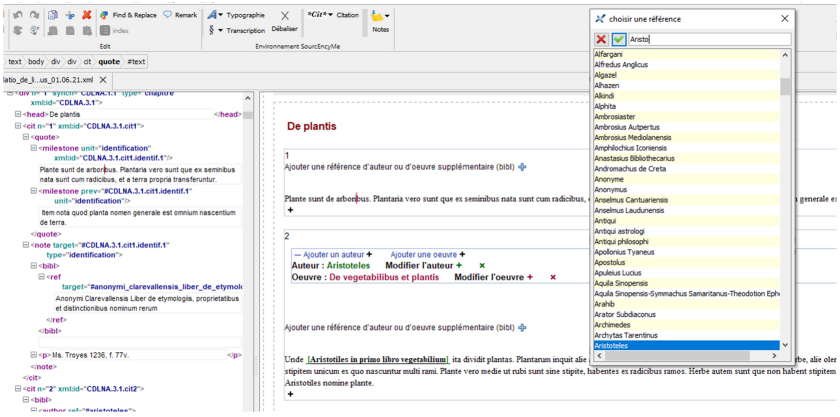

Fig. 12 : L'environnement de balisage sous XXE : vue de gestion des autorités

Enfin une troisième vue permet d'ajouter des annotations et des identifications (figure 13).

Le second outil est un environnement de balisage pour les mémentos, qui permet de créer de nouvelles fiches bio-bibliographiques ou d'éditer celles qui existent. Dans la mesure du possible, nous avons essayé de nous rapprocher le plus possible de la structure implémentée dans les thesauri du Pôle document numérique de Caen et dans l'environnement de balisage de notices de thesauri, avec quelques différences liées à la spécificité de la structure de SourcEncyMe, et notamment à l'importance accordée aux informations sur la transmission des sources manuscrites et à leur identification.

Il est ainsi possible, au sein du mémento auteur, de trouver des liens vers les mémentos des œuvres comme ici dans celui de Thomas de Cantimpré (figure 14).

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

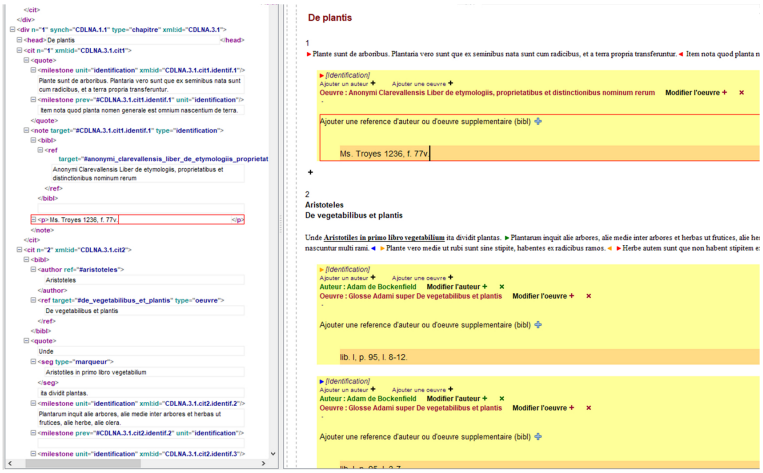

FIG. 13 : L'environnement de balisage sous XXE : vue d'identification et d'annotation

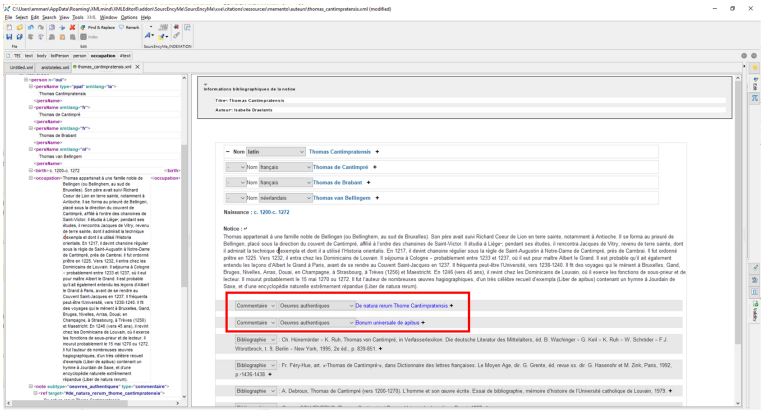

FIG. 14 : L'environnement de balisage des mémentos : liens entre fiches

Le cas échéant, on trouvera également la liste des œuvres pseudépigraphiques à la suite des œuvres authentiques comme dans le mémento auteur d'Aristote (figure 15).

L'encodage direct en XML-TEI permet d'ajouter des informations qui n'existaient pas dans la base de données initiale, comme la mention de révision des fiches, associée à la date et à l'identifiant de la personne responsable.

Pour ce qui est du mémento œuvre, on peut préciser les informations d'*incipit* ou d'*explicit*, et on peut faire la liste des témoins manuscrits (figure 16).

7 Évolutions et développements : enjeux et pistes pour l'avenir

Parmi les évolutions attendues, la configuration espérée du plugin PluCo devrait permettre aux collaborateurs d'interroger les mémentos à distance et de les enrichir, voire

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

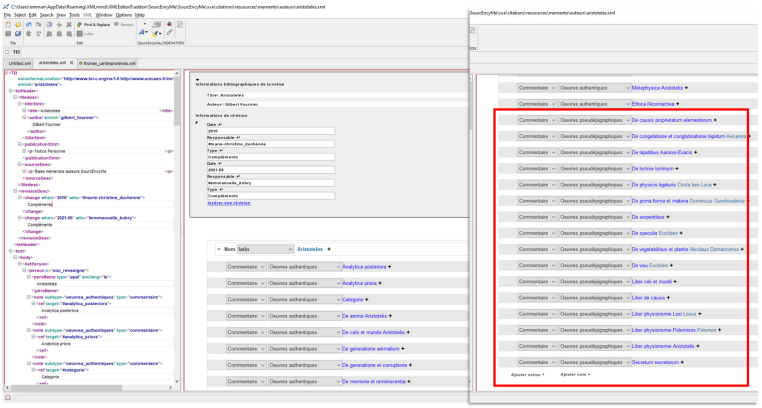

Fig. 15 : L'environnement de balisage des mémentos : liens vers les œuvres dans le mémento auteur

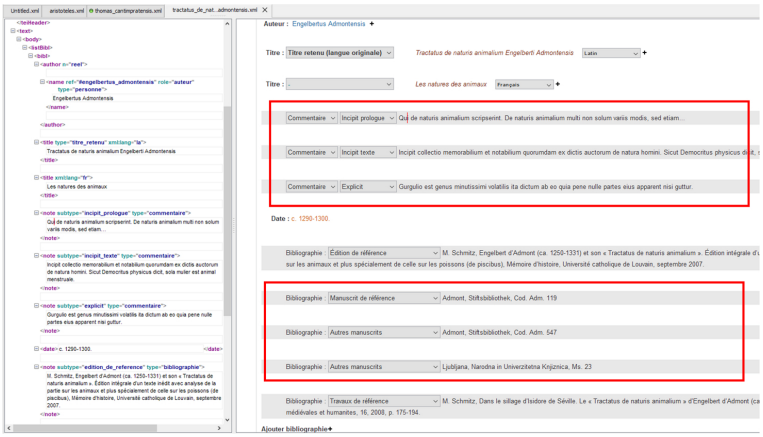

Fig. 16 : L'environnement de balisage des mémentos : informations sur les témoins manuscrits

d'éditer les textes. Un défi important est en effet celui du travail collaboratif, ce que permettait la plateforme initialement créée. Pour retrouver cette dimension, plusieurs alternatives sont actuellement à l'étude. On peut par exemple utiliser un système de *versioning* pour l'édition des fichiers et la mise à jour de ceux-ci sur le serveur central.

Nous prévoyons également de moderniser l'aspect et le design du site, et d'y associer un bref Webinar décrivant l'utilisation du site.

Les collaborations sont elles aussi un enjeu, elles sont essentielles à la survie du projet et à son élargissement. Dans cette perspective, citons par exemple une réflexion menée à l'échelle nationale sur la mutualisation des outils et des schémas d'encodage. L'IRHT est en effet engagé, avec le pôle « Document numérique » et les PUC, ainsi qu'avec l'École des Chartes, dans une réflexion sur la convergence des schémas d'encodage de sources pour l'édition électronique, associée à la mutualisation des outils d'encodage

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

et de publication. Cet effort de réflexion, qui a été ralenti par la crise sanitaire en 2020-2021, sera prolongé et amplifié dans le cadre de l'Equipex BIBLISSIMA+ à partir de 2022. Par ailleurs le travail réalisé autour des mémentos auteur et œuvre participe d'une réflexion globale sur l'alignement et la mise en lien des listes d'autorités construites ou à construire dans chaque projet s'occupant de sources anciennes. Cet aspect est aussi amené à être développé dans le futur.

À l'intérieur-même de l'IRHT, un enjeu important est de multiplier les liens et la compatibilité avec les données d'autres corpus en ligne, dans un double effort de standardisation et de création d'identifiants pérennes et uniques, regroupant, entre autres, les appellations multiples pour les œuvres comme pour les cotes de manuscrits. Dans le passé, l'IRHT a conçu tous ses outils séparément. Depuis quelques années, un mouvement d'harmonisation s'est amorcé et a été renforcé par la nécessité de traitement et de standardisation des autorités et des référentiels, une ambition qui faisait le cœur de l'Equipex BIBLISSIMA, adossé à l'IRHT dès 2012. Une solution se met peu à peu en place autour de l'identifiant unique pour unifier les cotes de manuscrits, mais on est loin d'une standardisation des outils.

Les cotes de manuscrits sont recensées dans la base de données « Medium » (medium.irht.cnrs.fr), qui contient près de 80.000 cotes et devient, à l'IRHT, le portail central des liens vers les reproductions de manuscrits, les dépôts dans les bibliothèques, et les diverses bases de données et corpus d'information sur les manuscrits produits à l'IRHT. Chaque manuscrit doit y recevoir désormais un identifiant unique. Ainsi, toute cote de manuscrit traité dans un projet IRHT (comme p. ex. la *Bibliographie annuelle du Moyen Âge tardif,* en ligne : BAMAT-O, ou l'incipitaire en ligne *In principio*) a vocation à s'y trouver alignée et à recevoir un identifiant unique[24] sur lequel seront alignées les variantes formelles de la cote utilisées dans la bibliographie. Quant à la base « provenance » des œuvres, des manuscrits et désormais des imprimés anciens, « Bibale », elle totalise 53.000 notices pour 215.000 items « manuscrits » identifiés – avec cependant un grand nombre de doublons voués à disparaître[25] ; elle s'accroît constamment et multiplie ses collaborations et ses projets avec les bibliothèques et dépôts de manuscrit français et étrangers, souvent *via* des accords portés par l'Equipex BIBLISSIMA. Une première étape consiste en l'harmonisation et l'alignement des référentiels œuvres et auteurs (« noms canoniques ») de SourcEncyMe et d'ICHTYA (Caen, voir plus haut) en vue de l'intégration dans les métadonnées de BIBALE.

Dans l'ensemble des textes transmis par les savants médiévaux, théologiens et philosophes, la Bible représente une part très considérable. Elle véhicule avec elle tout un apparat de commentaires sous forme de gloses, souvent alléguées par les encyclopédistes car elles constituent une partie fondamentale de leur bagage intellectuel. Il est donc hautement souhaitable de multiplier les liens et l'interopérabilité avec le projet sur la glose ordinaire de la Bible « Gloss-e[26]. La thématique des « autorités » et les objectifs d'identification d'un patrimoine textuel médiéval sont en effet communs à SourcEncyMe et à Gloss-e tout comme leur objet, la citation et l'intertextualité. Toute

[24] Une forme de cote générée automatiquement qui permet le regroupement et l'alignement de toutes les formes de cote se rapportant au même manuscrit.

[25] https://bibale.irht.cnrs.fr/

[26] *Glossae Scripturae Sacrae-electronicae* https://gloss-e.irht.cnrs.fr/ sous la direction de Martin Morard.

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

la pensée médiévale, et en particulier la pensée exégétique, use de jeux de correspondances entre les textes d'autorités et transmet un héritage sous forme de citations, dans un paysage doctrinal commun d'autorités hiérarchisées.

Un dernier enjeu de SourcEncyMe est la critique d'attribution *via* l'identification des citations. S'il constitue la part la plus considérable des métadonnées en évolution, il ne peut être atteint sans l'aide de chercheurs qui travaillent sur des traditions textuelles particulières ou sont spécialistes d'un auteur ou d'une œuvre en particulier et désirent participer au projet. Nous avons besoin de leur compétence pour mener des campagnes d'identification pour chacune des près de 2000 œuvres mentionnées dans les encyclopédies ; eux-mêmes tireront profit d'un repérage systématique de la réception de tel auteur ou œuvre dans les compilations encyclopédiques. C'est ainsi que pourront s'accroître les métadonnées reliées aux noms canoniques des œuvres et des auteurs, constituées au minimum par le renvoi à l'édition de référence à l'intérieur des mémentos ou des identifications de sources, mais aussi de liens vers les manuscrits numérisés et les éditions anciennes en ligne, ainsi que vers des sites d'érudition spécialisés relatifs à une source mentionnée dans le corpus.

En outre, la question de la poursuite de tentatives d'automatisation partielle de l'identification des sources continue à se poser, mais elle se heurte au manque de moyens humains et ne pourrait se concrétiser, pour le moment, que *via* une collaboration avec une autre équipe dont ce serait le projet.

Nous avons désiré donner à cette présentation le sens d'un retour d'expérience concret intégré dans une réflexion plus ample sur les conditions matérielles et intellectuelles des humanités numériques. En effet, l'adaptation aux possibilités et à l'évolution des outils est un enjeu constant dont l'ampleur ne pouvait être imaginée initialement. Ceux qui se sont lancés dans l'aventure ont constaté qu'elle n'avait pas tenu toutes ses promesses, dont la plus séduisante, celle de faire gagner du temps. C'est l'inverse : créer des outils numériques nécessite temps, énergie, argent. L'obsolescence des outils immatériels ainsi créés est une menace imposant la prévision de mises à jour des supports matériels et de pérennisation des données, en dépit d'un discours public mettant l'accent sur l'innovation et la création de nouveaux projets. Pour bon nombre de projets, ces conditions se sont traduites par un gaspillage financier et parfois par une perte de sens, car à devoir, pour obtenir des fonds, justifier son activité d'érudit par la mise en ligne quantitative de textes et accepter de mettre en ligne un *work in progress*, beaucoup ont eu l'impression de travestir une expertise longuement acquise. Le changement de culture et de pratiques impose aussi d'apprendre des techniques étrangères à son premier domaine de compétences. Une alternative d'avenir est néanmoins possible, celle de réserver des moyens à une ingénierie scientifique spécialisée en analyse de sources et mise en œuvre par un personnel capable de saisir les enjeux scientifiques des projets, d'aider les chercheurs à formaliser des modélisations de données et de configurer des outils pour l'encodage de sources.

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352