Pour une encyclopédie interactive du latin médiéval : le Semantic Web au service de la lexicographie médiolatine

Bruno Bon, Krzysztof Nowak

Pour une encyclopédie interactive du latin médiéval : le *Semantic Web* au service de la lexicographie médiolatine

Des projets de recherche

Lors de la précédente table-ronde des rédactions de dictionnaires du latin médiéval rattachés à l'Union Académique internationale (Bruxelles), réunie à León en 2010, François Dolbeau, directeur du *Novum Glossarium Mediae Latinitatis (NGML),* avait souligné l'opportunité, pour les équipes qui le souhaitaient, de participer au projet de programme européen COST « Medioevo Europeo[1] » coordonné par la SISMEL (Firenze) et l'Institut de Recherche et d'Histoire des Textes (IRHT, Paris). Plusieurs d'entre elles ont répondu à cet appel, à titre principal (France, Pologne, République Tchèque) ou secondaire (Allemagne). Invités à titre d'experts pour la première réunion du groupe de travail sur les textes et dictionnaires (*Working Group* 3, WG3) à Heidelberg en 2011, nous avons participé à une discussion sur les différents moyens susceptibles de lier entre eux dictionnaires et corpus de textes. Nous y avons défendu l'intérêt du traitement automatique des langues (TAL) et du *Text Mining*. Afin de prolonger la discussion dans le cadre restreint des dictionnaires de latin médiéval, nous avons organisé deux jours d'atelier en février 2012 à Zurich (*Mittellateinisches Seminar*[2]), où nous avons présenté les nouveaux enjeux d'une publication électronique collective. Avec notre collègue Renaud Alexandre, nous y avons finalement été encouragés à construire un prototype pour la prochaine réunion du WG3 à Saint-Jacques de Compostelle (novembre 2012), et si possible avant celle de Munich (septembre 2012), pour que nous puissions l'y présenter aux collègues lexicographes absents de « Medioevo Europeo ».
Mais il convient de reconnaître que ce travail n'aurait pas vu le jour sans le support d'institutions de recherche capables d'en financer l'avancement, au contraire du programme COST. Outre le programme français de l'Agence nationale de la recherche (ANR) OMNIA[3], plusieurs fois évoqué dans notre revue, les deux programmes polonais *eLexicon Mediae et Infimae Latinitatis Polonorum*[4] et *Fontes Mediae et Infimae Latinitatis Polonorum (1000-1550)*[5] sont destinés à développer, pour le latin médiéval, dictionnaires et corpus de textes numérisés. Avec le soutien de l'IRHT (France) et de l'Instytut Języka Polskiego (Institut de la Langue Polonaise, Pologne), qui participent activement au développement des humanités numériques dans notre domaine, ces projets ont permis la mise en place d'une efficace collaboration.

Wiki et interactivité

A l'origine du projet ANR OMNIA, l'idée de développer une encyclopédie interactive du latin médiéval à partir de la numérisation de dictionnaires existants, en particulier du *Glossarium mediae et infimae latinitatis* de Charles du Cange (DuC) et du *NGML*, est à l'honneur d'Alain Guerreau (Centre de Recherches Historiques, Paris). A cette fin, nous proposons aujourd'hui d'utiliser un service de type « wiki », bien connu des internautes par ses multiples applications, à commencer par l'encyclopédie multilingue « Wikipedia », fondée sur « MediaWiki ». A bien y réfléchir, le choix de ce logiciel s'impose comme une évidence pour une édition électronique commune des dictionnaires de latin médiéval. C'est d'abord, en effet, une interface expressément conçue pour une information présentée en articles repérés par des vedettes, qui correspond donc très bien au contenu lexicographique ; c'est ensuite une interface qui permet l'interrogation simultanée d'ensembles distincts, préservant ainsi l'intégrité des instruments qu'elle réunit.
Les avantages de « MediaWiki » sont nombreux : gratuit et entièrement libre, ce logiciel prêt à

1 www.medioevoeuropeo.org
2 Y ont participé, à l'invitation de la Suisse (Carmen Cardelle de Hartmann, Philipp Roelli et Peter Stotz), pour l'Allemagne, Helena Leithe-Jasper et Johannes Staub ; pour la Castille et le León, Estrella Perez-Rodriguez ; pour la France, Renaud Alexandre et Bruno Bon ; pour la Pologne, Krzysztof Nowak et Michał Rzepiela ; pour la République Tchèque, Pavel Nyvlt et Zuzana Silagiova.
3 www.glossaria.eu
4 *Elektroniczny słownik łaciny średniowiecznej na ziemiach polskich :* www.scriptores.pl
5 *Korpus języka łacińskiego na ziemiach polskich (1000-1550) :* www.scriptores.pl

l'emploi dispense ses utilisateurs des contraintes du développement, assuré par une communauté suffisamment importante pour être garanti sur le long terme ; objectifs fondateurs du projet, le support multilingue et les outils de collaboration sont implémentés au cœur du logiciel, non par raccroc ; enfin, la simplicité d'utilisation se conjugue avec la richesse des liens internes. Au contraire, les quelques inconvénients, inhérents à une interface adoptée, ne sont pas de nature à en réduire l'intérêt : on relèvera en particulier une certaine limitation de l'arborescence des données représentées.

Un outil d'édition

Notre proposition s'inscrit dans la perspective prometteuse du « Web sémantique », d'abord appelé *Semantic Web*, puis « Web de données » pour éviter l'ambiguïté trompeuse du mot *Semantic.* Pour simplifier, il s'agit de donner du relief au contenu HTML, comme le XML donne du relief au texte brut, afin de concevoir l'internet comme une gigantesque base de données. Dans un dictionnaire, la vedette, la définition et les références devront donc être encodées comme telles, exactement comme pour une édition électronique digne de ce nom. Dans ces conditions, le *Semantic Web* ou « Web de données » est le moyen idéal d'utiliser les richesses de l'encodage existant de nos dictionnaires numérisés. Parfaite illustration de la vivacité du développement de « MediaWiki », le module « Semantic MediaWiki[6] » (SMW) est spécialement conçu pour le « Web 3.0 ».
Le prototype[7] que nous présentons ici comprend six articles[8], issus de quatre dictionnaires[9] modernes de latin médiéval. Pour conserver l'individualité de ces instruments de travail, toutes les pages de l'encyclopédie sont précédées d'un préfixe correspondant à leur dictionnaire d'origine (respectivement CZ, DE, EU et PL) : ce sont les « espaces de noms ». Ces pages sont de deux types (catégories), selon qu'elles correspondent à une référence (auteur, texte) ou à un article (lemme). Les pages d'auteurs et d'œuvres sont directement issues de l'index des sources de chaque dictionnaire, dont elles reprennent chaque fois une ligne ; aux renseignements fournis par l'édition imprimée, nous avons ajouté trois indications supplémentaires, nécessaires à l'évolution des types d'interrogation (date normalisée, type de texte, lieu de production) ; des liens vers les éditions en ligne sont prévus. Les pages de lemmes sont également importées des dictionnaires, par l'intermédiaire d'un filtre de transformation du format XML vers le format SMW ; outre une présentation classique, reproduisant la structure et les informations de la version imprimée, et bénéficiant néanmoins d'un lien actif vers les pages de références, l'encyclopédie comporte un onglet avancé, sous forme de diagramme (types de texte), de frise chronologique et de carte géographique, trois types d'affichage inédits pour des articles de dictionnaire ; enfin, un troisième onglet regroupe de nombreux liens directs vers d'autres instruments de recherche en ligne.

Un outil d'interrogation

La page d'accueil de l'encyclopédie doit répondre aux divers besoins de ses utilisateurs. Elle propose donc plusieurs types d'interrogation, de la consultation traditionnelle d'un seul ouvrage à l'interrogation la plus avancée. Bien entendu, il est d'abord possible de choisir le dictionnaire à interroger, et d'en consulter la liste des articles ou l'index des sources. On peut également passer aisément d'un dictionnaire à l'autre, pour un même article, en utilisant des liens directs. Mais pour une consultation de tous les dictionnaires à la fois, la fenêtre de recherche générale, qui bénéficie de l'auto-complétion, renvoie à des pages d'hyperlemmes, où sont regroupées les informations issues des différents instruments. Comme dans « Wikipedia », les liens vers les articles existants apparaissent en bleu, et les liens orphelins en rouge. Enfin, la recherche en plein texte permet aussi d'accéder aux autres champs interrogeables.
Outre ces procédures d'interrogation classique, notre prototype encyclopédique offre un nouveau

6 www.semantic-mediawiki.org
7 www.scriptores.pl/wiki, adresse provisoire.
8 Ces articles ont été choisis en regard de leur présence simultanée dans plusieurs dictionnaires et de leur longueur limitée, qui réduit l'ampleur du traitement manuel des références.
9 Outre le *NGML,* ont été utilisés pour ce prototype le *Lexicon mediae et infimae latinitatis Polonorum (LMILP),* le *Mittellateinisches Wörterbuch (MLW)* et le *Latinitatis medii aevi lexicon Bohemorum (LMALB).*

mode d'interrogation des dictionnaires, que l'on pourrait qualifier de « recherche visuelle » : la frise chronologique et la carte des citations de tous les articles correspondant à la requête offrent la possibilité de sélectionner les occurrences en fonction de leur localisation dans le temps ou dans l'espace, et d'en visualiser immédiatement le détail.
Enfin, dans le cadre d'une recherche avancée, l'utilisation du module « Semantic Drilldown » permet de filtrer tout le contenu de l'encyclopédie selon les principaux caractères encodés dans ses pages, et d'afficher une sélection de références (par zone, période ou type) ou de lemmes (par catégorie, définition ou domaine). Les résultats de la requête sont aussi présentés sous les différents formats évoqués plus haut (tableau, chronologie, carte).

Un outil de collaboration

Cette encyclopédie interactive doit enfin grandement faciliter la collaboration scientifique entre les équipes de lexicographes d'une part, et entre les rédacteurs de dictionnaires et leurs utilisateurs d'autre part. L'efficacité de « MediaWiki » dans ce domaine du « Web 2.0 » n'est plus à démontrer, qui supporte tous les niveaux souhaités de collaboration, par l'intermédiaire des comptes d'utilisateurs. Une gestion des droits différenciée permet, par exemple, d'interdire l'intervention des utilisateurs anonymes, de réserver la page de discussion aux utilisateurs enregistrés, et de n'accorder les droits complets (écriture, modification, suppression) qu'aux équipes de rédaction.
La création d'articles nouveaux, par insertion de lemmes déjà publiés ou en cours de rédaction, peut se faire directement ou par l'intermédiaire d'un formulaire[10]. La syntaxe du SMW étant notablement moins complexe que celle du XML, le passage du second vers le premier peut se faire automatiquement à l'aide d'une feuille de transformation. La correction – ou la suppression – des articles, loin du procédé très inefficace des *errata* et *addenda*, s'effectue aussi facilement par l'onglet correspondant. Pour plus de sûreté, toutes ces actions sont enregistrées (et sauvegardées) dans un historique, ce qui permet de revenir sans dommage sur une intervention malencontreuse.
La page de discussion, ouverte à tous les utilisateurs enregistrés, est le moyen le plus souple d'offrir, pour la première fois de l'histoire de la lexicographie latine, un espace visible et permanent aux collègues, chercheurs ou étudiants, qui souhaiteraient réagir à la lecture d'un article. Cette page est associée aux hyperlemmes, pour permettre des commentaires croisés : elle est donc unique pour chaque mot. Nous souhaitons vivement que ces pages de discussion permettent enfin de créer un échange réel entre les rédacteurs de dictionnaire et leurs utilisateurs.

Deux conditions seulement

Pour pouvoir s'intégrer dans cette encyclopédie interactive, ouverte à toutes les équipes de lexicographie médiolatine, les données numérisées ne doivent remplir qu'une seule condition indispensable : la liberté totale et gratuite. En effet, dans le cadre d'un « wiki », aucune collaboration ne peut être envisagée sans la garantie que tous les articles et références de la base de données pourront être librement téléchargés et copiés, sous licence libre, et à l'exclusion de toute utilisation commerciale.
Enfin, pour faciliter le travail concret d'insertion des dictionnaires dans l'encyclopédie, il conviendrait sans doute de préférer, après leur numérisation, un encodage de type XML. Mais n'importe quel autre type de fichier structuré peut être converti sans difficulté au format SMW, le seul texte brut, sans encodage, étant à exclure absolument.

Bruno BON (CNRS-IRHT, Comité Du Cange, Paris)
Krzysztof NOWAK (Instytut Języka Polskiego-PAN, Pracownia Łaciny Średniowiecznej, Krakow)

10 Il en va de même pour les pages de références, mais l'efficacité des diverses fonctions d'interrogation et d'affichage fondées sur le SMW suppose un enrichissement minimal (date, lieu, type) de l'index des sources de chaque dictionnaire, que seules les équipes de rédaction sont en mesure d'entreprendre. Il s'agit là du seul travail scientifique nécessaire à la mise en œuvre de l'encyclopédie.