Bulletin du centre d'études médiévales d'Auxerre | BUCEMA

14 | 2010
Varia

OMNIA: outils et méthodes numériques pour l'interrogation et l'analyse des textes médiolatins (2)

Bruno Bon

Édition électronique
URL : http://journals.openedition.org/cem/11566
DOI : 10.4000/cem.11566
ISSN : 1954-3093

Éditeur
Centre d'études médiévales Saint-Germain d'Auxerre

Édition imprimée
Pagination : 251-252
ISSN : 1623-5770

Référence électronique
Bruno Bon, « OMNIA: outils et méthodes numériques pour l'interrogation et l'analyse des textes médiolatins (2) », *Bulletin du centre d'études médiévales d'Auxerre | BUCEMA* [En ligne], 14 | 2010, mis en ligne le 14 octobre 2010, consulté le 24 avril 2019. URL : http://journals.openedition.org/cem/11566 ; DOI : 10.4000/cem.11566

OMNIA: outils et méthodes numériques pour l'interrogation et l'analyse des textes médiolatins (2)

Bruno Bon

1 Depuis le 1er janvier 2009, et pour une durée de quatre ans, l'UMR 5594/ARTeHIS (Dijon) est associée à l'École nationale des chartes (Paris) et à l'équipe de lexicographie latine de l'UPR 841 (IRHT, Paris) pour mettre en œuvre le projet ANR «*Omnia*, outils et méthodes numériques pour l'interrogation et l'analyse des textes médiolatins» (http://www.glossaria.eu/).

2 Tout au long de sa première année d'existence, le projet «Omnia» a eu la chance de bénéficier d'une remarquable implication collective de ses membres [1], pourtant surchargés d'activités de plus en plus «prioritaires» et le plus souvent administratives: sans leur extraordinaire énergie, les trois perspectives de développement présentées dans le dernier bulletin du Centre d'études médiévales n'auraient pas avancé aussi sérieusement [2], qu'ils en soient ici tous chaleureusement remerciés! On saluera également l'arrivée de deux jeunes étudiants, Amandine Postec et Nicolas Perreaux, venus opportunément renforcer les forces vives de ce groupe.

3 La récupération des outils existants (perspective 1), dont l'objectif est de constituer la base d'une future encyclopédie interactive du latin médiéval, est la partie la plus visible et la plus avancée du projet. Depuis décembre 2009, l'intégralité du *Glossarium* de Du Cange – y compris le petit glossaire de français médiéval – est librement accessible et interrogeable (en plein texte) sur le site de l'École des chartes (http://ducange.enc.sorbonne.fr/), dans une interface dédiée entièrement refondue par cette institution; sur ce corpus, l'essentiel du travail, ô combien indispensable mais fastidieux, pris en charge par l'École des chartes, consiste à corriger les fichiers XML fournis par le prestataire extérieur. Parallèlement, au début de l'année, la numérisation des lettres disponibles (L-P) du *Novum Glossarium Mediae Latinitatis* – dictionnaire international de latin médiéval en cours de rédaction – a commencé, également à l'École des chartes; un

travail préparatoire sur les deux derniers fascicules *(Phacoides-Plaka)*, rédigés sous forme informatique, devrait permettre d'accélérer le traitement des images des fascicules antérieurs, complétant ainsi, avant la fin de l'année, le corpus de base du projet.

4 Afin de mettre rapidement à la disposition de la communauté scientifique les premiers résultats de ses investissements, une interface de consultation commune aux deux premiers dictionnaires, en cours de développement, sera accessible prochainement sur le site du projet «Omnia» : la partie gauche de la page proposera une liste déroulante mêlant les lemmes disponibles dans les différents corpus d'origine, afin de reconstituer virtuellement le «dictionnaire» le plus complet possible; les principaux champs de recherche choisis par les rédacteurs européens de dictionnaires médiolatins, lors de leur réunion quadriennale de 2004 à Barcelone, feront l'objet d'un formulaire en haut de page; l'essentiel de l'affichage sera réservé aux résultats, par dictionnaire, dans un format complet ou abrégé. Dans l'esprit d'un projet évolutif et ouvert, la plateforme devant être capable d'accueillir tous les dictionnaires nationaux de latin médiéval qui souhaiteraient la rejoindre, l'architecture de l'ensemble – de type *LAMP, Linux-Apache-Mysql-Php* – répond à des impératifs de simplicité et de robustesse. Cette interface devra, par la suite, intégrer les fonctionnalités interactives destinées à développer une réflexion collective sur le lien entre structures de sens et structures sociales (perspective 2).

5 Le développement de procédures de recherche formalisées et assistées (perspective 3) ne se conçoit pas, pour le latin médiéval, sans une lemmatisation préalable des corpus textuels utilisés. C'est, de loin, le sujet qui a concentré le plus d'efforts cette année, dans le cadre de la mise au point du premier outil nécessaire à son développement: un fichier de plus de 76000 «variantes» médiolatines associées à leur lemme, leur flexion, leur marqueur grammatical *(POS, Part-Of-Speech)* et leur «définition». Une fois ces données rassemblées, il s'est d'abord agi de modéliser autant que possible la vérification des diverses informations – dédoublonnement, concordance entre lemme et «variante», attribution du *POS*, de la flexion, etc. Mais la nature même de la distribution du lexique interdisait d'en rester à un traitement automatisé, et c'est ensuite «à la main» qu'il a fallu reprendre cet énorme ensemble, partagé pour l'occasion entre la quasi-totalité des membres du groupe de travail. Ce très long et fastidieux travail de relecture devrait être achevé avant l'été 2010, pour laisser la place à la flexion quasi automatisée de toutes ces «variantes», et aboutir à un fichier de plusieurs centaines de milliers de formes lemmatisées. Il s'agira ensuite de réunir un ensemble de textes «représentatifs» (!) du latin médiéval, afin de les marquer semi-manuellement et de pouvoir enfin envisager de lancer l'apprentissage du *Tree-Tagger*.

INDEX

Mots-clés : OMNIA