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

15 | 2011
Varia

OMNIA : outils et méthodes numériques pour l'interrogation et l'analyse des textes médiolatins (3)

Bruno Bon

Édition électronique
URL : http://journals.openedition.org/cem/12015
DOI : 10.4000/cem.12015
ISSN : 1954-3093

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

Référence électronique
Bruno Bon, « *OMNIA* : outils et méthodes numériques pour l'interrogation et l'analyse des textes médiolatins (3) », *Bulletin du centre d'études médiévales d'Auxerre | BUCEMA* [En ligne], 15 | 2011, mis en ligne le 19 février 2013, consulté le 02 mai 2019. URL : http://journals.openedition.org/cem/12015 ; DOI : 10.4000/cem.12015

OMNIA : outils et méthodes numériques pour l'interrogation et l'analyse des textes médiolatins (3)

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 Pour sa deuxième année d'exercice, le projet « *Omnia* » n'a pas eu à se plaindre de l'engagement personnel de ses membres [1], qui n'ont ménagé ni leur peine ni leurs efforts pour le faire progresser efficacement : nous tenons, ici, à les en remercier sincèrement. Le départ administratif de notre ingénieur contractuel, Frédéric Glorieux, recruté en janvier 2010 par l'École des chartes, a d'autant moins affecté la bonne marche du travail collectif que nous avons eu la chance de bénéficier au même moment de l'arrivée de Renaud Alexandre, nouvel ingénieur de recherche dans l'équipe de lexicographie latine de l'IRHT. Enfin, dans le cadre d'une collaboration officielle avec le *Lexicon mediae et infimae latinitatis polonorum* – dictionnaire du latin médiéval polonais rédigé à Cracovie –, nous avons invité le rédacteur, Krzysztof Nowak, à participer à nos échanges et à nos travaux.

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 toujours la partie la plus avancée du projet. Depuis décembre 2009, l'intégralité du Glossarium mediæ et infimæ latinitatis de Charles 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/). Parallèlement, depuis décembre 2010, la numérisation des lettres disponibles (L-P) du Novum glossarium mediae latinitatis (NGML) – dictionnaire international de latin médiéval en cours de rédaction – étant achevée, une version provisoire des deux derniers fascicules (Phacoides-Plaka) est en ligne, également

gratuitement, sur un site de l'IRHT accessible à partir du portail du projet « *Omnia* » (http://www.glossaria.eu/). On y trouvera également un lien vers la base de données de l' *Index scriptorum* du *NGML*, dans sa version cumulative (1973-2005), dont l'intérêt bibliographique ne se limite pas à la consultation des éditions utilisées pour la rédaction du dictionnaire. Le premier semestre 2011 sera consacré à la formalisation d'un schéma XML susceptible de permettre non seulement l'interrogation efficace de l'ensemble des fascicules parus, en cours de nettoyage, mais surtout la rédaction sous ce format des fascicules à venir.

4 Comme annoncé dans le bulletin précédent, un prototype d'interface de consultation commune aux deux premiers dictionnaires est accessible depuis l'été 2010 sur un site de l'École des chartes (http://omnia.enc.sorbonne.fr) : sur la partie gauche de la page, une liste déroulante mêle les lemmes disponibles du *Glossarium* (en rouge) et du *Novum glossarium* (en vert), reconstituant ainsi un prototype de « dictionnaire » virtuel [2] ; sur les huit champs de recherche retenus, seuls trois sont actifs [3], outre la recherche en plein texte ; les résultats s'affichent par dictionnaire, dans un format complet. 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 concevant pas sans une lemmatisation préalable des corpus textuels, c'est encore le sujet qui a concentré le plus d'efforts cette année. La relecture du fichier de plus de 76 000 lemmes et variantes, associés à leur flexion, leur marqueur grammatical (POS, Part-Of-Speech) et leur traduction, a occupé les six membres latinistes du groupe jusqu'à l'automne 2010, chargés de normaliser autant que possible les indications devant permettre d'automatiser la flexion. Parallèlement, le développement du fléchisseur, destiné à produire les formes attendues dans les textes, à partir de la liste des lemmes, s'est révélé beaucoup plus long et compliqué que prévu, le linguiste et l'ordinateur n'ayant absolument pas la même conception de l'exception grammaticale ! Malgré tout, fin décembre 2010, tout était prêt pour lancer le premier apprentissage de *TreeTagger* : une liste de plusieurs centaines de milliers de formes associées à leur *POS* et à leur lemme, un extrait de la *TreeBank* – petit corpus de textes lemmatisés – de la base de latin classique « *Perseus* » (http://www.perseus.tufts.edu), et quinze mille mots de divers types de textes de latin médiéval à lemmatiser. L'analyse de ces résultats, au début de l'année 2011, permettra ensuite de reprendre l'apprentissage du lemmatiseur sur un ensemble de textes de plus en plus important.

NOTES

1. Bruno Bon, Olivier Canteaut, Marie-José Gasse-Grandjean, Frédéric Glorieux, Anita Guerreau-Jalabert, Alain Guerreau, Eliana Magnani, Nicolas Perreaux et Amandine Postec.

2. On prendra garde, par exemple, qu'en vertu de l'évolution des usages lexicographiques, la vedette d'un verbe est à l'infinitif dans le *Glossarium*, et à la première personne du singulier de l'indicatif présent dans le *Novum glossarium*.

3. Lemmes, variantes, citations.

INDEX

Mots-clés : OMNIA