Medialatinitas. Pour une intégration superficielle de ressources textuelles et lexicales en latin

Krzysztof Nowak[1], Bruno Bon[2], Renaud Alexandre[2]

[1]Institut de la langue polonaise - Académie polonaise des sciences (IJP-PAN) Kraków – Pologne

[2]Institut de recherche et d'histoire des textes (IRHT-CNRS) Paris – France

Abstract

Medialatinitas is a lightweight Web application which integrates dictionaries, corpora and encyclopaedic resources for Latin. The integration takes places principally on the level of the user-friendly interface. The main objectives of *Medialatinitas* are: improving access to distributed data; challenging separation of linguistic and encyclopaedic information in lexicographic description; compensating for deficiencies of existing lexicographic resources; building a community of users who apply computational methods in their study of Latin texts. As for the architecture, *Medialatinitas* is implemented as as mashup application: a user's query (as of now, only lemma search is supported) is processed and despatched to both local and distant services (RESTful APIs, SPARQL endpoints); the results are subsequently displayed on the main page as a set of separate widgets. As a whole the widgets contribute to an extensive description of Latin lemmas according to their grammatical, semantic and cultural properties.

Résumé

Medialatinitas est une application Web légère qui agrège des dictionnaires, des corpus et des ressources encyclopédiques portant sur le latin. Les ressources sont intégrées au niveau d'une interface utilisateur conviviale. Les objectifs principaux de *Medialatinitas* sont les suivants : faciliter l'accès à des données dispersées, effacer la séparation, au sein du discours lexicographique, de l'information linguistique et de l'information encyclopédique, compenser les insuffisances des ressources lexicographiques existantes, permettre la construction d'une communauté d'utilisateurs mettant en œuvre des méthodes numériques pour étudier des textes latins. Sur le plan de l'architecture, *Medialatinitas* est conçu comme un *mashup* : la requête d'un utilisateur (pour l'instant, seule la recherche par lemme est proposée) est traitée et transmise à des services locaux aussi bien que distants (API RESTful, points d'accès SPARQL) ; les résultats sont affichés sur la page principale sous la forme de modules distincts. Conçus comme un tout, les modules permettent une description extensive des lemmes latins en fonction de leurs propriétés grammaticales, sémantiques et culturelles.

Key words : latin médiéval ; interface de recherche ; intégration superficielle ; lexicographie électronique ; statistique lexicale ; visualisation des données

1. Introduction

La langue latine occupe une place prépondérante dans l'histoire de l'Europe. Son usage a été constant pendant plus de quinze siècles sur un territoire allant de la Suède à l'Italie et du Portugal à la Pologne, pour répondre à des besoins très divers (droit, théologie, liturgie, etc.) Cet emploi très large de la langue a eu des conséquences sur son homogénéité. Une quantité immense de textes témoigne ainsi d'une grande influence du temps, de l'espace et des genres de textes aussi bien dans l'usage que dans le sens des mots. Le besoin de comprendre ces textes a conduit les érudits à élaborer peu à peu un vaste corpus de littérature secondaire, dont les dictionnaires sont un élément capital.

Malgré tout, ces ressources demeurent aujourd'hui – même sous leur forme numérique – largement insatisfaisantes pour une étude du vocabulaire médiéval : les textes ne sont généralement pas diffusés en corpus, mais comme de simples collections de textes, répondant à des problèmes spécifiques, et les interfaces de recherche sont généralement très limitées. Pas d'unité non plus du côté des dictionnaires, qui couvrent des zones géographiques et chronologiques différentes, n'utilisent pas le même métalangage et ont chacun leur style éditorial. Bref, la nécessité de consulter tous ces outils les uns après les autres ne permet pas d'obtenir facilement une vue d'ensemble pour l'étude d'un mot.

Medialatinitas est une application web en cours de développement (Nowak et Bon 2015), dont le but est de mettre en œuvre une intégration efficace des ressources en latin ou portant sur le latin, qu'elles soient textuelles, lexicographiques ou encyclopédiques, par le biais d'une interface agréable et conviviale.

2. Objectifs, données, mise en œuvre

2.1. Objectifs

Medialatinitas a été créé dans plusieurs buts : regrouper virtuellement des ressources dispersées, faciliter la rédaction des dictionnaires, stimuler la recherche sur le vocabulaire latin (spéc. médiéval), diffuser des méthodes innovantes d'étude des textes latins, participer à la constitution d'une communauté de chercheurs souhaitant appliquer des méthodes numériques dans les domaines de la philologie, de l'histoire et de la linguistique latines.

2.1.1. Confort de l'utilisateur et recherche collaborative

En regroupant virtuellement des ressources éparses, notre objectif est d'éviter aux utilisateurs la perte de temps et d'énergie que représente une recherche en plusieurs endroits, tâche qui implique à chaque fois d'avoir une idée claire du champ couvert par les données, de leur fiabilité, de la syntaxe de requête, etc.

Il s'agit également de fournir des outils d'analyse, de permettre aux utilisateurs d'en avoir un usage autonome, et de leur offrir la possibilité de les améliorer ou d'en proposer d'autres. Au moment de concevoir une application, il est en effet impossible de penser à tous les usages qui seraient susceptibles d'en être faits. A terme, *Medialatinitas* permettra à ses utilisateurs de tester et proposer leurs modules sous la forme de blocs de code R et Javascript, qui pourront ensuite être implémentés dans l'application. Nous souhaitons également, aux marges de l'application, proposer une base de connaissances comportant un ensemble de manuels pratiques, de démonstrations et de liens, ainsi qu'une description complète du déroulement des tâches à mener à bien, depuis l'OCR jusqu'à l'interrogation du corpus.

2.1.2. Répondre à des questions anciennes, en poser de nouvelles

Chaque type de source a ses défauts. Les dictionnaires traditionnels de latin médiéval, pour ne citer qu'eux, reposent sur un dépouillement manuel, forcément biaisé, des sources (Guerreau-Jalabert et Bon, 2010 ; Bon, sous presse) ; ils ne traitent convenablement que les mots de fréquence relativement basse, et ne donnent qu'un aperçu limité de la fréquence des termes (par exemple, en utilisant la formule *saepissime* 'très souvent', pour un terme jugé très fréquent par le lexicographe) et de leur répartition dans le temps et l'espace. L'usage de corpus et d'outils numériques en combinaison avec les dictionnaires permettrait de compenser ces insuffisances.

En concevant *Medialatinitas*, nous avons donc souhaité effacer la barrière entre les différents types de ressources à solliciter simultanément si l'on souhaite comprendre la société médiévale à partir des mots qu'elle nous a transmis. Pour y parvenir, et pour compenser les défauts inhérents à chaque ressource considérée isolément, *Medialatinitas* permet leur interrogation simultanée et les regroupe dans un seul ensemble afin d'élaborer une description cohérente du potentiel signifiant du mot, ses propriétés grammaticales et syntaxiques, sa fonction culturelle.

A cet égard, l'ajout de données encyclopédiques est essentiel, et constitue – avec l'usage d'outils avancés de visualisation (Theron et Fontanillo 2015) et de statistique – la différence principale entre *Medialatinitas* et d'autres portails combinant des données lexicographiques et textuelles, comme *Logeion* (http://logeion.uchicago.edu) et le *Dictionnaire vivant de la langue française* (http://dvlf.uchicago.edu/). Comprendre un texte médiéval signifie comprendre l'univers de référence dans lequel il a été produit, un univers radicalement différent du nôtre dans le cas de l'Occident médiéval (Guerreau 2001). Cette compréhension étant insuffisante si l'on s'en tient aux dictionnaires (Geeraerts 2009), il nous a semblé opportun de proposer des compléments d'information à la fois plus généralistes et plus spécialisés, d'autant plus que le champ d'usage du latin médiéval a été très vaste (musique, théologie, droit, philosophie, etc.)

En outre, une intégration plus étroite de données encyclopédiques et lexicographiques est souhaitable également pour des raisons pratiques : la majorité des dictionnaires généraux de latin médiéval excluent les noms propres, or un décodage correct des toponymes et des anthroponymes est crucial pour comprendre les textes anciens. La possibilité d'interroger aussi les noms propres pourra se révéler utile si la motivation du nom propre vient d'un nom commun, et inversement (*aqua*, 'eau' ou *vicus*, 'bourg' figurent dans de nombreux toponymes).

2.2. Données

Medialatinitas utilise largement les ressources numériques libres déjà disponibles. Les données qui sont intégrées au sein de l'application web peuvent être divisées en trois groupes. D'une part, des ressources lexicographiques : des dictionnaires de latin classique, médiéval et moderne, élaborés par des scientifiques (par exemple, le *Novum Glossarium Mediae Latinitatis* [F. Blatt et al., 1957] ou le *Lexicon Mediae et Infimae Latinitatis Polonorum* [M. Plezia et al., 1953]) ou par des communautés d'utilisateurs (le Wiktionnaire latin [http://la.wiktionary.org]), ainsi que des dictionnaires de toponymes ou d'anthroponymes antiques et médiévaux (entre autres, *Pelagios Project* [http://pelagios.dme.ait.ac.at/api] et *Orbis Latinus* [http://orbis-latinus.geog.uni-heidelberg.de]). D'autre part, des corpus (par exemple *eFontes*, corpus de latin médiéval polonais [http://scriptores.pl/fontes], ou les *Croatiae Auctores Latini* [http://ffzg.unizg.hr/klafil/croala]) et des collections de textes (comme le projet *Perseus* [http://www.perseus.tufts.edu/hopper], la *Patrologie latine* de Migne, etc.). Enfin, des ressources encyclopédiques : encyclopédies (Wikipedia, notamment sa version latine), recueils de proverbes et citations (Wikiquote latin [https://la.wikiquote.org]), dépôts de documents et d'images (Europeana [http://europeana.eu]), bibliothèques numériques (Internet Archive [https://archive.org], Open Library [https://openlibrary.org]), listes d'auteurs médiévaux (auteurs cités dans le *Novum Glossarium* [http://glossaria.eu/scriptores/], VIAF [http://viaf.org]), portails d'archives scientifiques ouvertes (Hal-SHS [https://halshs.archives-ouvertes.fr]) ou ressources mixtes (par exemple BabelNet [http://babelnet.org]).

La diversité de ces ressources a une implication sur les modalités d'accès à ces données, et contribue à la complexité de la tâche d'intégration, dans la mesure où nombre de ces ressources sont exploitées en l'état.

2.3. Mise en œuvre : construction et architecture de l'application

Le modèle d'intégration de *Medialatinitas* est qualifié de superficiel, notamment parce que cette intégration est essentiellement mise en œuvre au niveau de l'interface utilisateur (UI). Cette combinaison de données externes fait par ailleurs de *Medialatinitas* un *mashup*, défini comme « a composite application developed starting from reusable data, application logic, and/or user interfaces typically, but not mandatorily, sourced from the Web » (Daniel et Matera 2014, 3).

2.3.1. Structures des données

Les versions numériques des dictionnaires, développées en interne sous la forme de fichiers XML TEI, reposent sur un schéma partagé d'encodage. Les corpus (externes ou internes) sont généralement diffusés sous la forme de fichiers XML comprenant au moins quelques métadonnées d'ensemble. Les textes dont nous disposons ont été pourvus par nos soins d'un marquage morphosyntaxique au moyen de TreeTagger (Schmid 1994), avec un fichier de paramètres latins qui repose en partie sur les textes de la Perseus Digital Library et de l'Index Thomisticus. A terme, nous comptons utiliser les paramètres destinés au latin médiéval issus du projet ANR Omnia (http://glossaria.eu/treetagger).

Les ressources externes sont, pour l'essentiel, interrogées par l'intermédiaire de leur API RESTful ou leurs points d'accès SPARQL, de telle sorte que *Medialatinitas* ne prend pas en considération les formats originaux des données. Les données hébergées localement, elles aussi, sont ou seront interrogées par l'application web via leurs API : les dictionnaires, installés dans une instance eXist-db, sont interrogés par leurs API respectives ; les textes issus d'une OCR interne et les collections moins structurées sont stockées dans des index Lucene d'une instance eXist-db, et interrogés au moyen d'une API RESTful ; les corpus textuels sont interrogeables par CQPWeb (Hardie 2012), qui n'offre pas encore d'API, et n'est donc pour l'instant utilisé que comme outil d'interrogation avancée.

Les services locaux servent également à enrichir les données et faire des calculs : WikiLexicographica (Bon et Nowak, 2013), une implémentation de Semantic MediaWiki (Krötzsch et al. 2006), combine plusieurs dictionnaires avec une dimension chronologique et géographique, et permet une visualisation enrichie des données ; une session R (R Core Team 2015) est fournie à l'application web au moyen d'une API OpenCPU (Ooms 2014), et permet de faire des calculs sur les ressources lexicales et lexicographiques ; le paquet rcqp (Desgraupes et Loiseau 2012) permet l'interrogation du moteur CQP, et les scripts de statistique lexicale d'A. Guerreau (https://github.com/medialatinitas) permettent de trouver les cooccurrents du lemme dans le corpus, de même que le paquet wordspace de S. Evert (Evert 2014) permet de calculer les similarités entre les mots.

2.3.2. Niveaux d'affichage

Trois niveaux d'affichage successifs sont proposés : un aperçu général, une vue étendue, et une vue avancée.

Sur la page principale, l'utilisateur ne voit qu'un formulaire de recherche simple ; il entre sa requête (pour l'instant, seule la recherche sur un lemme est implémentée), qui est envoyée aux

services et API locaux et distants. Les résultats sont récupérés, traités puis affichés sur une seule et même page, agencée comme une grille composée de modules distincts, chacun d'entre eux étant chargé d'afficher une partie de l'information sur le mot en question. La page, conçue comme un tout, a pour but de donner un aperçu général et varié du sens, des propriétés linguistiques et de la distribution du mot. Les modules implémentés à l'heure actuelle permettent, entre autres, l'affichage d'extraits de définitions issues de dictionnaires de latin classique et médiéval, d'extraits de concordance issus des corpus, des propriétés morphosyntaxiques (genre, désinences, etc.) issues des dictionnaires, de la distribution des formes du mot dans les corpus, des graphiques illustrant la distribution chronologique et typologique du lemme, de ses cooccurrents, des termes similaires, des traductions (issues de BabelNet), des listes de citations qui comportent le mot recherché (issues du Wiktionnaire latin), des titres d'ouvrages contenant le mot (issus d'Internet Archive), des images (issues d'Europeana) contenant le mot dans leur description, d'une carte des toponymes comprenant le lemme (issus entre autres de Pelagios). *Medialatinitas* utilise différentes formes d'affichage des données pour ces modules : tableaux, listes, graphiques ou cartes.

Une fois examinée cette page principale, l'utilisateur pourra se rendre sur la page de vue étendue pour se concentrer sur tel ou tel aspect du mot. Pour l'instant, seul un tableau fondé sur shiny (Chang et al. 2015) et permettant de faire des statistiques lexicales a été développé.

Enfin, les applications natives (CQPweb pour les corpus, eXist-db pour les dictionnaires, R shiny pour les calculs) seront accessibles à partir de la vue étendue.

3. Conclusions et perspectives

Medialatinitas est un premier jalon vers une intégration plus poussée de ressources lexicographiques, textuelles et encyclopédiques. Pour l'heure, cette intégration, à titre de test, est conçue comme superficielle, mais permet d'ores et déjà d'offrir un aperçu original de vocables latins. L'intégration des données devra être approfondie, afin de faciliter l'interrogation des lemmes (pour l'instant, corpus et encyclopédies ne font pas référence à un système commun d'identifiants, et les entrées de dictionnaires ne sont pas alignées). Par son aspect modulaire, l'application est ouverte à des évolutions ultérieures, et nous y voyons un moyen d'expérimenter un nouveau type de discours lexicographique à l'âge du Web de données linguistiques.

References

Blatt F., Lefèvre Y., Monfrin J., Dolbeau F. and Guerreau-Jalabert A. editors (1957-2011). Novum Glossarium Mediae Latinitatis. URL : http://www.glossaria.eu/ngml.

Bon B. (sous presse). Histoire et perspectives du 'Novum Glossarium Mediae Latinitatis'. In *Proceedings of the 7th International Conference on Historical Lexicography and Lexicology (ICHLL 2014)*. Peter Lang.

Bon B. and Nowak K. (2013). WikiLexicographica: Linking Medieval Latin Dictionaries with Semantic MediaWiki. In Kosem I., Kallas J., Gantar P., Krek S., Langements M. and Tuulik M. editors, *Electronic Lexicography in the 21st century, Thinking outside the paper: Proceedings of the eLex 2013 Conference*, pages 407-420. URL : http://eki.ee/elex2013/proceedings/eLex2013_28_Bon+Nowak.pdf.

Chang W., Cheng J., Allaire JJ., Xie Y. and McPherson J. (2015). *shiny: Web Application Framework for R*. URL : http://CRAN.R-project.org/package=shiny.

Daniel F. and Matera M. (2014). *Mashups: concepts, models and architectures*. Springer.

Desgraupes B. and Loiseau S. (2012). *rcqp: Interface to the Corpus Query Protocol*. URL : https://r-forge.r-project.org/projects/rcwb.

Evert S. (2014). Distributional Semantics in R with the wordspace Package. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, pages 110-114. URL : http://anthology.aclweb.org/C/C14/C14-2024.pdf.

Geeraerts D. (2009). *Theories of Lexical Semantics*. Oxford University Press.

Guerreau A. (2001). *Vinea*. In Goullet M. and Parisse M. editors, *Les historiens et le latin médiéval*. Publications de la Sorbonne, pages 67-73.

Guerreau-Jalabert A. and Bon B. (2010). Le trésor au Moyen âge : étude lexicale. In L. Burkart et al. editors, *Le trésor au Moyen âge*. Sismel, pages 11-31.

Hardie A. (2012). CQPweb – combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, 17(3) : 380-409.

Krötzsch M., Vrandečić, D., Völkel, M., Haller, H. and Studer, R. (2007). Semantic Wikipedia. *Journal of Web Semantics* 5(4) : 251-61.

Nowak K. and Bon B. (2015). medialatinitas.eu. Towards Shallow Integration of Lexical, Textual and Encyclopaedic Resources for Latin. In Kosem I., Jakubíček M., Kallas J., Krek S. editors, *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference*, pages 152-169. URL : https://elex.link/elex2015/proceedings/eLex_2015_10_Nowak+Bon.pdf.

Ooms J. (2014). The OpenCPU System: Towards a Universal Interface for Scientific Computing through Separation of Concerns. ArXiv e-prints. URL : http://arxiv.org/pdf/1406.4806v1.pdf.

Plezia M., Brożkowa C., Rzepiela M. editors (1953-2011). *Lexicon Mediae et Infimae Latinitatis Polonorum*.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. URL : http://www.R-project.org.

Schmid H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*, URL : http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger1.pdf.

Theron R. and Fontanillo, L. (2015). Diachronic-Information Visualization in Historical Dictionaries. *Information Visualization*, 14(2) : 111-36.