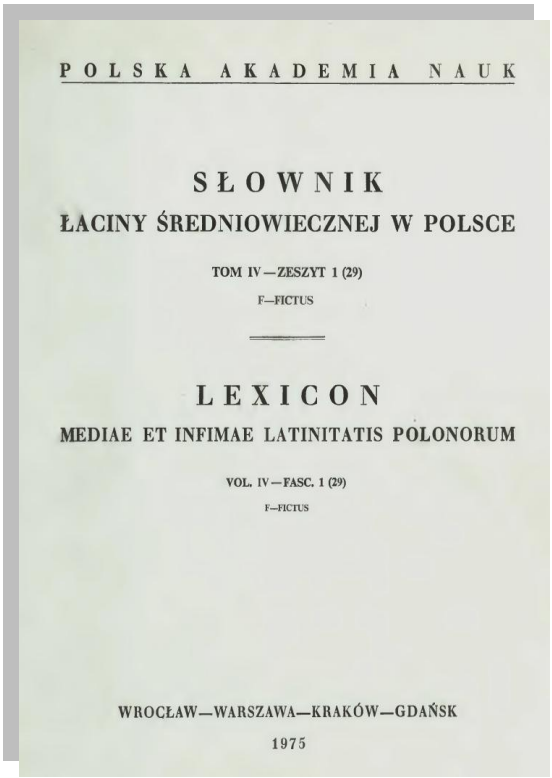
Les projets électroniques du Laboratoire de latin médiéval à l'Institut de la langue polonaise

Krzysztof NOWAK
Michał RZEPIELA
Institut de la langue polonaise,
Académie Polonaise des Sciences, Cracovie

I. Introduction


POLSKA AKADEMIA NAUK

SŁOWNIK
ŁACINY ŚREDNIOWIECZNEJ W POLSCE

TOM IV — ZESZYT 1 (29)
F—FICTUS

LEXICON
MEDIAE ET INFIMAE LATINITATIS POLONORUM

VOL. IV — FASC. 1 (29)
F—FICTUS

WROCŁAW—WARSZAWA—KRAKÓW—GDAŃSK
1975

En guise d'introduction aux projets électroniques de l'équipe du dictionnaire du latin médiéval polonais (*Lexicon mediae et infimae Latinitatis Polonorum*), il convient de rappeler brièvement les circonstances qui ouvrirent la voie à la mise en œuvre de ce dictionnaire. En 1920, face aux multiples déficiences du glossaire de la moyenne et basse latinité (*Glossarium Mediae et Infimae Latinitatis*) de Charles Du Cange, dont la première édition est parue au XVII[e] siècle et qui, amélioré à plusieurs reprises, fut jusqu'alors communément utilisé comme le dictionnaire le plus complet pour la période médiévale, l'Union Académique Internationale a lancé l'idée de rédiger un dictionnaire complètement nouveau. Lors d'une réunion préparatoire du « Comité Du Cange », responsable de la rédaction de ce dictionnaire, il a été demandé aux représentants des différentes académies adhérentes d'organiser le dépouillement dans leur propre pays des sources latines de l'époque médiévale, et de les envoyer ensuite au siège du Comité à Paris. Néanmoins, on s'est tout de suite rendu compte que le matériel ainsi rassemblé était trop riche et disparate, quant au cadre temporel, pour être réuni en entier dans un même dictionnaire, à moins d'omettre l'essentiel des dépouillements provenant des régions où le Moyen Age est habituellement défini comme « tardif ». C'était, entre autres, le cas de la Pologne. Il n'est

pas donc étonnant que le dictionnaire du latin médiéval polonais ait commencé à sortir parmi les premiers, même si, à cause des perturbations liées à la seconde guerre mondiale, le premier fascicule du dictionnaire du latin polonais n'a paru qu'en 1953.

Il convient de souligner que le patronage de l'UAI, ayant un caractère plus ou moins formel d'un dictionnaire à l'autre, a significativement contribué à la collaboration entre les équipes des dictionnaires nationaux. À partir du tournant des années 2000, cette collaboration s'est plus nettement orientée vers l'application des technologies nouvelles au travail lexicographique. L'objectif final y serait la création d'un outil permettant une consultation en ligne de l'ensemble des dictionnaires régionaux et internationaux du latin. Ajoutons que la mise en œuvre d'un tel projet a rencontré et rencontre évidemment encore certains obstacles. D'un côté, l'avancement de la numérisation des dictionnaires concernés n'en est pas au même point ; d'autre part, même si elles disposent de données numérisées, les équipes lexicographiques ne possèdent pas toujours les droits commerciaux sur leur travail. En outre, l'expérience montre que la coordination des travaux entre différentes équipes n'est pas facile si elles sont trop nombreuses, mais qu'elle devient en revanche beaucoup plus effective dans les groupes plus restreints, à savoir de deux ou trois équipes. À présent, c'est sans doute les équipes française et polonaise qui travaillent de la façon la plus étroite et qui sont les plus avancées dans le traitement des données électroniques de leurs dictionnaires. Nous nous rencontrons avec nos collègues du dictionnaire international (*Novum Glossarium Mediae Latinitatis*) au sein de plusieurs projets communs, mais nous conduisons également chacun nos propres projets, en partageant toujours nos expériences. Les personnes responsables (et à la fois «les protagonistes») de cette collaboration sont respectivement Krzysztof Nowak (IJP-PAN) pour le dictionnaire du latin polonais et Bruno Bon (IRHT-CNRS) pour la partie française.

Michał Rzepiela

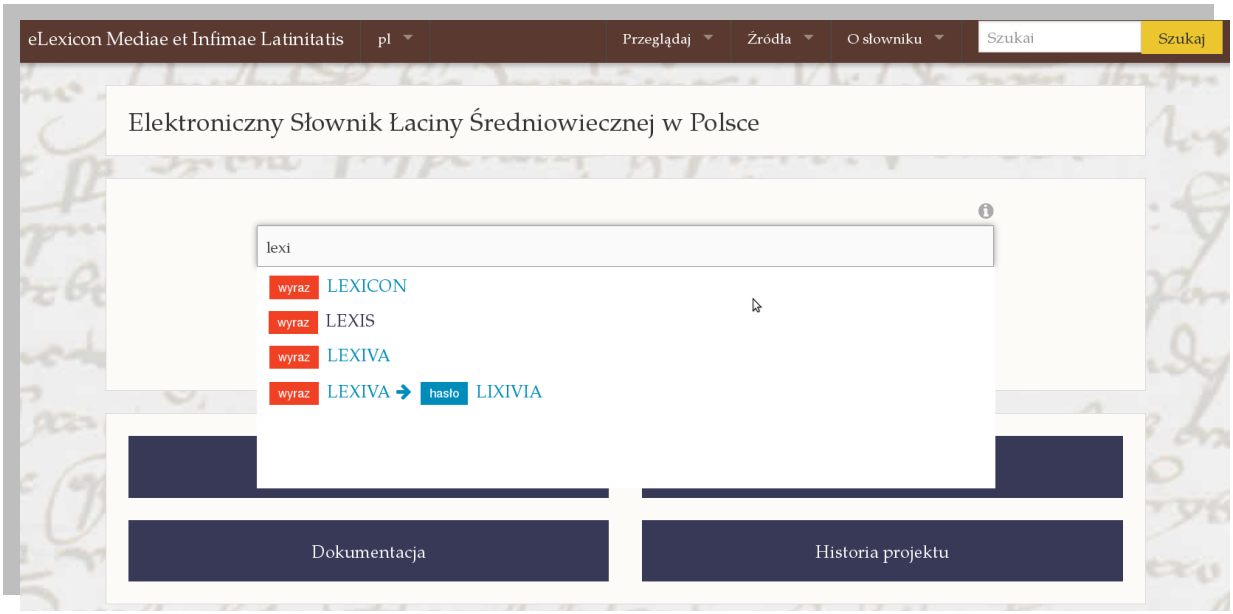
II. Le dictionnaire électronique du latin médiéval polonais[1]

Actuellement, le Laboratoire de latin médiéval à l'Institut de la langue polonaise (IJP-PAN) coordonne deux projets, sur le dictionnaire et le corpus du latin médiéval polonais. Le projet de dictionnaire électronique (lettres A-Q) a commencé en juin 2011 et vient de s'achever en juin 2014. Le travail du groupe de six personnes consistait à scanner les fascicules imprimés du dictionnaire, puis à corriger les fichiers océrisés par une annotation minutieuse. En premier lieu, une procédure d'annotation automatique a permis de supprimer les fautes d'OCR les plus fréquentes, tout en conservant les propriétés typographiques du dictionnaire qui paraissaient importantes pour l'utilisateur final. Elle fut suivie d'un encodage manuel chargé de rendre interprétable par l'ordinateur la grande richesse d'informations incluse dans le texte du dictionnaire. La sortie de cette phase de travaux fut enfin l'objet d'une transformation XSLT, pour aboutir à des fichiers XML plus ou moins compatibles avec le standard TEI.

[1] http://scriptores.pl/en/elexicon/.

L'interface du dictionnaire[2] a été construite pour répondre à trois objectifs principaux, et tout d'abord pour donner aux utilisateurs avancés la possibilité de formuler des enquêtes précises. C'était le but de l'encodage minutieux qui fut réalisé, et dont les chercheurs peuvent profiter dans le cadre de la « Requête avancée ». Le second objectif était d'élargir le public du dictionnaire en le rendant le plus accessible possible aux utilisateurs débutants. Ce fut réalisé à travers une nette séparation des différents utilisateurs au niveau d'un article, mais aussi grâce à un système de conseils et d'auto-complétion. Enfin, dès le début, le dictionnaire fut conçu comme un outil hybride, intégrant les autres ressources numériques : le corpus textuel du latin médiéval polonais et les fiches du dictionnaire numérisées[3] dans le projet RCIN[4].

L'interface en question propose plusieurs points d'accès au contenu lexicographique. La page d'accueil ne comporte que le menu et un seul champ d'enquête. L'utilisateur avancé peut accéder directement à l'interface de son choix, pendant que le débutant se servira plutôt du champ de base. Une liste de suggestions apparaît dès que l'utilisateur tape au moins trois lettres du mot qui l'intéresse.

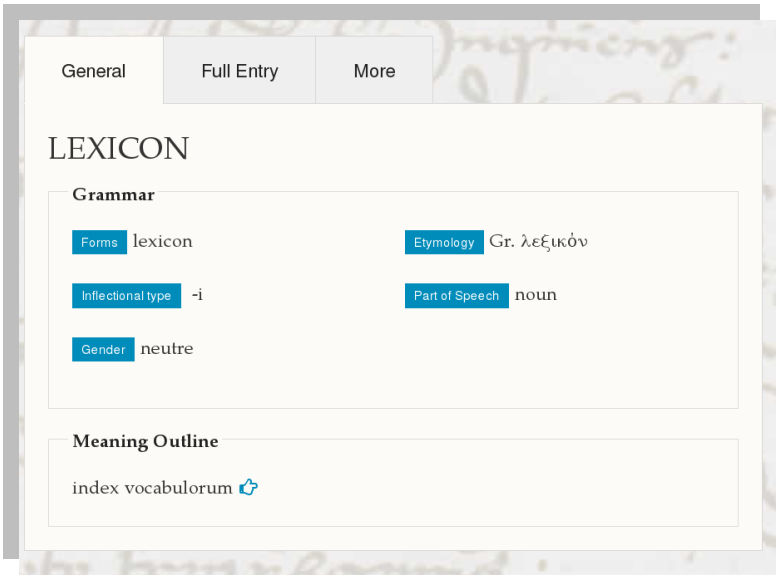


Si la forme recherchée correspond à un lemme attesté dans le dictionnaire, l'utilisateur est transféré à l'article correspondant, où deux affichages sont proposés, simple et complexe. Dans le premier onglet (simple), sous des chapeaux bien distincts, on trouve le résumé des propriétés morphosyntaxiques (conjugaison, partie du discours, *etc.*), de l'étymologie et du (ou des) sens du mot. La perspective avancée présente l'article entier au lecteur, mais pour rendre l'affichage plus convivial, on utilise un formatage varié, avec étiquetage en couleur, menu, *etc.*

[2] http://scriptores.pl/en/elexicon/
[3] http://www.rcin.org.pl/publication/31986.
[4] http://rcin.org.pl/dlibra/text?id=aboutRCIN.

Si la forme recherchée ne correspond à aucun lemme, l'utilisateur est transféré vers la page de désambiguïsation, où il trouve diverses propositions, selon le cas :

- la forme lemmatisée ;
- le résultat de la recherche dans les citations ;
- le résultat de la recherche dans les définitions lexicographiques ;
- la suggestion de mots similaires.



Le troisième point d'accès est offert par le formulaire de recherche avancée qui permet de profiter de toute la richesse de l'encodage. Ce formulaire, en deux parties, comporte :

- un champ dans lequel l'utilisateur peut définir une chaîne de caractères, limiter la recherche au champ de son choix et préciser la stratégie d'entrée, *i.e.* s'il s'agit d'une séquence entière ou partielle ;
- la liste des critères supplémentaires qui servent à filtrer les articles trouvés. Ces critères sont d'ordre tant morphosyntaxique qu'étymologique, chronologique, *etc.* Si l'utilisateur n'entre aucun mot dans le champ ci-dessus, la liste des critères peut servir d'outil d'exploration du lexique latin du Moyen âge en Pologne.

Comme il a été précisé plus haut, il est souhaitable que le dictionnaire serve de noyau à une infrastructure plus large pour la recherche médiolatine. Dès aujourd'hui, le lecteur est renvoyé à des ressources externes libres, dont il trouve les liens :

- sur la page de désambiguïsation, afin de suggérer aux utilisateurs d'élargir leur recherche, surtout si le mot n'a pas été trouvé ;
- dans chaque article du dictionnaire, pour enrichir la perspective avec les données des corpus, dictionnaires et encyclopédies.

III. Le corpus électronique du latin médiéval polonais[5]

Le deuxième projet de l'équipe de l'IJP-PAN vise à développer un corpus des textes latins polonais du Moyen âge à la Renaissance, entre 1000 et 1550 environ. Bien que de conception historique, ce corpus aura un caractère synchronique, car il permettra surtout d'interroger le latin médiéval. Mais l'inclusion des textes (les plus) tardifs permet d'offrir aux chercheurs l'opportunité de vérifier empiriquement les dates de la révolution de la Renaissance. Les cadres géographiques du corpus ont été définis aussi largement que possible. Pour rassembler les textes latins polonais, il fallait d'abord résoudre le problème des importantes modifications territoriales que la Pologne a subies pendant cette période, son territoire au XVI^{e} siècle étant très différent de celui du X^{e}. Pour contourner l'ambiguïté, les textes provenant des territoires voisins de la Pologne ont été pris en compte. Il est raisonnable d'espérer que cette définition élargie permettra de mieux mettre en valeur les phénomènes d'interférence entre le latin et les langues vernaculaires.

Le corpus en question sera généraliste, pour pouvoir représenter le latin dans la totalité de ses usages. Comme tel, il sera aussi représentatif et balancé. Dans la mesure du possible, le corpus sera constitué de textes entiers, ce qui permettra d'y inclure le vocabulaire tant des préambules que des clauses finales. Dans sa première version le corpus sera composé de cinq millions de mots, puis il sera enrichi en fonction des financements disponibles. En ce qui concerne le choix des textes, on y trouvera tous les genres de la production écrite du Moyen âge polonais. Les textes sont choisis d'après leur fonction communicative.

Actuellement il n'est pas prévu de développer d'outil informatique propriétaire pour les enquêtes dans le corpus. Au contraire, plusieurs plates-formes et outils existants sont testés, tels que Philologic[6], TXM[7] et CQPWeb[8]. Chaque texte est préparé sous deux formats : le fichier XML conforme au standard TEI, et le simple fichier TXT. Les deux types seront téléchargeables librement et gratuitement.

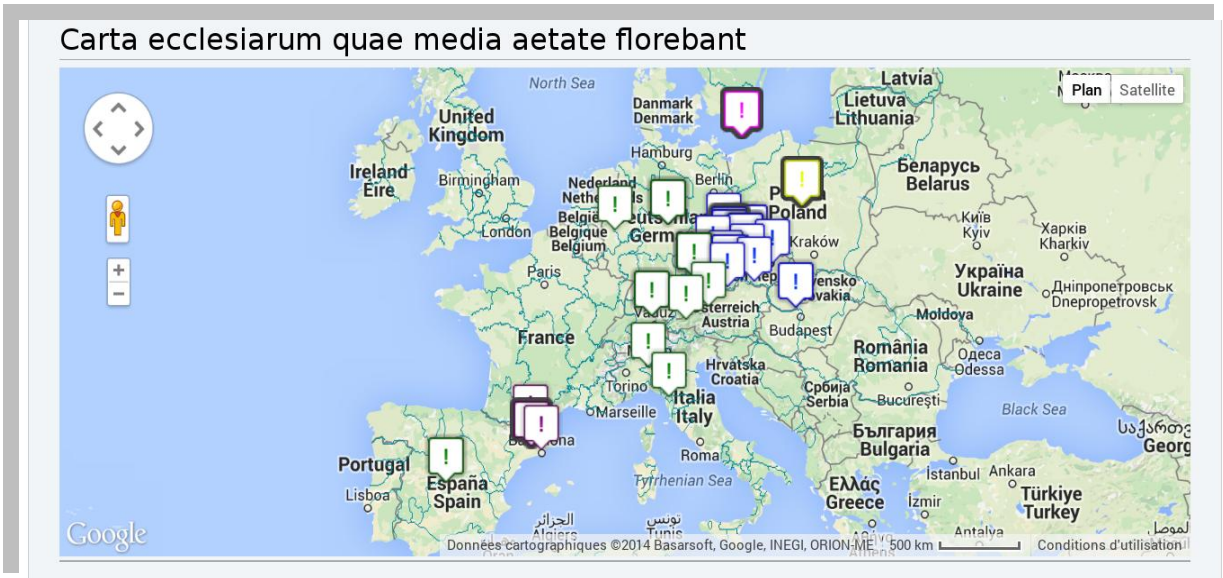
IV. Wiki Lexicographica[9]

Enfin le Laboratoire participe à plusieurs programmes européens de recherche et de collaboration. Dans le cadre du projet COST « Medioevo Europeo », Krzysztof Nowak et Bruno Bon ont développé, à partir de quelques dictionnaires de latin médiéval, le prototype d'encyclopédie interactive « WikiLexicographica », fondé sur le logiciel MediaWiki, utilisé par Wikipedia. Deux types (catégories) de pages sont prévus, selon qu'elles correspondent à une référence (auteur, texte) ou à un article (lemme). Les pages d'auteurs et d'œuvres sont issues de l'index des sources de chaque dictionnaire. Aux renseignements fournis par l'édition imprimée, ont été ajoutées trois indications

[5] http://scriptores.pl/efontes.
[6] http://artfl-project.uchicago.edu/.
[7] http://txm.sourceforge.net/.
[8] https://cqpweb.lancs.ac.uk/.
[9] http://scriptores.pl/wiki.

nécessaires à l'évolution des types d'interrogation (date normalisée, type de texte, lieu de production). Les pages de lemmes sont également importées des dictionnaires, par l'intermédiaire d'un filtre de transformation de XML vers SMW. Outre une présentation classique, reproduisant la structure 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, 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.

Bien entendu, il est possible de choisir le dictionnaire à interroger, et d'en consulter la liste des articles ou l'index des sources. On peut également passer 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. Outre ces procédures classiques, le prototype permet une sorte de « recherche visuelle » : la frise chronologique et la carte des citations offrent la possibilité de sélectionner les occurrences en fonction de leur localisation, et d'en visualiser le détail. Enfin, dans le cadre d'une recherche avancée, l'utilisation de « Semantic Drilldown » permet de filtrer 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 présentés sous les différents formats évoqués plus haut (tableau, chronologie, carte).



Le texte comprend aussi les extraits des publications suivants :

1 Krzysztof NOWAK, « The eLexicon Mediae et Infimae Latinitatis Polonorum. The Electronic Dictionary of Polish Medieval Latin », dans *Proceedings of the XVI EURALEX International Congress: The User in Focus (15-19 July 2014, Bolzano/Bozen),* dir. Andrea ABEL [et al.], Bolzano/Bozen, EURAC research, 2014, p. 793–806[10].

2 Bruno BON, Krzysztof NOWAK, « Wiki Lexicographica. Linking Medieval Latin Dictionaries with Semantic MediaWiki ». dans *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia,* dir. Iztok KOSEM [et al.], Tallinn – Ljubljana, Trojina, Institute for Applied Slovene Studies; Eesti Keele Instituut, 2013, p. 407–420[11].

3 Bruno BON, Krzysztof NOWAK, « Pour une encyclopédie interactive du latin médiéval : le Semantic Web au service de la lexicographie médiolatine », *Archivum Latinitatis Medii Aevi*, 70, 2012, p. 355–359.

Krzysztof Nowak

Les deux auteurs tiennent à remercier M. Bruno Bon (Comité du Cange, IRHT-CNRS) pour la relecture et la correction de cet article.

[10] Voir http://euralex2014.eurac.edu/en/callforpapers/Documents/EURALEX%202014_gesamt.pdf.

[11] Voir http://eki.ee/elex2013/proceedings/eLex2013_28_Bon+Nowak.pdf.