POURQUOI INFORMATISER
UN VIEUX GLOSSAIRE ?

PRÉSENTATION DU *DU CANGE* EN LIGNE [1]

Résumé : Les dix volumes du Du Cange, *glossaire du latin du Moyen Âge commencé au XVII^e siècle, viennent d'être informatisés à l'initiative de l'École des chartes qui l'offre gratuitement en ligne, téléchargeable. Approfondir la connaissance de ce glossaire ancien, toujours très utilement consulté, établir une édition électronique aux avantages décisifs par rapport au dictionnaire papier, constituer un corpus structuré pour des recherches linguistiques sur le latin médiéval, tels sont les objectifs du projet mis en œuvre. L'information permet ainsi de mieux mesurer la part de chaque auteur dans sa réédition augmentée jusqu'à la fin du XIX^e siècle, de même que le corpus travaillé pour des exploitations informatiques offre de multiples possibilités au linguiste, par exemple, à terme, la possibilité d'un lemmatiseur du latin médiéval. De fait, tout chercheur, linguiste, historien, diplomatiste, pédagogue, etc., peut mener ses expériences et contribuer à l'enrichissement de ce corpus devenu fécond pour la recherche.*

Le grand *Oxford*, le grand *Robert* et le *TLF*, le *Littré*, le *Godefroy*, le *Lacurne*, la plupart des grands dictionnaires encore consultés ont été informatisés. Manquaient les 10 volumes du *Du Cange*, ou *Glossarium mediæ et infimæ latinitatis*, un glossaire de la « basse latinité », autrement dit, le latin du Moyen Âge.

Commencé au XVII^e siècle, ce dictionnaire ancien n'est pas encore remplacé, il est utilisé par les médiévistes et les philologues. À la différence des projets précédents, le texte est non seulement gratuit en ligne, mais aussi téléchargeable dans sa version source, comme un logiciel libre. L'initiative est menée dans le cadre d'un programme de recherche initié par l'École nationale des chartes, il a bénéficié du financement du TGE Adonis en 2008 et, depuis décembre 2008, de l'Agence nationale de la recherche (programme OMNIA, en collaboration avec l'Institut de recherche et d'histoire des textes et le laboratoire ARTeHIS). Toutes les lettres sont désormais en ligne depuis décembre 2010 <http://ducange.enc.sorbonne.fr/>, dans un délai qui doit beaucoup aux expériences précédentes.

1. 2010, École des chartes, http://ducange.enc.sorbonne.fr/

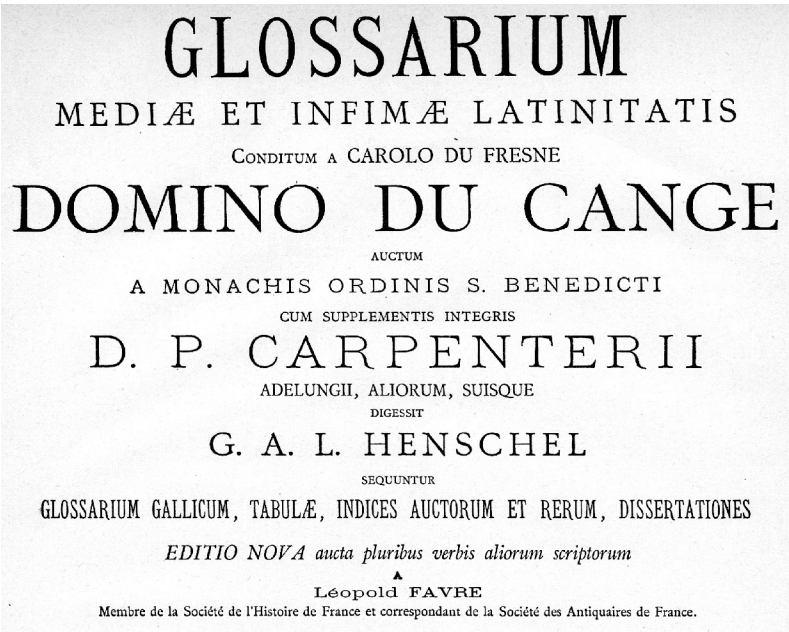
Détailler les techniques d'informatisation n'est plus indispensable, elles sont maintenant bien décrites, cependant savoir comment faire ne dispense pas de se demander pourquoi, et jusqu'où informatiser. La technique a été subordonnée à trois objectifs : approfondir la connaissance de l'ouvrage ; établir une édition électronique avec des avantages décisifs sur les volumes papier ; constituer un corpus structuré pour des recherches linguistiques sur le latin médiéval.

1 PRÉSENTATION ET HISTORIQUE DU *DU CANGE*

> Ducange, était d'une taille un peu au-dessous de la moyenne ; il avait la tête bien proportionnée, les yeux charmants et pleins de feu, une belle figure, les traits distingués et l'air noble. S'il ne jouissait pas d'une extrême opulence, il possédait néanmoins une fortune honnête, et n'en désira jamais une plus grande, répétant qu'un homme de lettres devait se contenter d'une aisance qui lui permit de satisfaire son goût pour les livres. D'une humeur égale, jamais incommode, ne fatiguant personne, se prêtant sans réserve à ceux qui imploraient son appui, communiquant avec facilité les fruits de ses études, il était plus enclin à mériter les récompenses qu'à les solliciter.
>
> BALUZÉ, 1689, traduit du latin, dans Du Cange, *Chronicon paschale*.

Le *Du Cange* est un glossaire portant sur le latin médiéval, en latin moderne. Il s'agit d'un glossaire, et non d'un dictionnaire, les mots courants ne sont pas traités, les articles ne comportent pas d'informations grammaticales, mais des gloses et des citations. Le latin médiéval, écrit dans toute l'Europe pendant un millénaire, est une langue à part entière qui se distingue du latin classique, notamment par son lexique. Enfin, initié au XVII^e^ siècle par du Cange, les gloses sont rédigées en latin (moderne), considéré à l'époque comme une langue savante, durable et internationale.
L'ouvrage fut réédité et augmenté jusqu'à la fin du XIX^e^ siècle. Si l'ensemble se présente comme un tout digne de confiance, il résulte cependant de différents apports qui ne sont pas anonymes. Chaque auteur à un style, des intérêts, une connaissance plus particulière de certains fonds, l'informatisation permet de préciser des impressions, et d'y mesurer la part de chaque auteur.

- 1.1 1678, Charles du Fresne, sieur du Cange (1610-1688), 3 volumes.
- 1.2 1733-1736, bénédictins de la congrégation de Saint-Maur, édition augmentée, 6 volumes.
- 1.3 1766, Pierre Carpentier (1697-1767), 4 volumes de supplément.
- 1.4 1840-1850, Louis Henschel, fusion des éditions précédentes chez Didot, 8 volumes.
- 1.5 1883-1887, Léopold Favre, réédition légèrement augmentée de Henschel, 10 volumes.


GLOSSARIUM
MEDIÆ ET INFIMÆ LATINITATIS
CONDITUM A CAROLO DU FRESNE
DOMINO DU CANGE
AUCTUM
A MONACHIS ORDINIS S. BENEDICTI
CUM SUPPLEMENTIS INTEGRIS
D. P. CARPENTERII
ADELUNGII, ALIORUM, SUISQUE
DIGESSIT
G. A. L. HENSCHEL
SEQUUNTUR
GLOSSARIUM GALLICUM, TABULÆ, INDICES AUCTORUM ET RERUM, DISSERTATIONES
EDITIO NOVA aucta pluribus verbis aliorum scriptorum
Léopold FAVRE
Membre de la Société de l'Histoire de France et correspondant de la Société des Antiquaires de France.

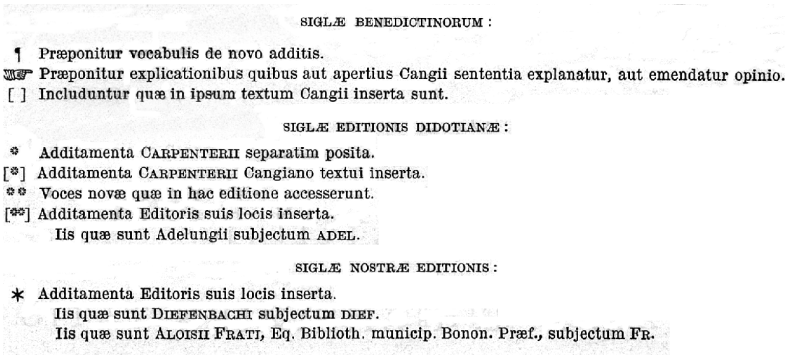
Page de titre du Glossarium dans l'édition de Favre, 1883-1887

La structure du glossaire doit beaucoup à son premier auteur, quelques éléments de sa biographie en éclairent certains aspects. Charles du Fresne, sieur du Cange, est né à Amiens en 1610. À neuf ans il entre au collège des jésuites de sa ville. Il continue des études juridiques à Orléans, on le trouve inscrit comme avocat au barreau de Paris en 1631. Un registre atteste de son mariage à Amiens en 1638, où il reprend la charge notariale de son oncle. Rien d'étonnant à ce que le *Glossarium* s'intéresse particulièrement au vocabulaire juridique. En 1669, Du Cange quitte Amiens avec sa famille, fuyant une peste, et s'installe à Paris où il mourra vingt ans plus tard, en 1688. Il commence par publier une *Histoire de l'empire de Constantinople* (1657), c'est-à-dire de l'« empire latin » (1204-1261) créé après la quatrième croisade. Cette première recherche permet de mieux comprendre son intérêt pour le grec byzantin, qu'il semble avoir étudié plus que le latin. Il continue avec un *Traité historique du chef de saint Jean-Baptiste* (1665), afin d'authentifier la relique revendiquée par la cathédrale d'Amiens, un crâne rapporté en 1206 de la croisade. Son édition de l'*Histoire de saint Louis, écrite par le sire de Joinville* (1669) est un assemblage curieux, avec de longs développements sur les couleurs héraldiques, les cris d'armes ou la hiérarchie des chevaliers, montrant une curiosité pour tous les détails. Le *Glossarium* comporte beaucoup de dissertations de ce type, par exemple sous l'entrée MONETA, où il énumère de nombreuses monnaies ayant eu cours, avec leurs équivalences et leurs poids. Il publie ensuite plusieurs monographies sur des historiens français, qui lui attirent la protection de Colbert. Le ministre pense lui confier la direction d'une collection sur tous les historiens et chroniqueurs de la France,

mais le gouvernement refuse le projet, auquel il reproche l'absence de plan. Libéré, il se consacre alors au *Glossarium ad scriptores mediae et infimae latinitatis*, trois volumes in quarto (1678). Outre quelques éditions de textes grecs et latins, il donne une *Histoire byzantine* (1680), et le *Glossarium ad scriptores mediæ et infimæ græcitatis* (1688), glosant en latin des citations en grec. Ses archives personnelles témoignent d'une activité érudite infatigable. Il commença une histoire de la Gaule, avec des renvois précis aux éditions des antiques, usant de la même rigueur bibliographique que celle mise en œuvre pour les citations de son glossaire. Il projeta un grand nobiliaire de France, une histoire de la Picardie et des ses familles nobles – sa carrière d'historien commença comme généalogiste. Du Cange n'est pas un grammairien, il s'intéresse aux mots pour les choses passées dont ils sont la trace, il n'a pas de théories ou de plans préconçus et s'épanouit dans l'ordre alphabétique.

Après des éditions allemandes, par exemple celle de 1710[2] à Francfort-sur-le-Main, l'ouvrage est repris par les bénédictins de Saint-Maur. Le *Glossarium* entre dans le projet savant des mauristes ; il est repris par deux moines, Maur Dantine (1688-1746), qui contribuera à l'*Art de vérifier les dates* (1750), et Pierre Carpentier (1697-1767). La collaboration aboutit en 1733 à une édition augmentée du glossaire de du Cange. Carpentier se fâcha avec les mauristes, et continuera le projet seul, publiant un supplément en 1766. À la sortie du premier volume de l'édition Didot en 1840, un érudit détaille cette histoire, et commence à compter les signes inscrits par l'éditeur qui attribuent alinéas et segments de texte à chaque auteur.

> […] des signes différents indiquent avec une telle précision ce qui appartient à chaque auteur, qu'au premier coup d'œil on distingue sans confusion le travail primitif, les additions des Bénédictins, les suppléments de Carpentier et les compléments dus aux recherches des nouveaux éditeurs. […] ce signe [le ** de Henschel] n'est pas répété moins de cinq cent dix-huit fois dans les cent soixante pages qui composent la première livraison.
>
> GÉRAUD, H. 1840.
> « Historique du *Glossaire de la basse latinité* de Du Cange »



SIGLÆ BENEDICTINORUM :

¶ Præponitur vocabulis de novo additis.
☞ Præponitur explicationibus quibus aut apertius Cangii sententia explanatur, aut emendatur opinio.
[ ] Includuntur quæ in ipsum textum Cangii inserta sunt.

SIGLÆ EDITIONIS DIDOTIANÆ :

* Additamenta CARPENTERII separatim posita.
[*] Additamenta CARPENTERII Cangiano textui inserta.
** Voces novæ quæ in hac editione accesserunt.
[**] Additamenta Editoris suis locis inserta.
Iis quæ sunt Adelungii subjectum ADEL.

SIGLÆ NOSTRÆ EDITIONIS :

∗ Additamenta Editoris suis locis inserta.
Iis quæ sunt DIEFENBACHI subjectum DIEF.
Iis quæ sunt ALOISII FRATI, Eq. Biblioth. municip. Bonon. Præf., subjectum FR.

Légende des sigles d'attributions du Glossarium dans l'édition de Favre (1883-87)

2. http://www.uni-mannheim.de/mateo/camenaref/ducange.html

	Nombre d'articles	Texte en millions de signes	Part des citations dans le texte	Citations, nombre par langue : latin, français, grec
Glossarium	90 388	44,7 Ms, 5004 p.	47 %	126 440, 22 518, 13 673
du Cange	22 980	15,4 Ms	48 %	48 215, 4710, 8 178
mauristes	34 875	16,9 Ms	38 %	42 526, 2 241, 4 850
Carpentier	26 411	14,0 Ms	49 %	31 418, 15 471, 644
Henschel	1 069	1,7 Ms	17 %	2 438, 68, 1
Favre	5 053	0,6 Ms	28 %	1 843, 28, 0

Répartition des citations du Glossarium *selon les auteurs.*

Cent vingt ans plus tard, un texte informatisé permet de mesurer la répartition des caractères selon les champs d'information, ainsi que selon les auteurs[3]. La première caractéristique du *Glossarium* concerne les citations. La moitié du texte est constitué de sources manuscrites référencées. Il s'agit généralement de documents d'archives, parfois de textes littéraires, beaucoup plus rarement de théologie. La reconnaissance des références n'est pas encore possible et demandera un gros effort de reprise manuelle. Les chiffres précisent aussi la répartition des contributions en trois gros tiers : la matrice initiale de du Cange, la patiente broderie des bénédictins, et le supplément de Carpentier. Chaque auteur ne cite pas les mêmes sources, ainsi du Cange propose plus d'extraits en grec, tandis que Carpentier insère plus d'ancien français. Les éditeurs du XIX^e^ siècle ont montré une patience tenace et scrupuleuse, ils apportent cependant peu de texte.

Cette édition électronique continue cette histoire, avec d'autres technologies, dans l'intention de commencer l'édition de référence de l'ouvrage pour ce siècle. Est-ce que la forme électronique promet plus de pérennité ? Le recul a désormais montré que les supports informatiques sont beaucoup moins durables que le papier (qui l'est d'ailleurs moins que le parchemin ou l'argile cuite). De plus, contrairement à un livre qui peut être oublié un demi-siècle sur une étagère, un fichier informatique a besoin d'un logiciel pour fonctionner. L'adoption de standards indépendants et documentés (XML[4]-TEI[5]) évite que l'information dépende d'un produit commercial, mais une organisation compétente reste cependant nécessaire pour garder une édition électronique

3. Le nombre de caractères attribué à un auteur tient compte du triple niveau de structure (articles, paragraphes, insertions) évitant par exemple d'ajouter pour du Cange un segment de paragraphe attribué aux bénédictins. Des imprécisions sur le nombre de signes sont possibles, suite à des variations techniques dans le comptage des espaces, sans affecter cependant les équilibres globaux.

4. eXtensible Markup Language, http://www.w3.org/XML/

5. Text Encoding Initiative, http://www.tei-c.org/release/doc/tei-p5-doc/fr/html/index.html

disponible en ligne. En dernier ressort, la meilleure protection pour le *Du Cange* électronique est d'autoriser la copie (à des fins non commerciales). Ainsi tout chercheur ou institution intéressée peut travailler avec les fichiers sources, assurant que le projet puisse continuer ailleurs quoi qu'il arrive. L'ATILF a par exemple manifesté son intérêt pour le *Du Cange*, afin d'ajouter les entrées en français à son portail lexical[6].
L'informatisation a permis de prouver que le *Glossarium* est surtout un monument d'Ancien Régime, une sorte de Versailles poursuivi sous Louis XIV et Louis XV. La révolution industrielle en a assuré la conservation et la diffusion par plusieurs éditions imprimées, mais ajouta peu de matériaux dans le plan initial. Notre siècle n'apportera probablement plus de texte, en revanche sa structuration progressera, nous entrons dans l'ère de l'édition continuée.

2 ACCÈS AUX ARTICLES ET AUX CITATIONS

> Le côté vert, c'est le « côté du cirque », la « saturnale permanente » (!). Et, de fait, on ne drague pas à gauche (gris) mais à droite (vert). Dans ce but, les meilleures places sont les places périphériques […] en faisant semblant de consulter le dictionnaire de Du Cange, ou celui de Godefroy, on dispose d'un observatoire privilégié pour regarder, prendre son temps, choisir, se préparer.
>
> PASTOUREAU, 1984,
> « Côté vert, côté gris. S'asseoir à la Bibliothèque nationale, travailler, lire, écrire, dormir, rêver, se souvenir d'avoir aimé »,
> *Médiévales*, n° 7, pp. 106-111.

L'informatisation d'un ouvrage peut s'effectuer selon plusieurs degrés de précision : numérisation en image (fac-similés), images indexées, recherche plein texte, typographie affichable, balisage des fonctions. Chaque approfondissement a un coût, l'informatisation du *Du Cange* encode la typographie, ainsi qu'une partie des fonctions lexicographiques. Était-ce nécessaire pour satisfaire le public ?
La numérisation en images suffit à dématérialiser un livre et à l'offrir en ligne ou au téléchargement. Plusieurs éditions électroniques du *Glossarium* étaient déjà disponibles.

1.6 Standford[7], gratuit, fac-similés avec recherche plein texte, sous forme de fichiers PDF téléchargeables, éditions de Favre (1883-1887) et de Didot (1840-1850).
1.7 Université de Mannheim[8], gratuit, beau fac-similé de l'édition de 1710 (le seul texte de Du Cange).

6. http://www.cnrtl.fr/definition/
7. http://standish.stanford.edu/bin/search/advanced/process?sort=title&browse=1&clauseMapped%28creatorBrowse%29=Du+Cange%2C+Charles+Du+Fresne%2C+sieur%2C+1610-1688
8. http://www.uni-mannheim.de/mateo/camenaref/ducange.html

1.8 Gallica[9], gratuit, fac-similé d'une réimpression (1937) de Favre.
1.9 Google Books[10], la plupart des éditions sont disponibles en images, mais les exemplaires sont souvent dépareillés.
1.10 Brepols[11], payant, fac-similé de l'édition de Favre (1883-1887) avec indexation des entrées.

Si une numérisation en image peut suffire à la lecture suivie d'un livre, pour un ouvrage de référence, le feuilletage est beaucoup plus fastidieux, car si le poids du papier n'est plus entre les mains, il reste sous forme électronique. Une image de page du *Glossarium* pèse de 3 à 600 Ko, 30 à 60 fois plus lourd que sa version texte brut. L'ouvrage complet pèse 600 Mo en images de basse qualité (fax, tiff4) ; et 5 Go dans sa version en ligne (jpg en niveaux gris). Le défilement des images pour retrouver un mot est beaucoup plus fastidieux que de tourner des pages de papier.

Le texte de Stanford et de Google, obtenu par reconnaissance automatique des caractères (océrisation), s'est avéré inexploitable. D'une part le texte brut a perdu toute mise en forme typographique (notamment l'italique des citations), d'autre part il aurait fallu corriger une faute toutes les deux ou trois lignes. À partir des seules images, une astuce aurait suffi à améliorer le confort, le « dictionnaire sporadique »[12]. En indiquant pour chaque page son premier mot, il est possible d'offrir une consultation avec très peu d'efforts d'indexation. Soit par exemple dans le tome 3 la page 537, qui commence par FOESA, et la page 538, démarrant à FOINESUM. Dans l'ordre alphabétique, FOILLIATA est entre les deux, donc sur la page 537. Par une simple recherche d'intervalle, il est très simple d'afficher la bonne page pour FOGIA, FŒTA, ou même une suite de lettres inconnues comme FOH. Encore faut-il que l'ordre alphabétique soit pertinent pour le latin médiéval.

Comme les autres langues médiévales, la graphie du latin médiéval varie selon les lieux et siècles, un même mot accepte de nombreuses variantes. La lexicographie de l'ancien français connaît ainsi quatre nomenclatures de référence : le *Godefroy*, le *FEW*, le *Tobler-Lommatzsch* et le *DMF*, avec une doctrine et une justification philologique allant en progressant avec le temps. Le *Du Cange* est bien plus confus.

Considérons par exemple l'entrée HOSTIARIUS 2. Les mauristes rapportent *hostiarius* au sens de « portier, huissier » et renvoient à l'article OSTIARIUS, écrit par du Cange, avec cette citation : « Estoit Huissiers et Chambrelens […] Fu à la porte pour ouvrir ». Du Cange ne renvoie pas bien évidemment à HOSTIARIUS qui fut écrit un demi-siècle plus tard, et aucun des rédacteurs postérieurs n'a ajouté ce renvoi nécessaire. L'élision du H ini-

9. http://gallica.bnf.fr/ark:/12148/bpt6k1175756.r=.langFR
10. http://books.google.com/
11. http://clt.brepolis.net/dld/
12. Nous reprenons le concept établit par Serge Heiden (ENS-LSH) afin de désigner les méthodes et outils de statistiques textuelles, à la fois sur le lexique comme sur les composants du document.

tial n'est pas un phénomène négligeable, sur 1644 vedettes dédoublonnées, un algorithme repère 335 équivalents commençant par toutes les voyelles, et même quelques cas en R et L, comme « HROCCUS […] Vide Roccus. ». Dans le *Mediae Latinitatis lexicon minus* (NIERMEYER, Leiden : E.J. Brill, 1976), *ostiarius* est la forme de référence pour les sens de type porte, huissier : « ostiarius, hostiarius : 1. *portier (dernier des ordres mineurs) [...] 2. portier, officier de ménage [...] 3. huissier, dignitaire antique [...] », à ne pas confondre avec les mots dérivés d'*hostie* (HOSTIARIUS 1). La recherche d'un mot médiéval mène à parcourir de nombreux tomes, éparpillés par le désordre alphabétique.

Le rapport à ce glossaire n'est pas celui du dictionnaire. L'historien n'y cherche pas une autorité qui définit, mais une citation, un indice, éclairant le mot qui l'arrête dans sa compréhension d'un document. Une grande table avec plusieurs volumes ouverts contient plus d'information qu'une image fixe sur un écran. Afin de convaincre le chercheur de passer à l'écran, l'informatisation doit ouvrir de nouvelles fonctionnalités de navigation et de recherche pour remplacer ce qui est perdu avec le papier.

À partir des images, deux voies peuvent conduire au texte : la saisie, ou la reconnaissance automatique des caractères. Les prestataires de numérisation ont organisé des ateliers de saisie à l'étranger à partir des années 1990, détectant les inévitables erreurs humaines, en doublant ou triplant les opérateurs. Les logiciels de reconnaissance de caractères (OCR) continuent à progresser. Sur un imprimé de bonne qualité et de typographie récente (à partir du XIX^{e} s.), avec un calibrage spécifique de la luminosité et un entraînement de la reconnaissance graphique, l'OCR peut permettre d'établir un texte acceptable, et surtout, de retrouver une part importante de la typographie, par exemple les petites capitales, ou des lettres en exposant, et généralement l'italique.

Cependant la typographie ne suffit pas à informatiser la structure d'un dictionnaire. Ainsi les renvois sont en italiques, mais aussi les citations, les mots en langues vernaculaires, certains titres d'ouvrages, les vedettes de troisième niveau… À la même convention typographique correspond beaucoup de fonctions lexicographiques, difficiles à distinguer sans reprise humaine. En 2010, le *Du Cange* informatisé comporte environ 90 000 entrées, 35 000 sous-vedettes, 50 000 renvois vérifiés, 150 000 citations encore à vérifier, et 150 000 segments italiques non définis. Cette structuration assure déjà une navigation commode, avec notamment une recherche à la fois sur les vedettes, les sous-vedettes et les renvois. Chercher *hostiarius* permet ainsi de trouver HOSTIARIUS 1, 2, 3, 4, et OSTIARIUS, mais aussi HUISSERIUS Armorum, ainsi que JANITORES ou ÆDITUUS qui renvoient à *Ostiarius*. S'ajoute enfin la recherche sur la totalité du texte, ou seulement des citations. La fluidité électronique permet de poser de nouvelles questions au *Du Cange*. Un diplomatiste a cherché des registres disparus depuis du Cange, un médiéviste fut étonné d'y trouver certains *exempla*, une philologue a constaté l'importance des références aux glossaires d'Isidore ou de Papias. Chaque besoin particulier peut amener un perfectionnement à l'interface.

3 INTÉRÊTS POUR LE LINGUISTE

> J'avais, il est vrai, en lexicographie, d'illustres prédécesseurs, Henri Estienne, Du Cange, Forcellini ; Du Cange surtout, que j'ai feuilleté sans relâche et pour qui je suis reconnaissant comme s'il était là me prêtant l'oreille. Je n'ai pas l'outrecuidance de me comparer à eux. Leur tâche, d'ailleurs, n'a pas été la même que la mienne ; car ils se sont occupés de langues mortes où tout est clos, et moi, j'ai eu affaire à une langue vivante où tout demeure ouvert.
>
> Littré, 1880, *Comment j'ai fait mon dictionnaire*.

Au-delà des historiens médiévistes, le *Du Cange* est souvent estimé, mais pas toujours pratiqué. Le latin n'aide pas à attirer un public plus large, d'autant qu'il ne s'agit pas de la langue de Cicéron, ni même de Thomas d'Aquin. Du Cange témoigne plutôt d'un mouvement européen de prise de conscience linguistique, remarquable dans sa curiosité pour l'étymologie : les langues romanes dérivent en partie d'un état non écrit du latin. Mieux, le *Glossarium* montre que le latin "vient" parfois des langues européennes : PIZZA « ex Italico », HUNDREDUS « voce Saxonica *hundreð*, *centum* » ou BURGWARDUS « ex Teutonico *Burg*, *Burgum*, et *Ward*, custodia ». Le vieux glossaire conserve une mémoire philologique encore reprise dans les dictionnaires actuels, l'informatisation en facilitera l'exploitation linguistique.

Le *Godefroy* (1863-1872) propose ainsi cette citation : « Et il soit ainsi, que ledit Pierre depuis un an en ça par impatience, fragilité ou diminution de son corps et de sa sensualité, soit devenu tout ydiote, etc » (article SENSUALITÉ). Le moderne est entraîné vers un contresens plaisant : comment la diminution de la sensualité rend-elle idiot ? Godefroy propose la définition : « capacité de sentir », disons *sensibilité*, qui laisse encore la phrase en contradiction. L'*impatient*, selon Godefroy lui-même, n'a pas seulement peine à attendre, mais plus généralement, à souffrir, à supporter. Si la *sensibilité* de Pierre diminue, alors ne devrait-il pas être moins *impatient* ?

Littré avait déjà retenu cette exacte citation, dans la section historique de l'article SENSUALITÉ (1863-1872). Il cite le *Du Cange* 3 173 fois, en général pour la section historique d'un mot, avec des citations en ancien français de Carpentier. Mais Littré (voir plus haut en épigraphe) reconnaît une dette plus profonde à la lexicographie de langues anciennes en général, et à Du Cange en particulier. Le lexicographe médecin sortait d'une édition d'Hippocrate lorsqu'il commença son dictionnaire. Firmin Didot rééditait le *Thesaurus graecae linguae* d'Henri Estienne (1572, 1831-1865, 6 volumes), qui cite les classiques avec précision, notamment pour Platon, dont il a établi la pagination de référence (*Stephanus*). Comme du Cange, Littré souhaita cette rigueur pour son ouvrage. La généalogie des dictionnaires à extraits cités part ainsi du grec, vers le latin et le français, pour se continuer en anglais dans le grand Oxford. Quant audit Pierre, Littré le place à côté d'un extrait d'Oresme : « Contrarieté en l'ame entre raison et la sensualité » (*Le livre des Ethiques d'Aristote*, 1370) et de Brunetto Latini : « Deliz sans respit n'est mie bons, porce que il est de natural sensualité, qui est commune as betes » (*Li livres dou tresor*, livre XXXVII. — De Delit [du délice, des plaisirs], 1254). Ces deux auteurs ont en commun d'avoir transposé l'aristotélisme

en ancien français, où le mot *sensualité* prend son sens en opposition à la raison, comme désormais pour nous. Cette hypothèse ne semble pas élucider l'emploi pour ledit Pierre.

La Curne de Sainte-Palaye (1697-1781) reprend toujours la même citation, en distinguant cette fois deux acceptions : « 1° Plaisirs sensuels […] 2° Sens, intelligence » (Lacurne, *Dictionnaire historique de l'ancien langage françois*). Si le bon sens de Pierre diminue, alors c'est une tautologie qu'il devienne idiot, la définition est beaucoup plus convaincante.

Cependant la première occurrence lexicographique et imprimée de cette phrase est dans le *Glossarium*, sous la plume de Carpentier (1766), à l'article « 3. SENSUALITAS, Gall. *Sensualité*. Sensus, intellectus. Lit. remiss. ann. 1376 in Reg. 110. Chartoph. reg. ch. 208 : *Et il soit ains, que ledit Pierre* […] Vide supra *Sensibilis* 2. et *Sensus* 1. » Les renvois précisent sans équivoque l'acception, 2. SENSIBILIS : « Laquelle Coline n'estoit pas bien Sensible, ne savoit pas bien faire ses besongnes. », « 1. SENSUS, Intellectus, νους, nostris Sens, bon sens ». *Sensualité* est rattaché à la signification de *sensible* encore courante en anglais : « avisé, judicieux, raisonnable ». Carpentier ajoute une autre information qui aide à interpréter le contexte, la référence au manuscrit, il s'agit d'une "charte" royale (*chartophylacis regis charta*), dans un registre de 1376 qui porte encore la même cote aux Archives nationales dans la série JJ, c'est une lettre de rémission (*lit. remiss.*). Par ce type d'acte est fréquent au XIV^e^ s., le roi accorde un pardon qui peut interrompre le cours de toute action de justice locale, accroissant ainsi son pouvoir par la miséricorde.

Les dictionnaires actuels ne semblent pas avoir retenu cette acception et cette citation. Le *Niermeyer* propose les traductions suivantes : « sensualitas : 1. faculté de sentir, sensibilité — ability to feel, sensitiveness. 2. esprit matérialiste — materialistic state of mind. ». Le *Dictionnaire du moyen-français* (ATILF) n'a pas encore rencontré cet emploi dans les dépouillements en cours, mais il enregistre bien les usages de *sensible* équivalent d'*intelligent*.

Sans promettre que la recherche dans le *Du Cange* soit toujours aussi fructueuse, l'édition électronique du *Glossarium* devrait faciliter sa consultation par un public moins familier du latin, en proposant un accès par un glossaire d'ancien français. Carpentier a en effet réuni une nomenclature, complétée par Henschel, d'environ 22 000 vedettes, renvoyant aux citations en ancien français dans les articles latins. Cela ne constitue pas un dictionnaire complet d'ancien français, mais ce fut le premier publié.

Cependant les trésors du *Glossarium* demeurent surtout entreposés en latin. Un autre exemple permettra d'apprécier le jugement de du Cange sur l'étymologie. Pour le mot *dague*, il propose une origine celte, ou plus exactement dans ses termes : *Cambro-Britannis* (Cambrie, le pays de Galles). Furetière (1690) résume ainsi les doutes de son temps : « Ce mot, selon Ménage, vient de l'Allemand *dagge & dagger*, qui signifient la même chose. » Cette phrase est textuellement dans Ménage (1650, 1670), pour la *dague* comme « petite épée », mais plus haut, il considère que le mot vient d'Écosse dans le sens « pointe de fer garnissant une hache d'armes ». Furetière continue à énumérer : « Du Cange dit que ce mot vient du Bas-Breton », « D'autres le dérivent à *Dacia*, parce que c'étoit leur arme ordinaire ; d'autres de l'Hébreu

dacack ». Littré (1863-1872) ajoute : « La forme portugaise a-daga pourrait indiquer une origine arabe. » Le *TLF* (1971-1994) contient 1313 occurrences de *Du Cange*, généralement dans la section origines, par exemple ici pour *dague*. Les hésitations avouées qui font honneur à ce dictionnaire n'envisagent plus les hypothèses de l'arabe et de l'hébreu, mais semblent donner raison à Du Cange contre le *FEW*.

> Orig. obscure. À l'hyp. d'un empr. à un lat. vulg. **daca,* fém. substantivé tiré de *daca spatha* « épée dace », par l'intermédiaire de l'ital. ou du prov. *daga* (*FEW* t. 3, p. 2[a] ; BL.-W.[1-5]), s'oppose le fait que l'ital. et le prov. ne sont pas attesté av. le XIV[e] s. (v. LÉVY *Prov.* et BATT.) ; d'autre part l'expr. *daca spatha* (ou *ensis*) n'est pas attestée en lat. (v. *TLL*). L'ancienneté du mot sur le territoire anglais (*cf. dagger, ca* 1200, *Statuts de Guillaume roi d'Écosse* ds **DU CANGE** ; lat. médiév. *daca* chez le grammairien angl. J. de Garlande, XIII[e] s., cité par A. Scheler ds *Jahrbuch für romanische und englische Literatur,* t. 6, pp. 153-154) conduit à chercher, avec COR., *s.v. daga* I, un étymon celt., mais que l'on ne peut identifier avec certitude (v. aussi KLUGE, *s.v. Degen*[2]).
>
> *TLF*, « DAGUE »

L'informatisation du vieux glossaire progresse pour rendre ce trésor disponible au linguiste, non pas seulement comme une édition, pour la lecture et la navigation, mais aussi comme un corpus de données structurées.

La nomenclature tout d'abord, est actuellement examinée par les lexicographes du projet OMNIA, afin d'établir une base lexicale du latin médiéval. Cette langue, contrairement au grec et au latin classique, ou aux langues modernes, ne dispose pas encore d'une ressource libre établissant la hiérarchie entre lemmes, variantes, et formes fléchies. Cette information est attendue pour plusieurs applications. D'abord, elle facilitera la consultation du *Glossarium*, en regroupant les articles selon des critères philologiques validés. Ensuite, elle permettra de lier n'importe quel mot d'un texte aux articles du dictionnaire. Un tel lexique peut aussi servir dans un moteur de recherche plein texte, afin de trouver toutes les occurrences d'un lemme, abstraction faite de la flexion et des variantes. Enfin, cette base permet de lemmatiser les textes, avec l'assistance d'un étiqueteur qui lève l'ambiguïté entre les formes par la nature morphosyntaxique d'un mot.

La lemmatisation est une étape nécessaire avant d'élaborer des méthodes et des outils de traitement plus avancé de la langue des documents : alignement de séquences pour la reconnaissance de formules et de citations ; catégoriseurs automatiques pour datation, localisation, attribution, ou typologie de documents. La numérisation du *Du Cange* s'inscrit dans le projet d'introduire la « textométrie »[13] comme une science auxiliaire de l'histoire, complémentaire de la diplomatique et de la philologie, afin d'assister l'édition et l'analyse de sources historiques médiévales.

13. Nous reprenons le concept établi par Serge Heiden (ENS-LSH) afin de désigner les méthodes et outils de statistiques textuelles, à la fois sur le lexique comme sur les composants du document.

CONCLUSION

L'informatisation du *Du Cange* a d'abord confirmé l'intuition des connaisseurs, c'est un assemblage de plusieurs auteurs, avec des approches différentes, l'attribution d'un article ou d'un alinéa peut importer à la compréhension. Le texte, les champs balisés et l'interface de consultation sont en perfectionnement continu pour faciliter la navigation dans une matière parfois éparpillée. L'outil se veut simple pour le novice, et efficace pour le chercheur assidu. Le corpus, enfin, est travaillé pour des exploitations informatiques, notamment vers un lemmatiseur du latin médiéval. Tout chercheur peut mener ses expériences et contribuer à l'enrichissement du corpus, selon les termes d'une licence libre, car plus la ressource sera connue et diffusée, plus elle sera protégée des détournements qui ne profitent pas aux intérêts de la recherche

Frédéric GLORIEUX
École nationale des chartes

BIBLIOGRAPHIE SOMMAIRE

BON, B. GUERREAU-JALABERT, A. 2002. « Pietas : réflexions sur l'analyse sémantique et le traitement lexicographique d'un vocable médiéval », dans *Médiévales*, n° 42, p. 73-88. http://www.persee.fr/web/revues/home/prescript/article/medi_0751-2708_2002_num_21_42_1540

DU CANGE, C. 1678-1887. *Glossarium mediae et infimae latinitatis*, éd. augm., Niort, L. Favre, 1883-1887. http://ducange.enc.sorbonne.fr/

GÉRAUD, H. 1840. « Historique du *Glossaire de la basse latinité* de Du Cange », dans *Bibliothèque de l'École des chartes*, 1-1 p. 498-510. http://www.persee.fr/web/revues/home/prescript/article/bec_0373-6237_1840_num_1_1_461649

TEI. 1999, 2002, 2007. « 9. Dictionnaires » dans *P5 : Principes directeurs pour l'encodage et l'échange de textes électroniques*. Oxford — Providence — Charlottesville — Nancy, C.M. Sperberg-McQueen and Lou Burnard. http://www.tei-c.org/release/doc/tei-p5-doc/fr/html/DI.html

WOOLDRIDGE, R. 1977, 1997. *Les Débuts de la lexicographie française : Estienne, Nicot, et le* Thresor de la langue françoyse *(1606)*. http://www.chass.utoronto.ca/~wulfric/edicta/wooldridge/