Pour un corpus de textes latins en ligne

Projekt : ein Korpus lateinischer Texte open access on line

L'histoire ancienne et médiévale ne peuvent se renouveler sans mettre sur pied des méthodes spécifiques nouvelles pour faire naître une sémantique historique débarrassée des présupposés du sens commun qui obstruent la connaissance de ces époques. Ces nouvelles méthodes sont en grande partie fondées sur l'emploi ad hoc de procédures statistiques, adaptées à la forme non-standard des distributions lexicales.

Ces méthodes n'ont de sens que si elles peuvent s'appliquer à des « textes » en quantité suffisante, et en format numérique. Le rassemblement de corpus importants est donc un passage obligé.

C'est aux possibilités et aux problèmes liées à ce rassemblement qu'est consacré le présent papier. On prendra garde à ne pas oublier cette perspective, qui seule donne son sens à une telle entreprise.

Toutes les manipulations formalisées potentiellement utilisables (statistiques classiques ou text mining notamment) nécessitent que le texte revête une certaine forme et, le cas échéant, inclue des indications particulières ; pour obtenir cette forme, on procède à ce que l'on appelle un prétraitement. De ce point de vue, des progrès significatifs ont été réalisés : le groupe de latinistes baptisé OMNIA a mené à bonne fin la constitution des outils nécessaires à la tokenisation, au postagging et à la lemmatisation du latin (principalement Bruno Bon, Renaud Alexandre, Anita Guerreau-Jalabert [IRHT-CNRS], Eliana Magnani, Marie-José Gasse-Grandjean, Nicolas Perreaux [ARTEHIS-CNRS], Olivier Canteaut, Frédéric Glorieux [École des Chartes] et moi-même [CRH-CNRS]). Il est donc bien temps de se préoccuper de la disponibilité d'un corpus approprié à la recherche.

Le contexte

Nous avons perdu la clé des textes anciens[1], et par anciens, il ne faut pas hésiter à entendre « antérieurs à 1800 ». La philologie classique s'est efforcée, depuis ses plus lointaines origines, au 16e siècle, de constituer des « dictionnaires » de type historique, c'est-à-dire donnant les équivalents actuels des mots anciens ; les historiens de leur côté, au moins depuis le 19e siècle, mettent en garde constamment contre la tentation de l'anachronisme. On pourrait donc croire être en terrain sûr, et bien des historiens et des philologues se l'imaginent ainsi ; certains vont même, lorsque cette doxa est contestée, jusqu'à perdre leur sang-froid... Pourtant, un minimum d'attention et de réflexion suffisent pour s'apercevoir que les bases de cette lexicographie historique sont largement contestables, pour ne pas dire radicalement erronées.

L'idée de « traduire » en français du 21e siècle un mot de Cicéron implique par absolue nécessité le présupposé qu'il existerait des « notions » communes au premier siècle av.-J.-C. et au 21e après, présupposé courant, qui n'est autre qu'une expression dérivée d'une sorte de platonisme vulgarisé : l'arrière-plan du langage est fait d' « idées » éternelles. Ce présupposé[2] est d'autant plus vivace que son champ d'application est bien moins la philologie qu'une conception du monde commune à la plupart des idéologies visant la stabilité de l'ordre établi...

L'historien rationaliste ne saurait en aucun cas reconnaître la validité de ce présupposé ; on devrait inculquer profondément dans la tête de tous les étudiants d'histoire ces remarques (toutes récentes) de Kurt Flasch : « Geschichtliches Wissen kennt keine Ewigkeit...Das Zeitlose ist das Unwahre »[3].

Je reprends l'exemple (devenu, à mon corps défendant, vaguement célèbre) de vinea au 11e siècle[4] ; j'ai montré qu'une simple « traduction » provoque le contresens ; ce point n'est pas contestable, ne serait-ce que parce qu'une telle « traduction » laisse nécessairement dans l'ombre la présence de vitis qui avale en quelque sorte une forte part des occurrences potentielles de ce que nous appelons « vigne » ; mais même en tenant compte de ce point, on est toujours très loin d'une correspondance acceptable ; il faut bien cependant finir par dire qu'il existe malgré tout une (faible) part de points communs, et c'est sur cette part (latérale) que s'appuient les afficionados de la traduction à tous crins. Cette situation est tout à fait générale : le taux de « recouvrement » (appelons le ainsi) du paquet de relations constitutif du « sens » d'un mot au 11e et de celui de son « équivalent » actuel est toujours faible, souvent très faible. La compréhension des textes anciens passe par de tout autres moyens que la « traduction ».

Les mots, comme tout élément individuel au sein d'un système, sont des nœuds de relations, raison pour laquelle les méthodes de type structuralo-statistique constituent la voie royale. Jusqu'à une date relativement récente, l'état des techniques ne permettait pas de dépasser l'aspect purement structural qui, à lui seul, ne peut pas déboucher sur grand chose. A certains égards, l'absence de postérité des grands travaux de Jost Trier[5] et d'Ernst Cassirer[6] peut largement s'expliquer de cette manière. Mais voilà : le cours de l'histoire a voulu que nous nous trouvions au moment où cet état des techniques a subi une profonde mutation, et où ce volet statistique, qui manquait complètement jusqu'à présent, devient parfaitement envisageable.

Ce même cours de l'histoire a aussi voulu que la société dans laquelle vient d'avoir lieu cette mutation soit dominée, pour ne pas dire écrasée, par une logique sociale hypermercantile, qui tend à tout juger à la seule aune du potentiel de profit, à court terme et spéculatif autant que possible. Du coup, les énormes découvertes techniques relatives à l'indénombrable masse de textes qui constituent ce que l'on appelle « internet » sont avant tout orientées par cette logique ; des préoccupations des historiens, absolument personne ne se soucie le moins du monde, ils ont à se débrouiller seuls. Et ils le font d'autant plus difficilement que, comme on l'a rappelé plus haut, l'historicisation rationaliste des textes anciens implique

1 Tous les médiévistes savent que les clercs médiévaux donnaient aux textes antiques un sens totalement anachronique ; une connaissance directe, plus ou moins intuitive, de l'emploi et du sens des textes médiévaux s'effaça rapidement dans la seconde moitié du 17e siècle : ce fut justement l'instant où Charles DuCange entrepris son célèbre glossaire ; il était encore capable de saisir une partie des significations, mais sentait bien que cette connaissance intuitive était en train de disparaître.

2 Ce que j'appelle le paradigme de la version latine, ancré au plus profond de la cervelle de la plupart des latinistes, jamais mis en cause, et pourtant catastrophique au delà de toute mesure.

3 Kurt FLASCH, *Kampfplätze der Philosophie*, Frankfurt a. M., 2008, p. 127.

4 A. G., « Vinea », in Monique GOULLET & Michel PARISSE (éds), *Les historiens et le latin médiéval*, Paris, 2001, pp. 67-73. *L'avenir d'un passé incertain*, Paris, 2001, pp. 195-202.

5 Jost TRIER, *Der deutsche Wortschatz im Sinnbezirk des Verstandes*, Heidelberg, 1931 ; *Aufsätze und Vorträge zur Wortfeldtheorie*, La Haye-Paris, 1973.

6 Ernst CASSIRER, *Philosophie der symbolischen Formen*, 3 vol., Berlin, 1923-1929 (tr. fr. Paris, 1972).

l'usage méthodique de présupposés non-conformes (litote) et que par dessus le marché leurs compétences statistiques sont en-dessous de zéro.

Mais, si les techniciens se sont trouvés pour ainsi dire immédiatement aux prises avec d'invraisemblables masses de textes dans les langues contemporaines, il n'en va pas du tout de même pour les historiens : là, la question de la constitution des corpus numérisés est un préalable massif et déterminant ; on va tenter ici de voir ce qu'il en est en 2011 des textes en langue latine.

Le cadre fondamental (en France) : le droit d'auteur dans sa version française

Bien que cela puisse paraître paradoxal étant donné l'ancienneté même des textes dont nous parlons, il est primordial de bien fixer les idées en matière de droit d'auteur (Urheberrecht, diritto d'autore, derecho de autor, copyright), en France pour commencer.

La France est signataire de la Convention de Berne (1886) et du Traité de l'Organisation mondiale de la propriété intellectuelle (1996). L'ensemble des dispositions législatives qui s'appliquent en France au droit d'auteur sont reprises et unifiées dans le Code de la propriété intellectuelle (1992), constamment tenu à jour et disponible très commodément sur Legifrance (consultation en ligne et téléchargement)[7].

Le principe de base est simple, et ne doit jamais être perdu de vue : le volet patrimonial du droit d'auteur naît d'un acte de création, et appartient par définition à l'auteur et **à lui seul**, qui peut déléguer l'exploitation du volet patrimonial par contrat (créant ainsi des « ayant-droits »). La date fondamentale est celle de la mort de l'auteur. Jusqu'en 1889, ces droits s'éteignaient à la mort de l'auteur ; depuis 1889, il durent encore 50 ans ; depuis 1999, ce délai a été porté à 70 ans. Plus précisément : 70 ans depuis le premier janvier de l'année civile qui suit la mort de l'auteur. Autrement dit, en 2011, toutes les œuvres des auteurs morts avant le premier janvier 1941.

Il faut ajouter immédiatement qu'une traduction est considérée comme équivalente à une création, et entraîne les mêmes droits, avec les mêmes délais.

Il faut aussi ajouter quelques exceptions. En France, la plus importante est celle des auteurs « morts pour la France », pour lesquels ce délai est prolongé de 30 ans. Inversement, les textes législatifs sont du domaine public du jour de leur publication. Le seul cas susceptible de prêter à discussion est celui des œuvres dites posthumes. A ma connaissance, strictement rien dans la législation (ni dans la jurisprudence ?) ne s'applique aux textes produits avant l'époque où la notion de « publication » commença à prendre un sens, soit le milieu du 15e siècle. Dans le cas ordinaire, la loi prévoit un délai de 25 ans à compter de la publication. Mais au bénéfice de qui ? des « propriétaires des droits à un titre quelconque ». S'il n'y a eu aucun contrat, je ne vois pas comment cette formule pourrait désigner d'autres que les héritiers. Donc les héritiers du marquis de Sade ou de Lamartine peuvent prétendre aux droits sur des publications de manuscrits inédits. Qui connaît des héritiers d'auteurs médiévaux ? De toute manière, et dans la pire des hypothèses (dont la plausibilité, e.g. pour des textes scolastiques anonymes du 13e siècle, me semble strictement nulle) le délai n'est que de 25 ans.

Tout cet ensemble se résume parfaitement avec la plus grande concision en ce qui concerne notre objet : **les textes dont nous traitons ici sont tous du domaine public, sans aucune exception**.

Corollaires essentiels : 1. la possession matérielle d'un support quelconque n'entraîne aucun autre droit que le droit de propriété ordinaire appliqué à l'objet-support ; 2. la reproduction d'un texte, quel que ce soit le procédé, le coût, le temps de travail nécessaire, etc., n'entraîne aucun droit. Plus précisément encore ici : **la numérisation n'entraîne pas plus plus de droits qu'une reproduction anastatique ou une photocopie, i.e. aucun**.

Glose : s'agissant de ce que l'on appelle « édition » au sens érudit du terme, il faut distinguer divers aspects auxquels s'appliquent des droits totalement différents. L'introduction et les notes, toute traduction,

7 La manière la plus simple et compréhensible d'entrer en matière est de se reporter à l'article « droit d'auteur » sur fr.wikipedia.org, et bien entendu d'examiner dans la foulée les articles correspondants en allemand, italien, espagnol et anglais, pour le moins [il serait souhaitable de disposer de précisions sur les différences - faibles d'ailleurs - entre les divers pays]. Traité de Berne : http://www.wipo.int/treaties/fr/ip/berne/trtdocs_wo001.html ; traité de l'OMPI : http://www.wipo.int/treaties/fr/ip/wct/index.html ; Code de la propriété intellectuelle, chapitre droit d'auteur : http://www.legifrance.gouv.fr/affichCode.do?idArticle=LEGIARTI000006278868&idSectionTA=LEGISCTA000006161633&cidTexte=LEGITEXT000006069414&dateTexte=20091206

ressortissent du droit d'auteur simple ; inversement, le texte lui-même, comme d'ailleurs les variantes (elles-mêmes résultats de modifications fort anciennes), sont du domaine public sans exception. De même d'ailleurs, ce que l'on appelle « structure d'une base de données » est protégé par le Code de la propriété intellectuelle[8], mais cette protection ne s'applique à aucun égard aux éléments contenus dans cette base.

Au total, il est facile de voir qu'un éditeur (au sens commercial) procédurier trouvera sans doute difficilement un avocat qui lui promette le succès devant les tribunaux pour tout cas autre qu'une publication postérieure (en 2011) à 1986 d'un texte ancien jamais publié antérieurement ; à ma connaissance, un cas de ce dernier genre n'a jamais été plaidé, et ainsi il n'existe malheureusement aucune jurisprudence. Il serait hautement souhaitable qu'un tel différend soit jugé au moins une fois. Mais, pour tout autre cas, aucune discussion n'est possible, et divers cas jugés l'ont tous été dans le même sens, avec des attendus extrêmement brefs.

On ne saurait achever ce paragraphe sans souligner qu'en pratique, très regrettablement, cette situation juridique fort simple est très généralement ignorée, en particulier dans les milieux scientifiques les plus directement concernés, et que divers éditeurs malhonnêtes profitent tant qu'ils peuvent de cette ignorance pour impressionner et intimider ces milieux scientifiques, non sans quelque succès[9]. Il paraît donc souhaitable de considérer comme une obligation fondamentale un rappel de ces principes en tête de tout site de mise en ligne de textes anciens. Il y a peu de modèles, on peut cependant citer la page (d'ailleurs en elle-même fort riche et bien faite) de l'historien de la philosophie canadien Peter King (http://individual.utoronto.ca/pking/resources.html).

Notons enfin, dans le même sens, que l'on rencontre constamment des procédures qui visent à faire reconnaître diverses « interdictions » : les commerçants filous bien connus, qui bardent leurs CD et leurs sites de rodomontades et de gesticulations aussi mensongères que comminatoires et prétendent subordonner l'usage de leurs textes à la reconnaissance de leurs soi-disant « droits » (=commerciaux), comme d'ailleurs, en sens inverse, la BNF sur son site Gallica, qui prétend interdire l'usage commercial de ses fichiers pdf. Tout cela contrevient à la loi, et n'a donc strictement aucune valeur juridique (nullité de plein droit : cliquez et signez tant que vous voudrez, cela ne vous engage à rien).

Insistons : l'appartenance au domaine public implique la liberté absolue de diffusion et de copie ; ce qui signifie qu'aucun éditeur, au sens scientifique ou au sens commercial - c'est la même chose -, ne dispose du moindre droit ni ne peut prétendre en réintroduire d'une manière ou d'une autre ; quel que soit le support matériel, papier, cdrom, internet, n'importe qui peut copier, utiliser, diffuser tous les textes du domaine public, autant qu'il le veut et sous la forme qui lui convient[10]. Sans la moindre restriction d'aucune sorte. Ce qui signifie que toutes les sources que nous allons examiner à présent sont copiables de manière totalement légale. Toute indication du contraire est dénuée de fondement et de tout effet.

---

8 On entre par : nhttp://www.legifrance.gouv.fr/affichCode.do;jsessionid=F1F8F30D898A5C6409E483C710B21131.tpdjo17v_3?idSectionTA=LEGISCTA000006161660&cidTexte=LEGITEXT000006069414&dateTexte=20080129 Il faut noter accessoirement que l'extension maximale des droits sur les bases de données est de 15 ans à partir de leur création (= 1996 pour 2011). Ce droit est accessoire du droit d'auteur, la loi précise explicitement qu'il s'applique « sans préjudice du droit d'auteur ». Qu'un texte soit ou non inclus dans une base de données originale ne modifie en rien son statut.

9 En français familier, on appelle cela un bluff minable. Une question intéressante est de savoir pourquoi un tel succès. C'est une affaire d'ordre sociologique ; antérieurement au développement massif de l'électronique, la publication papier représentait un gage de sérieux, et la publication par un « grand éditeur », un marche-pied vers la notoriété. Au demeurant, en Allemagne, un individu ayant soutenu son doctorat avec succès ne peut utiliser le titre révéré de « Doktor » qu'après publication de sa Dissertation. Mais cette situation reposait seulement sur un système technique (exclusivité de la diffusion par l'imprimerie) et n'avait aucun fondement intellectuel : la moindre expérience permet de savoir le caractère tout à fait indéterminé de la composition des sacro-saints « comités de lecture » et autres « comités éditoriaux », auxquels on accède par des règles tout autres que scientifiques. Ce système n'est d'ailleurs pas antérieur à un 20e siècle bien avancé. Le nouveau système technique l'a privé de fondement, il est en passe de devenir obsolète très rapidement. Mais l'inertie habituelle des milieux académiques se combine à la rapacité et au culot des éditeurs pour le prolonger tant que possible ; on ne peut faire aucun pronostic sur la durée de cette survie artificielle. La vogue grandissante de (pseudo-)critères « bibliométriques » risque de contribuer à cette prolongation dérisoire.

10 En respectant, signalons-le au passage, le volet *moral* du droit d'auteur, qui est (en France du moins) intangible et perpétuel (on n'a pas le droit de diffuser un texte en modifiant celui-ci quant au fond ; par exemple, l'épuration de passages « crus » de tel ou tel texte est interdite... Cas jugé vingt fois, toujours dans le même sens).

Ces préalables indispensables étant posés, passons au plus gros morceau : l'état des lieux, autrement dit la réponse à la question « en 2011, de quels textes latins numérisés disposons-nous ? »

On examinera quatre aspects principaux : 1. les textes en bon ordre récupérables sur internet, 2. les textes disponibles sur CD, 3. les problèmes de l'OCR, 4. les lacunes. On entendra par « textes latins anciens » tous les textes écrits dans cette langue des origines (7e s. avant) au 18e après.

Les ressources en open access

Il n'y a pas pléthore de sites généralistes spécifiquement latins. Le seul que je connaisse est www.thelatinlibrary.com. Très actif entre 1998 et 2007, un peu en sommeil depuis. On peut très facilement aspirer tout le site, on obtient un répertoire de 220 mégaoctets, pour plus de 3000 fichiers. Ceux-ci sont en html, et on peut les convertir soit avec un programme ad hoc, soit en utilisant la procédure « enregistrer sous » dans firefox ou opera. L'époque dite classique est fortement couverte, la suite plus épisodiquement, mais s'étend jusqu'au 19e siècle. Cela doit sans doute représenter un corpus d'environ 30 millions de mots. Un fichier indique la source de toutes les numérisations. Selon Wikipedia, ce site est maintenu par l'académie Ad Fontes (http://www.adfontes.com), mais je n'ai découvert aucune confirmation sur le site en question.

On trouvait, il y a quelques années (?), en annexe au programme « lector latinus », d'Abram Ring, un gros fichier zip contenant une collection de fichiers-textes en open access ; ce fichier n'est plus disponible (2011), sauf à acheter pour 25 $ le programme susdit[11]. Mais il a été largement récupéré et peut aisément circuler. Dézippé, l'ensemble pèse environ 230 mégaoctets, et comporte plus de 3800 fichiers. Partie en format txt, partie en format rtf. Il fait pour une large portion double emploi avec le site précédent, dont il a repris tous les éléments en ligne en janvier 2006. Comporte notamment une grande quantité d'inscriptions latines antiques, et de textes « musicaux ». Les fichiers sont classés par périodes d'un, deux ou trois siècles (onze en tout). Peut-être 35 millions de mots.

Les sites purement généralistes n'accordent en général au latin qu'une place dérisoire. Le « project Gutenberg[12] » propose plus de 35000 fichiers, dont seuls 71 sont en latin. La bibliotheca augustana[13] (Augsburg) propose une grande quantité de textes de toutes époques, mais la récupération doit se faire à la main, je ne sais pas si personne a tenté l'opération. On doit mentionner le site wikisource, plus exactement, pour nous, vicifons (http://la.wikisource.org/wiki/Pagina_prima) ; je vous recommande instamment la note de la page http://la.wikisource.org/wiki/Patrologia_Latina. Quelques textes que l'on ne trouve pas ailleurs, comme le commentaire de Macrobe sur le songe de Scipion, la traduction latine de l'introduction aux catégories de Porphyre, ou l'histoire scolastique de Pierre LeMangeur, le moriae encomium d'Erasme. On peut raisonnablement penser que ce site a l'avenir devant lui.

Pour le reste, on peut essayer de dresser une liste (très brève et forcément partielle) des sites plus ou moins spécialisés.

Pour le latin antique, le projet Perseus propose un corpus classique de 5,5 millions de mots. (http://www.perseus.tufts.edu/hopper/collection?collection=Perseus:corpus:perseus,Latin+Texts&redirect=true) Le site IntraText a été créé et est maintenu par une société italienne (Eulogos spa) qui, sous la direction de Nicola Mastidoro et avec la collaboration d'une multitude de religieux de divers ordres, a dépensé beaucoup d'énergie pour structurer son corpus d'une multitude d'hyperliens et de statistiques. Le site offre des textes de plus de trois cents auteurs latins, mais il faut les récupérer un par un et par petits morceaux (avec la fonction « enregistrer sous » = CTRL+Maj+S sous FF, CTRL+S sous Opera), ce qui limite fortement son intérêt. Quelques textes peu courants : http://www.ipa.net/~magreyn/. Textes intéressants d'époque moderne : http://www.philological.bham.ac.uk/library.html. Un site très copieux de textes de droit romain : http://webu2.upmf-grenoble.fr/Haiti/Cours/Ak/index.htm (va jusqu'aux premières « lois barbares », d'après l'édition des MGH). Quelques textes : http://penelope.uchicago.edu/Thayer/E/Roman/home.html.

Bien entendu, il convient de faire une large place aux sites qui proposent des chartes et autres

11 La personne qui a rassemblé ces textes est parfaitement libre de vouloir en tirer profit : c'est le principe de la liberté du domaine public ; mais quiconque dispose de ce fichier est libre de le rediffuser, c'est tout autant le principe du domaine public !

12 Histoire et description du projet sur fr.wikipedia.org. Il existe, en plus du site principal (anglophone), des sites allemand, français, espagnol...

13 http://www.hs-augsburg.de/~harsch/augustana.html

documents du même type, en ligne. Le plus important, et de loin, est celui de l'Artehis-CNRS de Dijon, dans le cadre du projet CBMA (Cartae Burgundiae medii aevi http://www.artehis-cnrs.fr/Le-projet-CBMA). Il s'agit d'un projet très récent (2004), dont les membres ont réussi, en moins de cinq ans, à numériser et à mettre en ligne et en téléchargement une impressionnante quantité de chartes (http://www.artehis-cnrs.fr/page-documentaire-CBMA). Les travaux d'Eliana Magnani, Marie-José Gasse-Grandjean et Nicolas Perreaux ont abouti à un résultat hors norme ! 35 cartulaires, 21 mégaoctets de texte, soit environ 3,5 millions de mots, sont ainsi disponibles au téléchargement en format txt, tout en étant consultables au travers du logiciel PhiloLogic. En Italie, la Lombardie se distingue avec le site piloté par Michele Ansani, Codice diplomatico della Lombardia medievale (http://cdlm.unipv.it). Une vaste entreprise, qui vise à une édition diplomatique, disponible en ligne, de toutes les chartes correspondant aux anciens diocèses de la Lombardie médiévale, du 8e à la fin du 12e siècle. Le site peut être aspiré, mais les fichiers html nécessitent un sérieux nettoyage, bourrés qu'ils sont de considérations diplomatiques variées. Ce travail réalisé, on a plus de 4800 textes pour plus de 3,7 mégaoctets. L'École des Chartes a entrepris la mise en ligne d'une série de cartulaires franciliens http://elec.enc.sorbonne.fr/cartulaires/ ainsi que d'un ensemble d'autres sources historiques importantes[14]. Le « projet DEEDS » de Michael Gervers permet la consultation de plus de 8000 chartes anglaises datées tirées de la plupart des cartulaires anglais http://res.deeds.utoronto.ca:49838/research/. L'ARTEM est un autre projet très ancien (1971), qui a abouti à la mise en ligne (avec l'aide de l'IRHT-CNRS) de plus de 5000 chartes dont les originaux, antérieurs à 1121, sont conservés dans les dépôts publics français http://www.cn-telma.fr/originaux/colophon/.

Un cas tout à fait particulier est celui des MGH. Beaucoup de médiévistes s'étonnent, depuis longtemps, de l'impossibilité de disposer des textes en version numérisée. Ces textes sont tous du domaine public (cf plus haut) et l'apparat critique l'est également pour la majorité des volumes, les auteurs étant décédés avant 1941. Donc, n'importe quel individu, institution ou entreprise peut entreprendre une numérisation générale de l'ensemble des textes de la totalité de la collection sans avoir d'autorisation à demander à qui que ce soit. Et, éventuellement, proposer le résultat à la vente... ou le mettre en ligne en accès libre.

On s'attendrait cependant à ce que la vénérable institution, une des plus prestigieuses d'Europe, fasse elle-même cette mise en ligne en open access. On a vu apparaître le site MGH digital[15] ! mais il ne s'agissait que des images des pages des éditions papier. Cependant, récemment (?) la situation paraît avoir évolué. En effet, sur le bandeau supérieur de la page de recherche, figure un onglet « html » ; et là, on obtient, selon les volumes, deux types de réponse : soit « Dokument nicht verfügbar » soit la même page, mais en format texte parfaitement numérisé et recopiable. Rien ne semble prévu pour un téléchargement, il faut donc utiliser la fonction « enregistrer sous », autant de fois qu'il y a de pages dans un volume. C'est une situation absurde. Il serait plus que souhaitable que les médiévistes allemands fassent en sorte que la totalité de ces textes soient numérisés et téléchargeables, le gain pour la recherche serait considérable.

Il faut terminer cette trop rapide revue par l'indication d'une situation globale : il existe une multitude de sites consacrés spécialement à un auteur, ou à un petit groupe, et sur lesquels on trouve un ou plusieurs textes ; la situation varie constamment, et je ne connais pas de site qui en donne une liste vraiment utilisable ; on peut en trouver en examinant les liens indiqués dans les articles-auteurs de Wikipedia, en prenant garde à consulter méthodiquement les quatre langues principales ; les moteurs de recherche classiques et bien connus sont également utiles.

Le site consacré à un auteur qui mérite une mention très spéciale est celui consacré à Thomas d'Aquin[16]. Enrique Alarcón a en effet récupéré les fichiers issus du travail pionnier du père Roberto Busa, et les a mis en ligne en accès libre : on dispose ainsi, très commodément, du « corpus thomisticum », qui comprend l'ensemble des œuvres de Thomas d'Aquin, ainsi qu'un très copieux complément de quelques auteurs scolastiques du 13e siècle. A cela, Enrique Alarcón a joint une collection de matériaux annexes fort utiles, notamment le très intéressant Thomas-Lexikon de Ludwig Schütz (auf deutsch ! = les principaux termes de Thomas soigneusement commentés). Au total, 139 mégaoctets de matériaux, un record.

La Vulgate est disponible sur plusieurs sites, il s'agit toujours de quelque chose de très voisin de la Vulgate dite clémentine, qui résulte de la normalisation pontificale à la fin du 16e siècle[17]. Tout le monde sait

14 On doit aussi signaler que l'École des Chartes met en ligne en accès libre le texte complet de la dernière édition du DuCange http://ducange.enc.sorbonne.fr/, dont l'intégralité peut être téléchargée.

15 http://bsbdmgh.bsb.lrz.de/dmgh_new/

16 http://www.corpusthomisticum.org/

17 http://www.drbo.org/lvb/ http://vulsearch.sourceforge.net/html/ plusieurs autres...

(l'observation a commencé à la fin du 13e siècle, chez les dominicains du couvent Saint-Jacques), que la Bible latine médiévale est un vrai fouillis de variantes, ce qui rend particulièrement compliquée l'identification des citations, qui constitue pourtant une obligation pour tout éditeur moderne. Il existe depuis longtemps des travaux consacrés aux Bibles illustrées, mais il est rarissime que les auteurs de ces analyses se soient préoccupés des textes exacts qu'ils avaient sous les yeux. Des sites traditionalistes proposent les textes de la messe tridentine et du bréviaire ancien, sous des formes cependant peu commodes[18].

Le site de Peter King, déjà cité, fourni 165 textes, pour 24 mégaoctets. Je prends un seul exemple, celui de Nicolas de Cues. On trouve une partie de ses textes dans la Bibliotheca Augustana, déjà mentionnée, mais il existe aussi un site spécial, où se trouvent les œuvres complètes (http://urts99.uni-trier.de/cusanus/content/suche.php) ; sur la page d'accueil, on trouve cette définition de l'objectif : « sie [die Werke NK] durch die Digitalisierung umfassender zu erschließen und zu ihrer weltweiten Verbreitung beizutragen. » Na schön und gut, on se demande bien alors pourquoi on ne peut pas récupérer lesdits textes autrement que par copier-coller... Il s'agit là en fait d'une situation générale, et qui tend à se généraliser davantage chaque jour.

Les CDRoms

Je n'aborde ici que ceux sur lesquels on trouve des fichiers plus ou moins facilement récupérables en mode txt. Chacun sait qu'il en existe une collection où le cryptage des fichiers textes les rend totalement opaques. Notons seulement que, dans certains cas ponctuels, on peut cependant récupérer les morceaux que l'on souhaite (c'est une simple question de temps et de patience...), en faisant des copies d'écran en mode texte, grâce à des logiciels spécialisés (comme SnagIt), qui permettent en somme le copier-coller même lorsque la fonction de copie est désactivée.

Le moment de l'apparition et de l'essor des CD a été très bref, une dizaine d'années, du début de la généralisation des lecteurs de CD (début des années 90) à l'essor d'internet (début des années 2000). Du coup, il en existe peu et, s'agissant d'un domaine très spécial comme les textes latins, on a donc vite fait un tour complet.

L'essentiel se résume à quatre entreprises, d'ampleur et de notoriété très différentes : la PL, les AASS, la Poetria Nova et la Quellensammlung zur mittelalterlichen Geschichte.

Tous les médiévistes connaissent et fréquentent plus ou moins régulièrement la PL. L'entreprise a consisté en un numérisation stricte en mode texte de la totalité de toutes les informations contenues dans les volumes de Migne. Rien n'a été ajouté ! il est même impossible de prétendre à l'originalité de la « base de données » présente sur les CD, puisqu'il s'agit, par construction, d'un décalque exact de la structure des volumes d'origine. La seule originalité réside dans le logiciel de consultation (recherche, lecture). Tout le reste appartient strictement au domaine public. En pratique, on dispose de cinq fichiers, à peu près lisibles. 920 mégaoctets. Il faut défalquer les titres, les index, les notes, et les très copieuses préfaces, introductions et autres textes d'érudits modernes ; je n'ai qu'une idée très vague de ce que cela représente, sans doute pas plus de 20% ; resteraient 740 mégaoctets, peut-être 120 millions de mots (????). Contrairement à ce que l'on pourrait croire en lisant la notice, ces fichiers ne comportent aucune balise ; les indications internes sur les titres, les auteurs, les périodes, ne sont pas très faciles à récupérer : cela réclame l'écriture de programmes ad hoc, mais c'est tout à fait faisable.

La structure des fichiers des AASS est très voisine, 437 mégaoctets. Cette grande série est infiniment moins fréquentée que la PL : il est très difficile de s'y retrouver, et l'on ne sait jamais très bien ce que l'on trouve... Je n'ai aucune idée de la proportion, dans cette masse, des textes latins utilisables antérieurs disons au 17e siècle, je n'ai pas davantage d'idée sur la manière dont on pourrait essayer d'exploiter certains au moins de ces textes dans une base de données structurée. Il reste qu'il s'agit pourtant d'une série classique, à laquelle font référence de nombreux instruments de travail, il faudrait sans doute parvenir au moins à extraire certains textes précis.

La Poetria Nova est une entreprise italienne (Sismel) de bonne facture. En examinant la structure du CD, on ne découvre pas sans une certaine surprise que tout est en clair ! L'ensemble des vers est contenu dans un simple fichier rtf, et toute la structure d'organisation dans une base relationnelle MSAccess. 43 mégaoctets,

18 http://www.traditio.com/off.htm http://divinumofficium.com/www/horas/Help/download.html (ces url m'ont été indiquées par Renaud Alexandre, que je remercie).

plus d'un millier d'œuvres de près de 600 poètes, 5,2 millions de mots : un must !

L'entreprise berlinoise Heptagon a créé voici quelques années une série de trois CD Quellensammlung zur mittelalterlichen Geschichte. Il s'agit presque exclusivement de textes de chroniques tirés directement des MGH, au total une centaine (pour chaque texte, on trouve également une traduction allemande). Ces CD sont vendus à des prix raisonnables (autour de 50 €) et le logiciel de consultation, original, est tout à fait honorable. Environ 25 mégaoctets, mais les fichiers, en clair, sont cependant encombrés de caractères parasites qu'il faudra nettoyer. Très utile cependant tant que les MGH ne seront pas disponibles.

C'est à peu près tout ce que je connais (à compléter !!!!). Un autre CD italien, consacré à la légende dorée, est crypté ; on peut cependant utiliser (fonctionnellement) le copier-coller ; comme il n'y a « que » 157 vite, le tout est récupérable en une journée.

L'OCR propre et le « dirty OCR »

Il n'est sans doute pas nécessaire de s'étendre sur la « méthode Google » et les diverses institutions qui utilisent des méthodes identiques. Il s'agit de mettre en place une chaîne de production de type industriel où les opérations sont complètement automatisées. Les livres entrent d'un côté et ressortent de l'autre, accompagnés d'un fichier-image qui a été soumis au traitement d'un logiciel d'OCR, qui a produit du « searchable pdf ». Divers logiciels permettent d'extraire de ce pdf le texte qui a été produit. La qualité est éminemment variable, en grande partie en fonction de la qualité de la typographie d'origine. L'expérience montre cependant que la fonction « recherche » des logiciels de lecture de pdf fonctionne avec un taux de réussite assez élevé, surtout si l'on fait attention. Mais on ne peut pas constituer d'index, encore moins se livrer à des traitements un peu plus complexes.

La question (importante, pour ne pas dire plus) qui se pose ici est de savoir quel genre de procédure de post-traitement pourrait permettre d'accélérer massivement la correction, c'est-à-dire diminuer drastiquement la proportion des cas nécessitant une intervention directe d'un opérateur. Il s'agit là, notons-le au passage, d'un problème de première importance, qui ne semble pourtant pas avoir jusqu'ici suscité beaucoup de réflexions et d'expérience[19].

Un étudiant ou un jeune chercheur qui souhaite entreprendre l'étude fine d'un auteur ou d'un corpus sait qu'il n'a guère d'alternative à une océrisation aussi soignée que possible, suivie d'une correction attentive. Dans tous les cas, on ne saurait assez répéter que le résultat dépend fondamentalement de la qualité des fichiers-images, c'est-à-dire avant tout de la qualité du scan. Un réglage très fin des paramètres du scanner est une obligation, il faut faire une grande quantité d'essais pour déterminer les paramètres qui vont produire le fichier qui « passera » le mieux à l'OCR. Ensuite, on a le choix : soit le logiciel du commerce (considéré comme) le plus efficace, soit un logiciel libre de pure reconnaissance en mode image. Pour obtenir FineReader Pro 10, on devra débourser 150 €. Avec ça (je parle par ouï-dire), on peut obtenir du txt ou du pdf searchable, et la qualité du résultat semble souvent excellente.

L'alternative libre est Gamera (http://gamera.informatik.hsnr.de/addons/ocr4gamera/index.html) ; d'abord créé et développé aux USA, le logiciel est aujourd'hui maintenu dans une université allemande (Krefeld / Mönchengladbach). Le principe est simple : une première phase consiste à apprendre au logiciel la valeur de tous les signes qu'il peut individualiser sur une image ; puis - second temps - à le laisser travailler. On voit bien 1. que l'étape clé est celle de l'apprentissage, qui peut être relativement longue ; 2. que la méthode est spécialement adaptée pour toutes les typographies disons « anciennes » qu'un logiciel standard ne reconnaîtra pas, le cas classique étant celui du s ancien (toujours lu f), de toutes les ligatures et autres dessins de lettres inhabituels. Pour océriser des pages du Monde ou du courrier commercial, ce n'est sans doute pas optimal, mais des éditions du 16e ou du 17e passeront certainement dix fois mieux de cette manière. Il est vraiment étrange et regrettable que l'on dispose ici de si peu d'expériences.

Les lacunes

Je n'ai pas réussi à trouver les canons des conciles œcuméniques des premiers siècles. Des documents fondamentaux pour l'histoire de tout l'Occident médiéval et moderne. Et ce malgré l'abondance des sites

19 L'idée la plus simple consiste à utiliser un correcteur orthographique, ce qui n'est cependant pas immédiat, étant donné l'extrême variation graphique du latin médiéval : on ne sait jamais bien s'il s'agit d'une variante ou d'une cacographie. Des procédés plus sophistiqués, genre text mining, sont envisageables.

consacrés à des textes « chrétiens », de diverses confessions. Je n'ai jamais entendu parler d'un projet consistant à numériser la grande collection des conciles de Mansi. On a les images de 51 volumes sur Gallica, rien de plus.

Je n'ai pas trouvé le Rationale divinorum officiorum de Guillaume Durand, encore moins les Postilles de Nicolas de Lyre, ni le Catholicon de Jean Balbi (un des premiers textes imprimés par Gutenberg !), tous des textes (fin 13e, tout début du 14e) très utilisés et extrêmement célèbres pendant toute la fin du Moyen Age, que l'on peut considérer comme indispensables à l'étude des 13e-16e siècles (et au delà). Je n'ai découvert que tout récemment une numérisation (en dirty OCR...) des deux volumes essentiels du Corpus Juris Canonici d'Emil Friedberg (Columbia University). La grande encyclopédie de Vincent de Beauvais n'est disponible qu'en partie.

Tous ces textes ont au moins un point commun : ils ne sont PAS l'œuvre de « grands auteurs » ; comme, d'autre part, on ne leur attribue aucune « valeur documentaire », ils n'intéressent personne ! (que font les historiens de la liturgie, de l'exégèse, du droit canon ?)

Le secteur des chroniques est tout aussi vide. En dehors de quelques textes des MGH (cf plus haut), on n'a rien, en particulier pour l'Angleterre (l'immense série des chroniques anglaises, d'un énorme intérêt ; on connaît Mathieu Paris, il y en a des dizaines d'autres), mais la situation est tout aussi désolante pour la France, l'Espagne ou l'Italie.

Ces considérations sur les lacunes ont, par la force des choses, un caractère quelque peu subjectif. Mais elles permettent cependant de dresser un constat simple et crucial : des pans entiers - et non des moindres - de l'ensemble des ouvrages en latin qui dorment sur les rayons des bibliothèques sont encore indemnes de toute numérisation (quel que soit le mode de mise à disposition considéré). C'est un point que l'on doit bien avoir présent à l'esprit quand on réfléchit en termes de construction d'un corpus latin généraliste.

Indexation

La moindre expérience en matière de constitution et d'utilisation de corpus de textes permet de se rendre compte que l'intérêt d'un corpus réside pour une très large part dans la pertinence de son indexation. Sans indexation, a n'a qu'une masse indistincte, autant mettre tous les textes dans un même fichier, les manipulations seront beaucoup plus simples. Mais une mauvaise indexation est une gêne sensible, des découpages (catégorisations) erronés pouvant très facilement amener à des conclusions complètement fallacieuses.

Or, si l'on est optimiste et que l'on imagine possible de rassembler la plus grande partie des textes disponibles, on obtient (évaluation très incertaine) quelque chose comme quinze mille textes, plus quelques dizaines de milliers de chartes (plus de 100000 ?). Pour tout ce qui est antérieur à 1200, la proportion de ce qui est disponible est importante mais, dès le cours du 13e, cette proportion baisse et finit par devenir très faible du fait de l'explosion de la masse de « l'écrit documentaire ».

Jusqu'ici, on s'est le plus souvent contenté d'une sorte de « fiche d'identité », soit le nom de l'auteur et le titre de l'œuvre ; les textes étant plus ou moins répartis en « grandes masses » chronologiques. Rien que cela pose déjà de très gros problèmes ; la proportion d' « auctor incertus » est considérable, les titres sont le plus souvent des inventions modernes, et les attributions comme les datations sont imprécises et souvent contestables et contestées. Et je ne parle pas de la répartition géographique, pourtant cruciale étant donné l'extension spatiale du domaine latin.

On songe habituellement en termes de « grandes catégories », du type : prose / vers, sermons, lettres, vite, textes « législatifs », « philosophiques », « théologiques », « exégétiques », discours, chroniques, romans, théâtre, bref ce que la théorie littéraire ordinaire appelle « genres ». Mais chacun sait le caractère incertain, daté, contestable voir dangereux, de ces étiquettes.

C'est la question très générale des classifications universelles. Un vieux rêve qui remonte au moins à la Renaissance, et que des bibliothécaires ont plus ou moins réussi à concrétiser au 19e siècle (Dewey et la CDU). Une vieille rengaine, que la Library of Congress a modernisé, et qui est passée tout droit dans la « base Rameau ». C'est ingérable et inutile, certains disent nocif. Il suffit, un jour de fatigue, d'essayer d'utiliser les « mots matière » de l'index de la BNF pour s'apercevoir au bout de deux minutes que l'on n'y comprend rien, que c'est opaque et que l'on perd son temps... Cela n'intéresse que les bibliothécaires, pas les lecteurs.

L'explosion d'internet, prise en main par des novices et des néophytes, a fait ressurgir le fantasme. Qui se trouve naïvement baptisé « ontologie(s) ». Avec tout le fatras qui va avec, genre metadata, « web sémantique » et j'en passe. Il y a belle lurette que les moteurs de recherche laissent ces en-tête de côté, sachant

leurs biais et leur inutilité : ils indexent en plein texte, c'est le seul moyen d'obtenir un résultat acceptable.

Dans notre domaine, nous avons affaire à une excroissance pathologique, dénommée TEI. Qui en fait ne concerne qu'un tout petit groupe, analogue à une secte anglo-saxonne. Aux origines (lointaines : 1987), quelques principes de bon sens, destinés à faciliter l'échange de fichiers-textes. Mais rapidement est apparue une tendance absurde à vouloir créer un nouveau système de description universelle (même si, en théorie, l'idée n'est pas reconnue). Si l'on veut suivre toutes les règles, on doit utiliser un manuel de plusieurs centaines de pages, et l'on se retrouve avec des fichiers où les balises tiennent trois ou quatre fois plus de place que le texte lui-même, résultat auquel on ne peut bien entendu parvenir qu'au prix d'une débauche de temps et de travail. Pour une utilité voisine de zéro.

Cette dérive ne doit pas empêcher de poser raisonnablement le problème. L'idée de placer quelques informations élémentaires en tête d'un fichier-texte, en les insérant dans une poignée de balises xml (éventuellement analogues à celles de la TEI, favorisons la concorde...), tient parfaitement la route. Mais jamais rien de plus. Il va de soi, que pour un corpus donné, le jeu d'indications et de balises doit être homogène.

Toute la difficulté consiste à choisir ces indications de manière à faciliter l'utilisation de CE corpus - pas d'un autre. Et, soulignons-le, cette difficulté est considérable. Car chacun sait que l'on retrouve le plus souvent en fin de recherche les éléments de structure que l'on a soi-même insérés. Or, comme on l'a rappelé en préambule, l'objet même de la constitution de ce corpus est d'éliminer, ou au moins de réduire drastiquement, la part des présupposés. Il faut donc consacrer à cette difficulté toutes la réflexion et les efforts dont on est capable.

C'est là le point clé : il faut refuser toute indexation passe-partout, bonne à tout bonne à rien[20]. **L'indexation ne peut être utile que pour autant qu'elle est appropriée hic et nunc, à un corpus particulier**. C'est la seule perspective qui évite toutes les illusions (pseudo-)universalistes.

L'expérience et plus encore les discussions montrent qu'il existe, par-ci par-là, des bases de données de type enseignement-recherche qui donnent satisfaction. Presque toujours, il s'agit de bases pour lesquels les « utilisateurs » sont en même temps les « fournisseurs » ; sachant ce qu'ils y ont mis, et comment, ils ont une idée assez précise de ce qu'ils peuvent y trouver et comment. A ma connaissance (restreinte, malheureusement), la structure la plus efficace s'inspire de deux considérations : 1. des champs obligatoires en nombre très très limité, 2. un champ « description libre », en format libre, que bien entendu le logiciel pourra indexer et rendre très accessible[21]. La difficulté principale réside (une nouvelle fois), dans le choix et la définition des termes à utiliser dans les champs obligatoires.

Avant de songer à examiner les procédures d'indexation pertinentes dans le cas d'un corpus de textes latins, il faut se demander quels critères peuvent permettre de juger un mode d'indexation.

En gros, j'en vois trois.

☞ Le premier est l'univocité. Aliis verbis : la combinaison des codages (champs) sera différente pour TOUS les éléments du corpus, il n'y aura pas deux textes indexés de la même manière. Ce qui va de soi pour les champs « identité » (auteur-titre-date), mais non pas pour les champs « description ».

☞ Le second est l'équilibre. Qui résulte de considérations de technique statistique. Il faut éviter tout autant les mots-clés qui ne s'appliqueraient qu'à quelques unités, que ceux qui couvriraient disons plus du cinquième ou du quart des occurrences, et il faut tenter de faire en sorte que les effectifs correspondant aux divers mots-clés (mutuellement exclusifs) d'un champ renvoient à des effectifs voisins. C'est en général à peu près impossible, mais il faut y tendre.

☞ Le troisième est la facilité d'indexation, sachant que les procédures seront variables, et qu'il faut impérativement que plusieurs personnes indexent de la même manière.

---

20 Un corpus de textes latins ne concerne qu'un public insignifiant par rapport aux pratiques courantes sur internet ; et ce corpus n'est pas destiné à l'agrément, ni à l'information retrieval ordinaire, il est là pour servir de base à des opérations de recherche, qui elles-mêmes sont destinées à sortir des routines et des impasses du sens commun ; on ne voit pas comment, dans ces conditions, un balisage prétendument adapté à toutes les langues et à toutes les époques pourrait être d'une quelconque utilité.

21 Ce point du champ rempli librement est crucial ; il sort complètement des pratiques ordinaires des systèmes xml, qui tendent à normaliser ; mais le système xml est cependant assez souple pour en permettre l'usage sans aucune restriction. Les techniques informatiques d'exploitation d'un ensemble de courts textes (genre « abstracts »)sont très développées, la seule difficulté est de choisir celles qui sont le plus appropriées étant données les spécificités de cet ensemble.

Enfin, la question des méthodes. Sans doute peut-on, ou faut-il, imaginer des procédures « mixtes », c'est-à-dire la combinaison d'une intervention d'un opérateur et d'un procédé d'indexation automatique (type : text-mining, variante : clustering). Ce qui, entre parenthèses, nécessite presque l'utilisation du critère 2 énoncé ci-dessus. On n'indexera jamais un corpus de plusieurs milliers de textes si l'on doit le faire entièrement « à la main » ; il est impossible d'imaginer une indexation ultra-élémentaire en moins de deux minutes (et encore...), et ce genre d'exercice est intenable plus de trois heures par jour. Soit un grand maximum de 100 textes par jour (sans doute irréaliste). Soit 500 par semaine, soit 20 semaines à temps plein (et sans rien faire d'autre) pour 10000 textes... Inversement, des procédures éprouvées de machine-learning peuvent permettre, avec un temps de préparation raisonnable, de coller des étiquette fiables sur 4/5 des textes, quel que soit l'effectif. Il en resterait 2000 à vérifier, ce qui pourrait correspondre à quatre semaines.

On doit tout de même savoir que l'on est bien loin d'être dans la situation la plus inconfortable ! car le text mining est susceptible d'apporter un outil efficace : l'iconographie n'a rien d'équivalent....

Maintenant, s'agissant d'une base de textes latins, on doit tenir compte des outils spécifiques existants. Tous les érudits savent qu'un des moyens classiques d'identifier un texte, notamment dans un manuscrit, consiste à considérer son incipit et son explicit, et à rechercher ceux-ci dans les répertoires spécialisés : dans les manuscrits, la majeure partie des textes ne portent ni titre ni nom d'auteur[22]. L'IRHT-CNRS, au cours des années, a accumulé un répertoire sans équivalent ; mais cette institution a perdu la maîtrise de sa diffusion dans des conditions absurdes. Cependant, il existe au moins un CD qui comporte un fichier plus ou moins exploitable. Question à clarifier[23]. Plus récemment, deux membres de l'équipe de lexicographie latine du même IRHT, Bruno Bon et Renaud Alexandre, ont numérisé l'Index Scriptorum, qui répertorie la totalité des éditions de textes latins utilisés pour la confection du Novum Glosssarium, et l'ont transformé en base de données MySQL. Ce répertoire est en ligne, en accès libre[24]. On dispose ainsi d'un répertoire de plus de 11000 titres de textes latins couvrant la période 8e-12e siècles. C'est manifestement une base d'information unique, en ordre de marche, qui devra être utilisée pour l'aspect « identité » de l'indexation.

Cette question décisive de l'indexation est aujourd'hui un **front pionnier** de la recherche historique.

Remarque brève sur la non-fixation des textes

Tous ces textes, qui bien souvent n'avaient ni titre ni auteur, ont encore moins de forme claire et indiscutable. C'est un des soucis majeurs de la philologie, depuis ses plus lointaines origines. Le sens commun distingue, ou croit distinguer, de « bonnes » et de « mauvaises » éditions, en jugeant celles-ci sur le choix de variantes retenues comme texte principal. On finit par se perdre dans des méandres obscurs : on ne cite plus les manuscrits qui portent telle ou telle variante, mais le nom de l'éditeur qui a choisi l'une ou l'autre... Il y a des éditions où la graphie est fortement normalisée, selon des principes d'ailleurs très variables, et d'autres où l'on conserve une certaine souplesse.

Cet aspect n'est pas isolé ou anecdotique. En surface, les variations locales, puisque chaque copiste soit se trompe soit cherche à améliorer le texte ; au plan des textes, absence de titre et d'auteur, mais des indications tout à fait différentes de celles auxquelles nous sommes habitués, et relevant bien plus du contenu que de considérations extérieures ; au plan du sens, les structures profondes de production dudit sens étant complètement différentes des nôtres, et utilisant des critères de choix des mots et de structuration des champs sémantiques bien distincts des nôtres...

Un minimum d'imagination suffit pour voir que l'informatique peut offrir des moyens algorithmiques de gérer toutes ces variations, sans introduire de choix a priori, mais seulement en indiquant l'origine des variantes (en général un manuscrit ou un groupe de manuscrits). C'est une perspective qui mérite des recherches et la création de structures de fichier spécifiques. Ce n'est cependant pas une priorité, du moins par rapport au projet présenté ici.

22 Cet aspect visible « en surface » est en fait l'indice d'une structure profonde : NOTRE notion de « texte » n'existait pas au Moyen Age. Ludolf KUCHENBUCH & Uta KLEINE (éds), *'Textus' im Mittelalter*, Göttingen, 2005.

23 Ce répertoire a été constitué en grande partie à partir de catalogues de manuscrits, il recense ainsi une énorme quantité de textes inédits, ni datés ni attribués.

24 http://ngml.irht.cnrs.fr/index/index-scriptorum.php

Quelques considérations sur l'organisation pratique d'une base de données textuelles latines

Le risque majeur est de se lancer dans la fabrication d'une usine à gaz, ingérable, et qui reste toujours plus ou moins à l'état de prototype. Il faut viser la plus grande simplicité, condition pour que la base reste facile à faire fonctionner. La simplicité est d'ailleurs un gage de solidité. On pourra toujours se demander comment ajouter à une structure simple des fonctions plus ou moins sophistiquées, l'inverse ne serait pas possible.

Existe-t-il des modèles, ou tout au moins des sites dont on pourrait utilement s'inspirer ? Comme je l'ai écrit plus haut, je trouve proche de la perfection le site des CBMA, mais il est vrai que, dans ce cas, le nombre limité de fichiers à mettre en ligne évite les pièges de l'indexation. On retombe sur les sites de diffusion de fichiers-textes libres ; voici quelques années, le site italien www.liberliber.it était lui aussi voisin de la perfection, il a subi une dérive qui en fait un support publicitaire désagréable ; voir les diverses branches nationales du « Project Gutenberg », certainement l'ancêtre du genre (1971). Maintenant, le projet wikisource, qui paraît plus solidement organisé, plus structuré. Cependant, dans ce dernier cas, la volonté de faciliter (?) la lecture en ligne et de rendre impossible le téléchargement simple d'une œuvre complète est à mes yeux une erreur majeure, que je ne sais comment expliquer.

Ne perdons pas de vue un aspect quantitatif élémentaire. On a vu que les plus gros ensembles « pèsent » quelques centaines de mégaoctets. Pendant des années, j'avais généralement avec moi une disquette (1,4 méga...) qui contenait une vulgate complète ! Imaginons des fichiers-textes de bonne taille, disons 500k. 20000 fichiers valent 10000 méga, soit 10 gigas. En prenant une marge irréaliste, on n'arrive que difficilement à 50 gigas (même en construisant des index qui multiplient par trois l'espace nécessaire) ; or on vend dans tous les supermarchés des disques durs externes d'un téra, pour moins de 100 €. Bref, le stockage d'une base de textes quasi monstrueuse ne remplit pas un outil aujourd'hui commun. On ne saurait bien sûr en dire autant d'une base de fichiers-images. Donc, si l'objectif est de construire une base de textes aussi copieuse que possible, il faut savoir dès le départ que la question de l'espace de stockage ne se pose pas. Cette considération devrait faciliter la recherche et la création de sites-miroirs, objectif qui devrait figurer parmi les plus importants : la démultiplication des lieux de conservation est le gage principal (sinon unique) de la longévité[25] de fichiers électroniques.

Le démarrage du projet ne peut être assuré que par une petite équipe, organisée et déterminée : le besoin en crédits est limité, mais le besoin en compétences de latiniste-informaticien est crucial, car seuls des latinistes peuvent ne pas s'engager presque immédiatement dans diverses impasses[26] : aussi bien le matériau que les objectifs de l'opération ne correspondent à rien de ce que connaissent les informaticiens, même très compétents. Cependant le développement, lié à la diffusion et à l'usage, suppose obligatoirement que l'on passe très rapidement à un projet collaboratif, international cela va de soi, très ouvert. C'est une grande banalité de rappeler que la clé du succès de Wikipedia a été la facilité d'intervention de n'importe qui. Sans en aller exactement jusque là, il faut cependant tout faire pour que n'importe quel latiniste (il n'en reste pas tant que ça...) puisse collaborer sans avoir à se soumettre à une procédure particulière. Quitte, bien entendu, à ce que toute intervention ne devienne définitive qu'après vérification par un ou plusieurs « administrateurs ».

Il semblerait très souhaitable que tous les textes soient donnés sous une forme graphique à peu près commune (choix de la distinction ou non du i-j, u-v, etc.). Indispensable en tout cas si l'on veut créer un système d'indexation global. Simple affaire de tokenisation, un petit système d'instructions simples devrait permettre à tous les volontaires d'obtenir aisément des fichiers conformes. J'ai évoqué plus haut la possibilité d'un champ (dans l'en-tête) en format libre. Celui-ci, de toute manière, ne pourra pas être rempli par

25 Je préfère parler de longévité plutôt que de pérennité ; il est certainement plus important d'assurer la pérennité des données que celles des fichiers eux-mêmes. On touche ici la question de la complémentarité nécessaire papier/électronique, qui déborde de beaucoup le cadre du présent papier.

26 On effleure ici une autre question générale, en général posée totalement de travers par des individus ne disposant pas d'une expérience concrète, celle dite de la « pluridisciplinarité » ; d'une manière très globale, dans la plupart des domaines scientifiques, celle-ci n'est efficace que dans la tête d'une seule et même personne ; en matière scientifique, les avantages de la division du travail sont fort limités, les risques de dérive négative, considérables.

procédure. Mais c'est là que tout un chacun pourra progressivement enrichir le stock d'informations disponible, sans avoir à se soumettre à un formalisme contraignant.

L'essentiel, comme je l'ai dit, et j'insiste, est la possibilité de télécharger très facilement tous les textes disponibles, individuellement ou peut-être même par blocs. Mais cela n'empêche en aucune manière[27], bien au contraire, de proposer un accès structuré aux textes, sous la forme d'un système de gestion de bases textuelles à proprement parler ; ici encore le modèle est le site des CBMA, qui propose l'ensemble des textes interrogeables dans une base de données PhiloLogic.

Se pose donc la question du choix d'un logiciel approprié. 1. je suis tout à fait hostile à l'idée de créer un logiciel de plus, c'est une impasse, même dans les meilleures conditions (i.e. si l'on dispose d'une équipe d'informaticiens compétents et efficaces) ; 2. il n'y a pas pléthore de grands logiciels open source : je ne connais que trois candidats d'allure sérieuse, PhiloLogic (Chicago), Textométrie (Lyon) et ASV Browser-Toolbox (Leipzig). (http://philologic.uchicago.edu/ - http://textometrie.ens-lyon.fr/spip.php?rubrique7 - http://corpora.uni-leipzig.de/download.html et http://wortschatz.uni-leipzig.de/~cbiemann/software/toolbox/index.htm).

Une discussion approfondie des avantages et inconvénients des trois logiciels demanderait un papier complet. Chacun des trois dispose de solides atouts, et tout chercheur intéressé par les bases de données textuelles a tout intérêt à les connaître assez bien tous les trois, au moins comme sources de réflexion et d'inspiration. A mes yeux, la facilité avec laquelle on peut les modifier et/ou les « brancher » sur d'autres logiciels (statistiques, graphiques, etc) sont des points à prendre en compte en première ligne, car aucun logiciel n'est capable de répondre par avance à tous les besoins et à toutes les idées de n'importe quel chercheur[28] ; or il s'agit bien, s'agissant de textes latins, d'un outil de recherche, pas d'un simple, ennième, moteur d'indexation standard.

Rappelons à toutes fins utiles que l'insertion de liens dans une page html est un jeu d'enfants. Créer, dans le cadre d'un tel site, une page de liens ne réclame aucun effort technique. La difficulté serait plutôt de commenter chaque lien, car les sites où s'alignent sans distinction des dizaines ou des centaines de liens sont à peu près inutilisables. Et il est malaisé de tenir à jour une telle page, car les liens ne cessent de se modifier, et rien de plus irritant que les liens cassés.

27 Les deux problèmes techniques sont complètement distincts ; au demeurant, on peut parfaitement gérer une base de textes téléchargeables sur une machine, et installer un logiciel de consultation sur une autre, le tout de manière transparente, au travers de la même url. Notons aussi qu'il faudrait envisager la possibilité de télécharger une version des textes déjà prétraitée, c'est-à-dire lemmatisée et postaggée.

28 Je signale à tout hasard un petit logiciel libre open source (en python, fonctionne sous linux et W), qui établit très commodément des statistiques de base sur un texte ou un groupe de textes : http://neon.niederlandistik.fu-berlin.de/de/textstat/ Je sais par expérience que les utilisateurs (en particulier les débutants) apprécient beaucoup TextSTAT.

Conclusion

Quelques points clés.

1. tous les textes auxquels on s'intéresse ici font partie du domaine public. Toutes celles et tous ceux qui ont à cœur l'avenir des « études latines » doivent considérer comme un impératif moral de première importance de mettre un frein énergique aux gesticulations comminatoires et mensongères d'une poignée d'éditeurs avides.

2. tout différencie un tel corpus de ce que les linguistes et les informaticiens ont l'habitude de traiter sous ce nom : le corpus est définitivement clos, son ampleur est microscopique par rapport à ce que l'on manipule aujourd'hui, mais surtout son sens est opaque, et la mise en ligne a bien moins pour objet de permettre de rechercher des citations, que de fournir le matériau nécessaire aux perspectives actuelles de la recherche.

3. La situation actuelle des « mises en ligne » de textes latins se caractérise surtout par sa dispersion et son hétérogénéité. Il faut chercher un moyen d'y remédier.

4. La création d'un site regroupant le maximum de ce qui est disponible ne demande que des moyens matériels très réduits et ne pose aucun problème technique particulier, mais des problèmes intellectuels ardus, que seuls des latinistes sont en mesure d'affronter.

5. Un tel projet doit être orienté dans le sens de la création d'une dynamique générale, bien sûr internationale, sur la base d'une collaboration facile ; la création rapide de sites-miroirs pourrait y contribuer assez aisément.

Ne pas oublier l'adage médiéval : **scientia donum dei, non venditur**.

Alain GUERREAU CRH-CNRS avril 2011

[je mets le présent texte dans le *domaine public*, toute liberté est donnée de l'utiliser, de le reproduire, de le diffuser, sans aucune restriction. Eine deutsche Übersetzung wäre sehr willkommen. A.G.]