Colloque "Le patrimoine à l'ère du numérique : structuration et balisage" (Université de Caen, MRSH, 10-11 déc. 2009)

Leçons du projet *CAMENA – textes latins de l'ère moderne* : La mise en valeur des textes signalétiques du livre ancien

Wolfgang Schibel – projet CAMENA : Latin Texts of Early Modern Europe

Résumé en français:

L'expérience du projet CAMENA – textes latins de l'ère moderne nous a amené à ne saisir plus, en mode texte, que les petits textes signalétiques du livre ancien – les intertitres, les têtes, les mots clef en marge de page, les tables des matières. Une simple transcription des formes souvent variées ne suffisant pas, la rectification et le balisage de ces textes, qui sont protéiformes en leur typographie, leur orthographie, leur lexique (variation par synonyme!), leurs abréviations, leur ordre des mots, leur organisation des entrées d'index et leur mise en page, sont confiés à des latinistes. Le lecteur de CAMENA dispose de dictionnaires, de répertoires et de compilations du savoir du public érudit moderne (Thesaurus eruditionis). Les difficultés multiples que les éditions originales d'écrits latins modernes présentent tant au lecteur qu'au traitement informatique, ne semblent pas permettre des solutions rapides.

Résumé en anglais:

Our ten-years' experience in creating the online library CAMENA - Latin Texts of Early Modern Europe finally determined us to transcribe but the headings, tables and indexes of the old editions we were digitising. The enormous amount of variation present in early modern Latin texts – in typography, spelling, vocabulary (use of synonyms!), abbreviation, word order, organisation of index entries and layout – demands a treatment by expert hands, i.e. by Latinists capable of drawing up reliable, enhanced tables of contents and indexes. A cumulation of tagged entries can provide detailed subject and key-word access to the entire library. Moreover, a handy reference collection such as our Thesaurus eruditionis seems to be an indispensable aid for all but the most expert readers.

Mesdames et Messieurs,

Je vous remercie de bien vouloir examiner avec moi les enjeux de CAMENA – d'une bibliothèque en ligne qui est à égale distance d'un projet d'édition élaborée et d'une collection incohérente de simples reproductions en mode image. Aujourdh'hui, le champ de la numérisation est dominé par de grandes institutions politiques, scientifiques et culturelles, qui ont lancé des programmes de numérisation de masse. Les visées de ceux-ci ne diffèrent pas beaucoup de celles de *Google Books*. Il s'agit de numériser à toute allure toutes les éditions anciennes et d'en dériver des versions "texte" en utilisant des logiciels de reconnaissance optique de caractères (OCR). L'expérience de CAMENA nous permet d'identifier les difficultés caractéristiques présentées par le patrimoine latin imprimé du XVe au XVIIIe siècle – des difficultés qui ne seront pas surmontées par une démarche purement technologique. Le traitement de ces textes doit être muni des données de langue et de fait qui constituent l'érudition humaniste.

CAMENA à vrai dire n'a pas achevé de résoudre les problèmes particuliers du patrimoine latin d'une façon qui se prête à être adaptée par la plupart des entreprises pareilles. D'autre part, notre projet semble avoir réussi à attirer beaucoup de lecteurs intéressés. Il y a peut-être trois facteurs peu communs qui ont contribué à cet effet: 1. Étant de dimension modeste et d'une disposition qui peut être embrassée d'un seul coup d'œil, CAMENA évite de

détourner et déprimer les lecteurs par une surabondance peu organisée. 2. Présentant sans les rétrécir des pages en mode image bien lisibles et souvent belles, CAMENA permet une lecture agréable. 3. En communiquant l'orientation de leur travail et en insérant des informations utiles à l'intelligence des textes présentés, les éditeurs de CAMENA ont établi un interface entre les lecteurs d'aujourd'hui et les auteurs de l'ère moderne, qui aide à motiver et satisfaire les lecteurs.

Il me faut avouer que CAMENA n'est pas tout à fait homogène. J'exposerai les divers motifs et problèmes qui ont formé et modifié notre projet dès son début il y a onze ans. Les limites de ma compétence ne m'autorisent pas à présenter en détail les procédés automatisés de correction, normalisation et balisage structurel du texte continu. Mes collègues ont developpé quelques outils à rectifier et baliser les transcriptions. Comme conservateur de livres anciens et éditeur de textes latins de l'ère moderne, j'étais responsable de l'orientation et du contenu de nos projets.

CAMENA – textes latins de l'ère moderne
<http://www.uni-mannheim.de/mateo/camenahtdocs/camena_e.html>

présente près de trois cent mille pages en mode image, dont la moitié environ est également disponible en mode texte.

La proposition initiale de CAMENA date de l'an 1998. Dès le mois de mars 1996, une petite équipe informelle (Heinz Kredel, Emir Zuljevic, Wolfgang Schibel) avait mis en ligne une série de livres anciens de l'Université de Mannheim présentés en mode image. C'est ***MATEO – Altes Buch***,
<http://www.uni-mannheim.de/mateo/epo.html>

Nous étions parmi les premiers à établir une telle bibliothèque. Notre visée, alors, était de présenter à un large public des sources illustrant la culture littéraire et artistique de l'ère moderne. Aux pages reproduites nous ajoutions une introduction et une table des matières, parfois aussi un sommaire ou *argumentum* basé sur notre lecture de l'ouvrage. Voyez e.g.
<http://www.uni-mannheim.de/mateo/desbillons/esop.html>
et
<http://www.uni-mannheim.de/mateo/desbillons/aport.html>

On s'adressait au public érudit général plutôt qu'aux chercheurs spécialistes. On ne prenait pas en considération les métadonnées ou bien une structuration élaborée.

Notre premier projet soutenu par la DFG (Deutsche Forschungsgemeinschaft, la fondation allemande de la recherche) était ***Poemata*** – un corps de poésie néolatine provenant des pays allemands, qui comprend environ 65.000 pages imprimées:
<http://www.uni-mannheim.de/mateo/camenahtdocs/camenapoem_e.html>

Ce choix était motivé par ma collaboration à l'édition d'une anthologie de la poésie néolatine d'Allemagne, *Humanistische Lyrik*. C'est un gros volume de la *Bibliothek Deutscher Klassiker*, publié en 1997. Nous voulions disséminer la connaissance de cette poésie, qui jusqu'alors était ignorée même par les érudits. Outre les pages originales reproduites en mode image nous mettions en ligne le texte total saisi au clavier. Celui-ci remplit plusieurs fonctions: Aux moteurs de recherche et aux chercheurs il rend accessible le texte intégral, mot par mot ou phrase par phrase; il aide les lecteurs moins avertis à déchiffrer une graphie souvent difficile; il fournit une base à qui veut éditer et commenter un texte de façon didactique ou

scientifique; il pourra livrer, éventuellement, des matériaux à construire une banque de données du latin de l'ère moderne.

Dès le début nous ajoutions un balisage conforme à TEILITE. Je donnai alors une instruction détaillée à nos typistes qui étaient des jeunes latinistes résidant dans diverses régions d'Allemagne:
<http://www.uni-mannheim.de/mateo/camenatools/docs/textein.htm>

Nos règles déterminaient non seulement le balisage, mais aussi la constitution et rédaction du texte. On ne devait pas, par exemple, transcrire les tables alphabétiques ou les notes manuscrites. On devait normaliser la graphie – non pas des noms propres, mais des *appellativa*, des noms communs.

Le balisage des textes de la section *Poemata* ne vise pas à signaler la mise en page et la variation typographique de l'édition originale. Nous recommandons aux usagers de CAMENA d'étudier la page originale, la transcription n'étant qu'un support de travail. La préférence donnée à l'impression ancienne se manifeste par la qualité supérieure de la reproduction et dans l'affichage nu et sans mélange (e.g. d'options de maniement, de données signalétiques et structurelles) de l'image fac-similé. Les réactions de nos lecteurs confirment que ce design est bien estimé. C'est mon collègue Emir Zuljevic qui, dès le début de MATEO (1996), a proposé, réalisé et maintenu cette présentation.

Il y a un balisage ‚critique' pour ainsi dire, qui signale ou des corrections nécessaires du texte imprimé, ou ses variantes orthographiques remarquables, ou bien nos doutes sur le texte transmis. Notre balisage sémantique sert à classifier les noms propres et à expliquer les abréviations et les dates. Voyez e.g.
<http://www.uni-mannheim.de/mateo/camena/bald1/books/baldepoemata1_1.xml>
et le même texte en format HMTL:
<http://www.uni-mannheim.de/mateo/camena/bald1/books/baldepoemata1_1.html>

Mon collègue des années 2001 à 2005, Rüdiger Niehl, a utilisé ce balisage pour créer des index alphabétiques des noms propres qui se trouvent dans l'oeuvre immense du grand poète Jacobus Balde:
<http://www.uni-mannheim.de/mateo/camena/baldeindex.html>
Il y a l'index des noms bibliques et des saints, l'index des noms du mythe, l'index des autres noms de personne, l'index des noms d'ethnie ou de peuple, l'index des noms de corporation ou société, et enfin l'index des noms géographiques. Ces index témoignent de la vaste érudition de l'auteur.

Rüdiger Niehl a aussi créé un *index metrorum* de quelques éditions, par exemple:
<http://www.uni-mannheim.de/mateo/camena/bald9/baldetillius-met.html>
Comparez s. v. p. les index correspondants des écrits d'Elizabeth Jane Weston, poétesse latine de Bohême, originaire d'Angleterre:
<http://www.uni-mannheim.de/mateo/camena/weston1/westonparthenica-nam.html>

Seule l'édition de l'oeuvre de Westonia a été pourvue d'un *index argumentorum*:
<http://www.uni-mannheim.de/mateo/camena/weston1/westonparthenicaana.html>
qui utilise des termes d'indexation thématique et littéraire que j'ai recueillis pour l'analyse de la poésie néolatine:
<http://www.uni-mannheim.de/mateo/camenahtdocs/ubi.html>
Or, une telle analyse demande une lecture approfondie du texte. Quand je me mis à l'entreprendre, j'avais déjà étudié l'oeuvre de Westonia. Autrement, c'est guère possible de

dresser en peu de temps un index des thèmes, motifs et genres présents dans une édition de quelque volume.

La saisie et la rédaction des textes poétiques de CAMENA par des jeunes latinistes ne durait que deux ans, puisque la livraison de textes transcrits et balisés était presque toujours tardive. Il nous fallait désormais employer des copistes Chinois. Naturellement, des copistes qui ne savent pas le latin ne sont pas capables de normaliser et baliser le texte. De plus, ils ne peuvent déchiffrer la typographie que si elle est usuelle et distincte.

Les projets qui suivaient *Poemata*
- la collection ***Historica & Politica*** (2004-2006/2008, plus de 60.000 pages):
<http://www.uni-mannheim.de/mateo/camenahtdocs/camenahist_e.html>
et la collection ***CERA*** (Lettres des savants, 2006-2008, 55.000 pages):
<http://www.uni-mannheim.de/mateo/camenahtdocs/cera_e.html> -
en multipliant nos textes saisis au clavier, ne nous permettaient plus la rédaction intellectuelle et manuelle de tous les textes. Une production annuelle d'environ 15 mille pages en mode texte (d'un total de 35 mille pages en mode image) coupait court nos ambitions de balisage.

Grâce au Projet Persée (*Perseus Project*, Tufts University, Medford, Mass.) nous pouvions utiliser la banque de données MORPHEUS qui contient toutes les formes régulières du vocabulaire latin répertorié par Lewis & Short: *A Latin Dictionary*. Il y restent cependant bien des difficultés: Dans le dictionnaire de Lewis & Short est enregistré le lexique de l'antiquité. Ni les mots et noms propres latins ou latinisés du moyen-âge et de l'époque moderne ni les variantes orthographiques qui se trouvent à foison dans les livres imprimés avant 1700 figurent dans MORPHEUS. Par conséquent, la petite équipe de CAMENA ne pouvait corriger de façon systématique qu'un tiers, peut-être, des textes saisis au clavier en Asie orientale. Se sentant comme perdu dans un immense taillis, on n'appliquait désormais qu'un balisage structurel tout en tentant de réduire, à l'aide de MORPHEUS, les fautes de transcription et la variation orthographique. Dans le *header* de chaque fichier l'état du texte copié est indiqué par une notice, par exemple: „*structural tagging complete - no semantic tagging - MORPHEUS spell-check performed*". Ces développements ne s'accordant pas à notre instruction initiale, celle-ci fu partiellement retranchée. Somme toute, nos efforts de familiariser de jeunes latinistes avec l'écriture latine des modernes, de pousser la graphie normalisée du latin et de baliser les noms propres qui se présentent dans la littérature néolatine, après peu d'années ont cédé à la nécessité d'une manière d'agir plus économique.

En 2001 nous initiâmes la collection ***Thesaurus Eruditionis***
<http://www.uni-mannheim.de/mateo/camenahtdocs/camenaref_e.html>
qui devait représenter l'érudition de l'époque moderne et offrir au lecteur de la poésie néolatine les informations dont disposaient les contemporains. Cet ancien salon de lecture (85.000 pages) contient des ouvrages de référence qui rendent accessible le vocabulaire spécial et le savoir des disciplines.

Afin d'activer l'usage de cette source d'information, nous voulions ensuite répertorier les mots clef des ouvrages les plus riches du *Thesaurus Eruditionis* en donnant leurs définitions, leurs synonymes, leurs compléments et satellites (en anglais: *collocations*). Ce dépouillement d'ouvrages de référence latins devait créer un index général donnant des informations terminologiques de base, en expliquant en peu de mots le concept, la chose ou le fait signifié par le mot clef, et en étalant le vocabulaire voisin pour orienter la recherche.

C'était le dessein du projet ***Termini - Vocabulaire latin du savoir moderne*** qui démarra en 2004:
<http://www.uni-mannheim.de/mateo/termini/index_e.html>
On espérait alors qu'un tel répertoire de données sémantiques pourrait aussi faciliter le balisage des noms propres et du vocabulaire spécial – ou au moins dédommager le lecteur du défaut de balisage sémantique. On anticipait donc un procédé analogue à celui du projet *SCALE* du *Perseus Project*, qui développe des outils à effectuer des hyperliens qui associent aux mots clef du texte non balisé les nœuds d'information correspondants fournis par des fichiers d'autorité.

Hélas, l'ampleur de notre plan, la diversité de nos tâches et l'instabilité de notre petite équipe qui manquait d'un informaticien permanent, nous empêchèrent d'atteindre notre but. Un grand embarras imprévu, c'était la variation permanente des formes de l'expression chez les écrivains humanistes marqués par la rhétorique. Ils choisissent un synonyme pour éviter la répétition (*copia verborum*), ils changent l'ordre des mots pour ne pas ennuyer le lecteur, ils varient les abréviations et la ponctuation, ils aiment le détour et la surprise - non seulement dans le discours, mais aussi dans le dictionnaire, dans les tables des matières et dans les index alphabétiques. Notre plan d'exploiter l'uniformité anticipée de la série des données, mais aussi des mots ou formules de classement, était illusoire. Fin Septembre 2006, d'ailleurs, c'était le terme de ma retraite.

Dès 2008, le projet ***Lemmata*** suit le chemin raboteux que *Termini* a tracé.
<http://www.uni-mannheim.de/mateo/termini/db/index.php>
Son unique collaborateur est employé à mi-temps. Il est en train de construire un index des mots vedette expliqués dans une dizaine de dictionnaires généraux et spéciaux du *Thesaurus eruditionis*. Les catégories lexicologiques relevées n'embrassent pas l'échelle conçue par *Termini*. À l'heure actuelle, le dépouillement n'est pas encore assez avancé pour déployer le potentiel de la cumulation.

Il y a un an, nous établîmes une section d'éditions rares de l'humanisme italien de la renaissance qui s'appelle ***ITALI*** (25.500 pages):
<http://www.uni-mannheim.de/mateo/camenahtdocs/itali.html>
Un mécène généreux (*Istituto Italiano per gli Studi Filosofici*) soutenait la numérisation en mode image, tandis que je m'engageai à saisir au clavier les tables et intertitres des éditions.

Il y a quatre ans, en écrivant sur les sommaires et index des livres anciens (*Summaria ac indices*, dans mélanges Wilhelm Kühlmann: *Strenae nataliciae*, Heidelberg: Manutius Verlag, 2006), je compris enfin qu'une bibliothèque en ligne profitera beaucoup de la transcription de ces textes succincts dont l'effet peut être augmenté par un balisage intensifié et des regroupements. C'est là la proposition principale que mon discours doit déployer.

Permettez-moi de répéter la première section du résumé que vous connaissez:

A. Les développements de notre projet nous amènent, enfin, à ne saisir plus, en mode texte, le texte intégral, mais seulement les petits textes indicateurs du livre ancien. Ce nouveau dessein se prête à une bibliothèque numérique de dimension moyenne.

Je procède maintenant en expliquant successivement les autres propos de mon résumé.

B.1. Les aspects spécifiques de l'édition latine de l'ère moderne demandent un traitement adapté aux besoins des lecteurs d'aujourd'hui.

La mise en ligne d'un livre ancien numérisé est plus qu'une forme d'archivage, c'est un acte de dissémination. Une grande partie de notre patrimoine latin ne peut intéresser que les spécialistes, je le concède. Mais beaucoup d'ouvrages latins de l'ère moderne contiennent des informations importantes ou curieuses et des illustrations de qualité supérieure. La connaissance du latin, chez nous, étant plus médiocre qu'autrefois, on doit aider le lecteur à les découvrir.

B.2. Il s'agit de textes érudits qui comprennent des éléments de plusieurs langues et écritures, des notes en marge et en bas de page, avec une mise en pages bien complexe.

Voyez e.g. la composition lardée des pages de la grande grammaire de Gerardus Johannes Vossius:
<http://www.uni-mannheim.de/mateo/camenaref/vossius/vol4/jpg/s0146.html>

Les lecteurs contemporains estimaient l'apport d'une typographie diversifiée qui, en classant les éléments du texte, soutenait la lecture rapide.

Dans la plupart des livres destinés aux érudits sont entremêlés des éléments grecs en écriture grecque. Or, celle-ci n'est pas tout-à-fait de la forme typographique que nous connaissons; elle rappelle le style cursif des manuscrits des savants byzantins qui enseignaient le grec aux humanistes italiens. Aujourd'hui, même les chercheurs sont parfois incapables de la déchiffrer.

Voyez s. v. p. l'article „Achilles" du *Lexicon universale* de Hofmann (1698):
<http://www.uni-mannheim.de/mateo/camenaref/hofmann/hof1/s0039a.html>

Au lieu de garder l'exclusivité du savoir de l'helléniste, il faudrait présenter les éléments grecs, qui se trouvent interposés dans la plupart des textes latins modernes, en transcription exacte et lisible. J'ai élaboré un tel schéma de transcription. La page initiale de notre section *ITALI* offre un hyperlien qui conduit à la page „*Polytonic Transcription of Ancient Greek*":
<http://www.uni-mannheim.de/mateo/itali/af/transcription.pdf>

Notre édition de *Varia Philippi Beroaldi Opuscula* (1515?), par exemple, dans sa table fait usage de ce schéma:
<http://www.uni-mannheim.de/mateo/itali/autoren/beroaldo_itali.html>

Mais rentrons aux textes latins.

B.3. Ces textes fourmillent de variations orthographiques, d'abréviations, références et citations peu réglées.

La diversité orthographique héritée des manuscrits latins du moyen-âge persistait longtemps. Elle gêne le lecteur d'aujourd'hui qui est peu familiarisé avec la langue latine et ignorant de cette variation. Le plus souvent, la diversité n'est pas de valeur phonétique ou stylistique, mais indifférente. Nos copistes d'Extrême Orient la conservent fidèlement, bien sûr, mais nos lecteurs n'en bénéficieront pas.

La plupart des abréviations qui se trouvent dans les livres anciens sont ignorées par les lecteurs d'aujourd'hui. Il convient qu'elles soient balisées et développées. CAMENA a numérisé le répertoire très utile
Winiarczyk, Marek: *Sigla latina in libris impressis occurrentia*. Wratislaviae: Univ., 1995:
<http://www.uni-mannheim.de/mateo/camenaref/siglalatina.html>

Cependant, l'homonymie de beaucoup d'abréviations demande que l'on choisisse une interprétation convenable au contexte. C'est parfois difficile.

Les références bibliographiques dans les livres anciens d'ordinaire sont brèves, signalant un texte plutôt qu'une édition spécifique. Les noms propres et les mots de titre sont souvent donnés en abrégé. Le projet *Termini*, en vue de créer un répertoire d'ouvrages fort connus qui puisse servir à l'identification de références bibliographiques insuffisantes, a numérisé quelques volumes de *historia litteraria*, e.g. le catalogue de la bibliothèque de l'université de Leiden imprimé en 1716,
< http://www.uni-mannheim.de/mateo/camenaref/leiden.html>
et la *Bibliotheca vetus et nova* de Georg Matthias König (1678),
<http://www.uni-mannheim.de/mateo/camenaref/koenig.html>
qui recense tous les auteurs connus au bibliothécaire de l'université d'Altdorf (près de Nürnberg).

B. 4. Les formes des noms de personne ou de lieu y varient.

C'est un problème de vastes dimensions. La coexistence de formes vernaculaires et latines entraîne une multiplication des formes désignant une seule personne ou un seul lieu, puisque il en résulte une quantité de formes mixtes. En outre, il y a les formes plus ou moins abrégées des noms propres composés. Il y a aussi la tendance d'helléniser, si non le nom entier, à moins une partie ou la graphie seule (e.g. en substituant un *y grec* pour un *i*, un *th* pour un simple *t*). De plus, les nations de l'Europe moderne avaient la coutume d'adapter à leur langue nationale les prénoms étrangers appartenant aux traditions communes d'origine biblique, antique, germanique ou celtique.

Il y a plusieurs banques de données très riches qui contiennent des notices d'autorité créées, pour la plupart, par la coopération des bibliothèques d'un pays. Il faut les mettre à la disposition des projets de numérisation et balisage qui en pourraient tirer profit.

C.1. Puisque la transcription, la rédaction et le balisage du texte intégral coûtent cher, nous choisissons de ne traiter plus (dans la section *ITALI*) que les petits textes d'accès.

Ce travail peut être exécuté par une seule personne qualifiée en philologie classique. Dans nos projets, les difficultés qu'entraînait l'emploi d'aides qui, souvent, ne connaissaient pas à fond le latin et manquaient d'application soutenue, étaient énormes. Distribuer les tâches, donner les instructions nécessaires, contrôler et corriger le travail accompli – tout cela nous coûtait trop de temps. Par contre, le travail d'une personne bien qualifiée comprendra tout - l'évaluation des textes d'accès, leur sélection, leur transcription partiellement normalisée, leur cumulation et regroupement, le balisage du texte d'accès ainsi constitué, la production d'index spéciaux. On évitera donc un chevauchement des efforts de plusieurs collaborateurs.

Les éditions qui présentent le savoir des disciplines, en général offrent des intertitres et tables très riches. J'estime que leur volume fait à peu près sept pourcent de l'impression, en moyenne. Cependant, vue la conformité fréquente des entrées qui figurent dans la table des matières (sommaire), en tête de chapitre ou d'une section quelconque, et dans la table alphabétique, il n'est d'ordinaire pas nécessaire de les saisir tous.

C.2. Ces textes d'indication sont le plus souvent des composants intégraux de l'ouvrage, très utilisés et estimés par les lecteurs contemporains.

Les imprimeurs ou éditeurs les annonçaient dans le titre même, parfois en pompe:
„*Cum amplissimis indicibus ...*" (Paris, 1552),
„*Omnia ... concinnis indicibus aucta*" (Rotterdam, 1693),
„*copiosissimis scripturarum, rerum, glossarum indicibus locupletata & illustrata*" (Paris, 1693).

En 1640, un savant allemand fait observer:
„Aujourd'hui l'on ne sortit guère un livre – écrit en latin ou en vernaculaire – qui ne soit pas pourvu de sommaire ou synopse, d'une liste des auteurs cités, d'un index des sujets, de notices marginales, de têtes de chapitre et d'autres éléments de cette sorte. Il n'est certainement pas besoin de les louer et recommander comme nécessaires et utiles."
(Bernhard von Mallinckrodt: *De ortu ac progressu artis typographicae dissertatio historica*. Cologne, 1640, Cap. XVII. *Comparatio antiquae et hodiernae typographiae. Quarta differentia*, p. 104.)

Au milieu du XVIe siècle, Conrad Gesner, dans son ouvrage *Pandectarum sive partitionum universalium libri XXI* (Zurich, 1548. *Liber I De grammatica, tit. XIII, ps. De indicibus librorum,* fol. 19v sqq.), recommande à l'écrivain savant de tirer profit des sommaires et index imprimés qui l'aideront beaucoup à compiler les informations nécessaires et les allégations convenables à son dessein.

Ce ne sont point des témoignages isolés. La technique et pratique des extraits, *ars excerpendi*, était une partie essentielle de la discipline scolaire. On s'accoutumait dans ses jeunes années à prendre note, en lisant, des mots ou phrases mémorables, et de les arranger sous des rubriques (titres). En conséquent, l'indexation des textes imprimés souvent reflète la disposition des collectanées dans les cahiers d'études appelés *adversaria* ou *ephemerides*.

C.3. Concurremment, il faudra normaliser la graphie de ces textes pour faciliter la compréhension aussi bien que la recherche manuelle-automatique des mots.

La recherche en plein texte exécutée par des moteurs de recherche n'a peut-être pas toujours besoin de l'orthographe. Dans le cas du latin, cependant, elle en profitera beaucoup. Cette position est controversée, je le sais. Veuillez bien noter que ce sont seulement les variantes triviales que nous avons normalisées sans les documenter, tandis que les variantes de valeur phonétique (e.g. une contraction) ou stylistique (e.g. un archaïsme) et les formes diverses des noms propres ont été conservées par l'encodage ‚original' vs. ‚regular'. À ceux qui censurent cette démarche, j'adresse deux arguments:

1. Les humanistes cherchaient à restituer et pratiquer la prononciation et la graphie des anciens. L'orthographie de leurs éditions d'auteurs classiques ne diffère point de celle qu'ils observaient dans ses propres écrits. Ne sommes-nous pas autorisés, au même titre, à revêtir les textes latins des modernes de la même orthographie que nous appliquons aux textes anciens, aujourd'hui? Beaucoup d'éditeurs de textes latins des humanistes ont fait ce choix. Voyez e.g. la bibliothèque en ligne *Poeti d'Italia in lingua latina* qui reproduit les textes plus ou moins normalisés d'éditions récentes. (Je doute d'ailleurs qu'il soit convenable de parler d'une langue ‚néolatine': Les bons écrivains latins de l'ère moderne ont observé la latinité des anciens à tel point qu'ils se sont attirés la moquerie d'être leurs singes.)

2. Les chercheurs assez rares qui s'intéressent à la graphie et la prononciation de l'époque moderne doivent étudier les éditions originales, bien sûr. Les lecteurs de CAMENA ont à leur disposition les images distinctes des pages originales qui valent plus qu'une transcription diplomatique.

CAMENA contient l'ouvrage *Orthographia Latini sermonis vetus et nova* de Claude Dausque de Tournai, imprimé à Paris en 1677:
<http://www.uni-mannheim.de/mateo/camenaref/dausque/dausque1/p1/jpg/as001.html>
L'auteur traite l'histoire de la graphie du latin en spécialiste de paléographie et d'épigraphie.

Tant s'en faut qu'il s'occupe d'établir une norme actuelle d'orthographie latine; plutôt, il admet les variantes.

Quant aux intertitres, têtes et tables des livres anciens, la normalisation que nous proposons y paraît d'autant plus accceptable que ces petits textes sont destinés à offrir au lecteur des points d'accès. Ne sommes-nous pas autorisés, donc, d'en tirer le plus grand profit possible en les adaptant aux besoins du lecteur?

Cela vaut certainement mieux qu'employer un métalangage artificiel, en utilisant un vocabulaire fabriqué ou une classification actuelle. Ces systèmes documentaires évoquent un ordre du savoir qui est incompatible aux sources. Une telle duplicité ne peut que confuser les lecteurs. Ils préféreront les vedettes originales aux étiquettes attachées, sans doute. Comparez e. g. le schéma des données structurelles, "Strukturdatenset", imposé récemment par la Deutsche Forschungsgemeinschaft.
<http://dfg-viewer.de/strukturdatenset/>

D.1. Suivant les principes directeurs pour l'encodage TEI-P5, nous allons esquisser un balisage qui réunit des éléments des sections 3.7 « Listes » et 3.8.2.1 « Pre-existing indexes » et du chapitre 9 « Dictionnaires ».

Ce sont là les promesses de mon résumé anticipé écrit en septembre 2009. Maintenant, tout en concevant plus nettement l'usage efficace des éléments d'indexation, je ne vois pas bien comment y parvenir par un balisage qui est déterminé par la lettre et la structure du texte original.

Afin de planer un chemin d'accès commode à l'ouvrage d'érudition latine, je propose d'offrir aux lecteurs deux ou trois tables d'échelle différente où sont exposés la structure et le contenu du livre numérisé. Le plus possible, on utilisera des textes indicateurs tirés de l'édition originale. En défaut de ceux-ci, on admettra des sommaires tirés d'ouvrages de recherche plus récents ou bien des dénominations signalétiques en langue commune ou spéciale, qui soient conformes à l'usage reçu (en distinguant e. g. le privilège de la permission, l'illustration de l'ornement).

La première table, c'est la liste des intertitres signalant les grosses parties du livre – une division ou disposition plutôt qu'une table des matières.

La deuxième table reproduit en plus les têtes des divisions inférieures dont est composée l'edition, c'est-à-dire des chapitres, des lettres, des poèmes.

La collection *ITALI* de CAMENA présente de telles tables. Voyez e. g.
Boccaccio, Giovanni: *Genealogia deorum gentilium*. Basel, 1532:
<http://www.uni-mannheim.de/mateo/itali/autoren/boccaccio_itali.html>
Tous les intertitres de la première table y sont pourvus d'un lien qui amène la partie respective de la deuxième table plus détaillée. D'ici, un hyperlien conduit à l'image qui reproduit la page initiale de la division signalée.

La troisième table contiendra en outre des indications plus détaillées, e.g. les mots clef ou brèves phrases qui se trouvent en marge de page, ou bien les entrées d'index extraites d'une table alphabétique en fin de livre, ou des mots clef glanés du texte intégral, souvent mis en relief dans l'edition originale.

Remarquons en passant que la structure du livre ancien, en moyenne, est plus complexe que celle de nos livres modernes, et qu'un collaborateur de qualification médiocre aura grande peine à démêler les structures irrégulières ou simplement imprévues. Voici un exemple:

Sabinus, Georg (1508-1560): *Poemata Georgii Sabini ... aucta, et emendatius denuo edita* [per Ioachimum Camerarium]. - [Leipzig]: Voegelin, [1568?]. - [21], 532 pp. ; 8o. – Vers la fin du livre, l'imprimeur nous avertit:
<http://www.uni-mannheim.de/mateo/camena/sabin1/jpg/s542.html>
"Maintenant, nous ajoutons quelques lettres écrites par Sabinus, que nous avons obtenues plus tard. Elles sont aussi propres à être publiées que dignes d'être lues."

Regardons d'autre part un exemple d'une simplicité classique:
Thou, Jacques Auguste de: *Historiae Sui Temporis*. - Paris, 1606-1609. 2o.
<http://www.uni-mannheim.de/mateo/camenahist/autoren/thou_hist.html>
et plus spécialement Tom. 2: Lib. XXVII-XXXXIX, 1560-1572. - 1606. CAMENA offre un *Inhaltsverzeichnis*, une table des matiéres: Mais ce n'est là qu'une énumération des vingt-trois livres qui constituent le deuxième tome; il n'y a aucune information de contenu:
<http://www.uni-mannheim.de/mateo/camenahist/thou1/t2/Thuanus_historiae_2-toc.html>
Les pages préliminaires du volume ne contiennent pas de table ou de disposition thématique quelconque. Les pages du texte principal n'exposent qu'une note marginale indiquant l'an où sont arrivés les évènements racontés. Voyez e.g.
<http://www.uni-mannheim.de/mateo/camenahist/thou1/t2/jpg/s150.html>
et <http://www.uni-mannheim.de/mateo/camenahist/thou1/t2/jpg/s151.html>
De plus, il n'y a pas de relief typographique. Donc, cette grande chronique écrit par le témoin savant et acteur politique Jacques Auguste de Thou est d'aspect bien austère.

Heureusement, à la fin de chaque tome, on trouve un index alphabétique, qui donne, en moyenne, une dizaine de références par page du texte narratif. Puisque une page comporte en général plus de cinquante noms, la sélection de dix mots clef par page est bien utile. L'index nous donne le nom au nominatif, tandis que le texte courant souvent contient une forme infléchie. Dans l'index, les noms sont parfois accompagnés d'indications biographiques (lieu d'origine, fonction) ou d'événement, e.g. (dans tome 2): "Cadomensis arx fere capta 91.a – Cadomum captum 73.b.c. 91.a. *& seqq.* [...] Rotomagus capitur 72.a. 160.d – obsidetur 158.c. d - Monachi expelluntur 90.b – tumultus 629d". (Les lettres a, b, c et d ajoutées à la pagination signifient les portions quaternaires de la page.)

De Thou, qui raconte l'histoire de son temps dans l'idiome antique des Romains, tient à latiniser les noms propres. On y trouve des appellations très rares, e.g. le nom de personne *Quadrigarius* dénotant un M. Chartier. En 1634, le savant Jacques Dupuy publiait une longue liste, où sont décodés ces noms latins:
Nominum propriorum ... quae in ... Thuani Historiis leguntur Index:
<http://www.uni-mannheim.de/mateo/camenaref/dupuy.html>
Ici, nous lisons „Cadomum, *Caën*." et „Rotomagum, *Rouën*."
CAMENA offre le texte intégral des quatre tomes de l'édition de Thuanus sans donner une orientation préliminaire. Il faudrait saisir au clavier les tables alphabétiques des quatre tomes et regrouper les entrées suivant l'ordre du texte continu.

D. 2. Notre balisage servira à l'enrichissement, la cumulation et le regroupement des entrées des tables et index divers du livre traité.

Comment définir un balisage adéquat des entrées? Les règles TEI P5 : 3.8.2.1 «Pre-existing indexes» se rapportent à la forme de l'index imprimé : L'index est une liste <*list*>, l'entrée un <*item*> de cette liste. Aussitôt que nous regroupons les entrées suivant l'ordre du texte courant, le balisage effectué ne correspond plus à notre présentation des entrées. - Le sommaire (la table des matières ou divisions) de l'édition imprimée d'ordinaire reprend les

intertitres ou têtes du texte continu. Notre deuxième table d'accès correspond à peu près à une table des matières; cependant elle n'est pas la transcription du sommaire imprimé, mais la série des intertitres et têtes glanés du texte courant imprimé. - Les mots clef ou brèves phrases qui se trouvent en marge de page, dans les règles TEI P5 (3.8 Notes, Annotation, and Indexing) sont traités en note (e.g. <note type="gloss" place="margin">Cadomum captum</note>), bien que leur fonction souvent soit celle de tête subordonnée signalant le thème d'un paragraphe. - Notre balisage sémantique des noms et des termes d'une entrée quelconque vise à produire des index spécifiques de la structure <*list*> et <*item*> ... - N'est-ce pas un maquis de procédure qui nous attend?

Heureusement, les règles TEI P5 offrent une solution de nos problèmes; au moins, elles contiennent un paragraphe propre à justifier notre démarche, laquelle, au lieu de conserver la lettre et la forme de l'édition originale, construit des tables et index à plus-value. C'est la subdivision "3.8.2.2 Auto-generated indexes" qui parle de la construction d'un index à partir du texte numérique intégral. "It can also be useful, however, to generate a new index from a machine-readable text, [...] to construct a fully adequate index, which might then be post-edited into the digital text, marked-up along the lines already suggested for preserving pre-existing index material." Il nous faut supprimer, bien entendu, les phrases "from a machine-readable text" et "into the digital text", puisque nous n'offrons pas le texte intégral numérisé de l'édition originale. Mais laissons là le balisage des entrées totales pour examiner le traitement des mots clef qu'elles contiennent.

Quant au balisage des noms, la parallèle de nom savant et nom vulgaire peut être notée par la balise *original vs. regular form* (<orig reg='[...]'>[...]</orig>); il y manquera cependant l'indication de langue. Les variantes orthographiques de l'un et de l'autre nom peuvent être encadrées de la balise équivalente *regular vs. original form* (<reg orig='[...]'>[...]</reg>). Mais une telle rencontre de deux balises presque synonymes ne satisfait guère.

Le classement des noms par attribut (*type*) et attribut subordonné (*subtype*) est essentiel. Il sera utile de classifier les noms de personne (<name type="person">) en ajoutant des valeurs *subtype*="*myth*" ou "*bible*" ou "*ecclesiastical*" ou "*military*" ou "*academic*" ou "*governmental*" ou "*noble*" ou "other". Ainsi, on pourra produire huit index spécifiques des noms de personne.

Les noms de lieu peuvent être traités de façon semblable: L'attribut *type*="*place*" est complémenté par le *subtype*="*myth*" ou "*bible*" ou "*region*" ou "*town*" ou "*mountain*" ou "*water*" ou "*island*". Il y manque encore l'indication de l'unité géographique ou territoriale qui comprend l'entité nommée, e.g. 'Normandie' pour représenter l'entrée: 'Caen, ville de Normandie'. Peut-on accommoder ces informations dans une seule balise?

Le vocabulaire spécial des disciplines du savoir moderne peut être classé moyennant l'attribut '*term'* accompagné d'un '*subtype'* signalant la discipline. On y suivra la division usuelle de l'ère moderne: *type*="*term*" *subtype* ="*theologia*" ou "*iurisprudentia* " ou "*medicina*" ou "*zoologia*" ou "*botanica*" ou "*mineralogia*" ou "*trivium*" ou "*quadrivium*" ou "*philosophia*" ou "*artes mechanicae*" ou, enfin, "*practica*" (comprenant les objets de la vie quotidienne qui ne se prêtent pas à une des classes antérieures).

La constitution d'index spéciaux et leur juxtaposition représentent, pour ainsi dire, une cartographie du savoir et des faits qui sont à la base d'une œuvre. Chaque index spécial – soit des termes de telle et telle discipline, soit des noms de personne ou des noms de lieu ou des mots grecs – nous permet d'embrasser d'un regard une classe distincte de données traitées dans le livre. Donc, l'index spécial augmente la valeur de l'indexation. En plus, le

balisage sémantique servira à diriger la recherche – qu'elle soit intellectuelle ou automatique – aux ouvrages de référence soit des disciplines soit des classes d'objets respectives. On ne se bornera pas à baliser, dans l'entrée d'index, le mot initial [en grec: *lê:mma*]; on traitera également tout autre mot de valeur indicatrice. Les index constitués ainsi offriront plus d'entrées que l'index imprimé.

Que faire des entrées d'index cumulatives contenant plusieurs entrées qui se rapportent à un seul mot clef (e. g. *Achilles - a matre muliebri habitu absconditus - Ulyssis astu revelatur* – etc.)? Si elles sont transposées ensemble - structurées comme un <*item*> comprenant une liste <*list*> de plusieurs éléments <*item*> - à tous les endroits référenciés, elles encombront la troisième table, qui est le sommaire augmenté que le lecteur veut parcourir vite. Il vaut mieux les dégager pour les y présenter seules, chaque entrée à son endroit.

Il faut aussi indiquer les illustrations ou figures imprimées et ajouter leur légende. Cependant, je ne voudrais signaler, dans notre troisième table d'orientation, ni les ornements (lettrines, vignettes) ni les notes manuscrites, celles-ci n'appartenant pas à l'édition imprimée, ceux-là ne se rapportant guère au contenu. Les notes manuscrites, on pourra les signaler de façon sommaire dans la notice introductive placée avant les textes d'accès extraits de l'édition originale. Quant aux ornements, on les apercevra facilement en parcourant les pages miniaturisées (*thumbnail images*) du livre complet.

Nous résumons: L'ordre de la troisième table qui donne accès au détail du contenu est conforme à celui du texte courant. Tous les éléments exposés seront balisés – de préférence conformément aux règles TEI P5, mais toujours eu égard à leur fonction primaire d'orienter le lecteur. On n'offre pas de transcription du texte intégral. L'usager de CAMENA s'en rapportera aux pages originales qui sont et lisibles et le plus souvent d'une mise en pages instructive et agréable. Le balisage structurel correspond à la mise en pages de nos textes d'accès.

D.3. Digression: Les notes signalétiques en marge de page et l'index alphabétique du livre ancien.

Afin de justifier notre plan, il convient d'analyser la forme et la fonction de ces petits textes qui sont ajoutés au texte courant pour guider le lecteur et faciliter la recherche sélective dans un texte majeur. Gérard Genette, dans son livre sur les paratextes intitulé *Seuils*, ne parle guère de ces textes d'accès. Je n'y ai trouvé qu'une note en bas de page (éd. 1987, p. 292), dans le chapitre sur *Les intertitres*: „A vrai dire, l'usage classique était plutôt de placer en tête une table des chapitres, et à la fin une table des matières proprement dite, sorte d'index plus détaillé." Genette, bien sûr, n'a pas analysé le livre savant mais l'œuvre de fiction.

Je me borne ici à quelques observations qui regardent le livre ancien. Dans les éditions publiées du vivant de l'auteur, les intertitres et têtes (en latin *capita*) précédant chaque partie ou chapitre du livre sont, bien entendu, le plus souvent formulés par l'auteur, tandis que les index en fin du livre sont d'ordinaire de fabrication postérieure, parfois hâtive, exécutée par les correcteurs ou *amanuenses* de l'imprimerie. Il faut donc examiner les index et les comparer aux intertitres, aux résumés (en latin: *argumenta*) que l'on trouve parfois en tête de partie ou de chapitre, aux indications marginales (*tituli margini appositi*) qui souvent fonctionnent comme tête subordonnée.

Voyons un index qui ne reproduit que la moitié de ces notes marginales - sans raison de sélection apparente:
Adam, Melchior: *Vitae*. Frankfurt <Main> et Heidelberg, 1615-1620. 5 parties: *Vitae*

Germanorum Philosophorum; [etc.] 2.815 pages:
<http://www.uni-mannheim.de/mateo/camenaref/adam.html>
C'est une précieuse collection biographique de l'Allemagne savante, dont l'index, malgré l'avertissement prometteur de l'éditeur (*Cum Indice triplici : personarum gemino, tertio rerum*) est d'une qualité bien inférieure.

Dans l'édition bien faite de
Pius <Papa, II.>: *Commentarii rerum memorabilium*. - Roma, 1584:
<http://www.uni-mannheim.de/mateo/itali/autoren/pius_itali.html>
nous trouvons un type assez commun d'index:
<http://www.uni-mannheim.de/mateo/itali/pius1/jpg/s750.html>
Ce n'est, somme toute, qu'une transformation des intertitres en marge de page, enrichie par de rares renseignements glanés du texte principal. Ici, un *titulus margini appositus*, en moyenne, a produit deux entrées de la table alphabétique par simple permutation des mots clef. Dans la table en fin de livre, il y a maintes fautes qui ne se trouvent pas dans les intertitres correspondants. L'on y observe la méthode dérivative et l'empressement du travail d'indexation. Au lieu de transcrire cet index, il vaut donc mieux saisir les indications en marge de page et permuter ces entrées de façon systématique. – En somme, avant de saisir un index alphabétique, il vaut bien la peine d'examiner sa qualité!

Il faut ajouter que la note marginale, peu commune aujourd'hui, mais fréquente dans les livres anciens, ne contient pas toujours une indication de thème. Souvent, c'est une notice bibliographique, parfois une traduction d'une phrase grecque entremêlée au texte latin.

Joseph de Jouvancy, pédagogue et historien jésuite du siècle de Louis XIV, auteur de *Historiae Societatis Jesu Pars Quinta, Tomus Posterior* (1710),
<http://www.uni-mannheim.de/mateo/camenahist/hsj/t52/jpg/as001.html>
a évidemment lui-même conçu les *tituli margini appositi*, les intertitres en marge de page, et réfléchi sur la constitution de l'index. Dans la notice qui précéde l'index
<http://www.uni-mannheim.de/mateo/camenahist/hsj/t52/jpg/s921.html>
nous lisons, que l'on a voulu réduire et raccourcir les notes en marge de page, *ne titulorum crebritas et amplitudo marginem oneraret*, „pour éviter de surcharger [...] la marge de page".
Voyez e.g. cette page-ci:
<http://www.uni-mannheim.de/mateo/camenahist/hsj/t51/jpg/s067.html>
L'index en fin de livre rend les *tituli*, parfois en forme amplifiée. De plus, il réunit les entrées relatives à une personne, ce qui est d'une valeur considérable dans un vaste volume de narration annalistique où les faits d'une personne sont racontés en divers endroits.

D.4. Enfin, on cumulera toutes les entrées des tables ou index de la collection numérisée pour les comparer et pour aligner ensuite les entrées semblables. Il en résultera des chaînes d'entrées indiquant les passages où figurent les mêmes mots clef.

En utilisant des logiciels à calculer le degré d'affinité de textes (*text similarity measure*) on pourra comparer de façon automatique toutes les entrées. Afin d'obtenir des résultats plus pertinents, il convient de sauter tous les mots non balisés. L'exploration en détail d'une bibliothèque étendue et homogène apportera aux chercheurs des informations essentielles qui jusqu'ici ne sont obtenues que par de longues recherches.

E. Épilogue: Suppositions et conditions

Pour finir, il me faut dessiner le cadre qui, selon mon opinion, sera le plus favorable aux travaux proposés.

On établira une bibliothèque en ligne assez homogène et d'ampleur moyenne (de quelques centaines à quelques milliers de volumes). Le cas échéant, une sélection judicieuse d'ouvrages tirés de diverses bibliothèques en ligne pourra constituer le corps d'une nouvelle bibliothèque à plus-value (*Creative Commons*). Avant de choisir une édition, il faudra évaluer non seulement l'ouvrage, mais aussi les textes d'accès de l'édition considérée.

Les collaborateurs seront des latinistes accomplis et curieux du savoir de l'ère moderne. On devra garantir la continuité de leur travail. Puisqu'il s'agit d'opérations complexes basées sur des connaissances étendues, des projets de brève durée et des emplois passagers ne sont pas adéquats. La création d'une bibliothèque en ligne, étant sortie de sa phase initiale de tentative, doit être une occupation professionnelle permanente, plutôt qu'un stage transitoire marqué par les apprentissages, l'embarras et l'insécurité.

De plus, les collaborateurs auront besoin d'une infrastructure propre à faciliter leur travail et à empêcher la duplication d'un effort déjà fait ailleurs. Le patrimoine latin de l'Europe exige la coopération internationale. Quand est-ce que nous jouirons d'un environnement électronique intégrant toutes les données lexicologiques et terminologiques, toutes les notices bibliographiques d'autorité et, en plus, le sténogramme des faits historiques – en somme, les diverses données qui, réunies et accessibles à tous, serviraient à éclaircir les questions de langue et de fait posées par les textes latins modernes. On tenterait ensuite de connecter un mot clef ou une phrase du texte avec les entrées d'information correspondante contenues dans ces bases de données - de façon automatique, bien sûr, n'offrant au lecteur que des propositions parmi lesquelles il devra choisir. Pour y parvenir, il faut pousser les efforts complémentaires de cumulation des fichiers d'autorité, de balisage des textes et de valorisation des tables des matières apportant de nouvelles données.

On verra si le grand projet IMPACT (IMproving ACcess to historical Text, 2008-2011), <http://www.impact-project.eu/>
sorti de l'initiative i2010 de l'Union Européenne, tâchera de créer une infrastructure pareille. Je crains que l'empressement des grandes institutions partenaires ne soit pas favorable à un travail de longue haleine. L'ambition de numériser en mode image et en mode texte toutes les éditions anciennes en peu de temps ne me semble pas bien fondée. La tradition culturelle demande la sélection des sources estimées par les contemporains et par la postérité, aussi bien que la médiation qui s'adresse au lecteur. Afin d'engager un public diversifié de chercheurs spécialistes, professeurs, étudiants et érudits curieux, il faudra le cibler et l'orienter en aidant le lecteur à vaincre les difficultés du texte présenté. C'est une tâche nouvelle, essentielle sans doute. Autrement, l'énorme effort de numérisation du patrimoine latin de l'ère moderne ne produira peut-être que des copies d'archives dormantes.

P. S. Il nous faut prévenir nos lecteurs, que la bibliothèque en ligne CAMENA ne peut plus être augmentée. On a bien conçu et élaboré en détail un propos visant à enrichir le *Thesaurus eruditionis* de CAMENA en numérisant des livres volumineux imprimés en Italie et en Europe occidentale – des livres qui se trouvent dans la bibliothèque universitaire de Mannheim. Mais l'université de Mannheim, au lieu de prolonger notre coopération sans reproche, le 12 mai 2009 s'est subitement retirée de la coopération. Ainsi, le projet CAMENA fut dérobé de la chance de soumettre son propos au jury de la DFG (fondation allemande de la recherche). Nos efforts d'enlever ce blocus ont échoués.