Numérisation et codage des caractères de livres anciens

Jacques André

Irisa/Inria-Rennes
Campus universitaire de Beaulieu
F-35042 Rennes cedex
Jacques.Andre@irisa.fr

RÉSUMÉ. *La numérisation des livres anciens n'a pas été abordée aussi bien que celle des ouvrages manuscrits anciens, du moins en ce qui concerne les caractères. Or, avant de faire de la reconnaissance de caractères, encore faut-il avoir un certaine connaissance de ceux-ci. Par ailleurs, restituer un ouvrage avec ses caractéristiques typographiques peut induire des recherches qui sinon seraient impossibles. Nous présentons alors le projet Cassetin dont le but est de faire un inventaire des caractères d'imprimerie utilisés depuis le XV*^{e} *siècle et de proposer une normalisation de leur codage (sous forme d'entités ou de caractères au sens d'Unicode) de façon à rendre portables les sorties d'OCR.*

ABSTRACT. *Digitalization of ancient books is far less concerned with characters than digitalization of ancient manuscripts. However, before recognition you need cognition. Furthermore, new researches may be launched from texts marked with typographic tags and would be quite impossible without the actually used types. The Cassetin project is proposed to make the inventory of types used from the 15*th *century and to establish an encoding (either as entities or as Unicode codes when available). This is a way to make OCRs' output more portable.*

MOTS-CLÉS : *caractère, codage, glyphe, ligatures, livres anciens, normes, OCR, projet Cassetin, typographie, Unicode.*

KEYWORDS: *ancient books, Cassetin project, characters, encoding, glyph, ligature, OCR, standard, Unicode.*

1. Introduction

L'édition savante, l'édition critique d'œuvres anciennes, a commencé bien avant l'imprimerie et a acquis ses lettres de noblesse et ses canons avec celle-ci. Les ordinateurs puis l'internet ont apporté des outils que nos collègues des sciences humaines utilisent en général très vite et très bien. Ainsi, le concept de codage leur permet d'utiliser le même texte à des fins différentes : un texte codé en TEI[1] peut être imprimé ou affiché avec tous les attributs typographiques de l'édition professionnelle tandis que ses balises de nature sémantique permettent d'y faire des recherches de noms propres, de dates, etc.

Nous nous intéressons ici à une classe particulière de documents : les livres (souvent anciens) qui ont été numérisés et qui sont accessibles aux chercheurs sur le web ou sur cédérom. De plus, nous nous intéressons à une classe spéciale de chercheurs, ceux en bibliologie, plus précisément en orthotypographie[2], disons les paléographes de la chose imprimée ! Or s'il existe des outils et des normes pour l'étude des manuscrits anciens, il n'en est pas de même pour les livres. En particulier, les sorties des OCR *(Optical Character Recognizers)* ne sont pas normalisées et ne permettent pas en général de conserver trace de ce qui a été effectivement imprimé, le texte ainsi saisi étant en quelque sorte appauvri. L'objet de cet article est donc d'en montrer l'utilité et de proposer un codage, donc un inventaire de ces caractères « perdus ».

2. Divers modèles de numérisation des livres anciens

De nombreux livres ont été numérisés et mis à disposition du public. Mais il en existe plusieurs classes.

2.1. *Mode image seulement*

Les pages du livre ont été scannées et les images de ces pages sont accessibles, sur le web, par tout le monde ou de façon restreinte (abonnement ou achat dans le cas de cédéroms notamment). On retrouve ici les principes des microfilms, mais la qualité varie d'un site à l'autre : il y a des images qui sentent le travail de masse[3],

---

1. *Text Encoding Initiative*, voir (Role, 1996).

2. L'orthotypographie est une discipline très ancienne (on attribue à Hornschuch en 1608 la création de ce terme) mais son sens a évolué et couvre aujourd'hui tout ce qui relève de l'art d'écrire correctement une langue, tant au plan de la grammaire que de sa forme graphique (imprimerie, écrans, manuscrits...) et notamment ce qui relève du « code typographique ». On trouvera dans (Méron, 2002) une longue bibliographie d'ouvrages relevant de ce domaine.

3. C'est le cas de la collection Gallica de la BNF (http://gallica.bnf.fr/) où les livres ont été parfois filmés ouverts, sans tenir compte de la courbure de la reliure déformant le texte, voire

d'autres qui sont très mal présentées[4] et certaines frisant la perfection[5]. Ces bases de données sont des outils très précieux pour consulter de nombreuses œuvres : Gallica permet ainsi à un étudiant, faisant par exemple son diplôme sur les influences de La Pléiade sur la langue française, de lire dans le texte les versions originales de Meigret ou de du Bellay et d'en disposer d'une copie de travail pour l'annoter ou pour surligner des passages (et ce de façon moins chère – et plus légale – qu'en « photocopillant » les rééditions fac-similé des éditions Honoré Champion par exemple). Mais il ne pourra faire aucune recherche « électronique », ne serait-ce qu'une recherche en « plein texte » sur ces images, sauf bien sûr en disposant d'un système de reconnaissance de caractères adapté, ce que l'on ne trouve pas dans le commerce actuellement[6].

2.2. *Mode texte seulement*

C'est sans doute encore le cas le plus fréquent des bibliothèques numériques. Citons par exemple, pour le français, certains textes proposés par l'InaLF[7] ou par l'ABU[8]. Mais ce mode texte peut prendre plusieurs allures selon sa finalité.

2.2.1. *Texte non formaté*

Le texte est saisi et affichable « au kilomètre », sans le moindre enrichissement typographique ou autre (sauf en général les fins de paragraphe ou de sectionnements tels que les titres, etc.), sans la moindre image. On parle parfois de texte source, de texte brut, de plein-texte, etc. Ce texte source ne contient que du texte ayant un sens langagier (et donc aucune « balise » ni aucun caractère de mise en page). Ce texte

---

le cachant ; par ailleurs, très souvent, la définition est très basse, rendant impossible tout agrandissement et donc tout essai de lecture de détails.

4. Typiquement cette *Histoire naturelle* de Buffon dont les pages pourtant bien scannées (sauf quelques-unes mal massicotées, coupant le texte) et aux illustrations très belles sont mises en page de façon naïve, digne d'un potache de cinquième (classeur à spirales) : http://www.oiseaux.net/cgi-bin/redir?http://www.oiseaux.net/buffon/buffon.tome6.html

5. On pense en particulier aux éditions d'Octavo dont par exemple, pour se placer dans le monde de la typographie qui sera le nôtre ici désormais, le *Manuale Tipografico* de Bodoni, le *Champfleury* de Tory ou une *Bible* de Gutenberg dont les images sont accessibles (dans les versions *Research Facsimile*) à très haute définition : http://www.octavo.com/ ! Par ailleurs, ces ouvrages sont présentés « livre ouvert », c'est-à-dire que les deux pages (paire et impaire) sont côte à côte (mais la courbure est corrigée) retrouvant ainsi les tracés régulateurs des canons de mise en page.

6. Par ailleurs, même s'il en existait, il serait peu probable d'aboutir à quelque chose d'utilisable vu la faible définition des images offertes sur ces sites pour grand public.

7. http://www.inalf.cnrs.fr. Certains de ces textes sont repris par le site Gallica de la BNF.

8. Association des bibliophiles universels : http://abu.cnam.fr/

est écrit en respectant un codage qui peut être soit les codes d'une norme[9] (comme ASCII[10], Latin-1 ou Unicode ; voir figure 1, ligne 2, soit les codes d'un standard propriétaire (comme le « code page » de Windows, figure 1, ligne 3.

2.2.2. *Mode de restitution*

Nous appelons ainsi les documents qui comportent, en plus du texte, des informations complémentaires permettant la restitution (papier ou écran) du texte (italique[11], gras, etc.), voire de la mise en page (paragraphes, listes, en-têtes, etc.).

2.2.3. *Mode structuré*

Le texte comprend une structuration logique souvent à l'aide de balises[12] (la TEI par exemple permet de spécifier que ce que Word considère comme une « liste » est en fait un « dialogue », voir figure 1 ligne 4), voire un marquage plus sémantique (attribut « boutique », par exemple). Il est en général facile d'en déduire un mode de restitution (c'est-à-dire de produire un texte balisé typographiquement).

Ces trois modes permettent de faire de la recherche de mots. Pour le mode non formaté, ceci se fait de façon très simple et efficace, avec évidemment les limites liées au codage lui-même : rechercher dans un texte codé en ASCII le mot « côté » ne donnera aucun résultat ou alors (selon la façon dont le programme de recherche traite la clé) aussi les occurrences de « cotés » et de « cote »[13].

Les autres formats permettent également de faire de la recherche de texte (à condition d'avoir un outil adapté au format considéré pour sauter les balises, marques spéciales, etc.), mais aussi des éléments de structure (dialogue par exemple), voire « toutes les boutiques » (et dans ce cas il faut des outils spécifiques

9. Voir (André Hudrisier, 2002 ; p. 13-88) pour la distinction entre norme et standard et pour les principales normes comme ASCII, Latin-1 ou Unicode.

10. L'ASCII n'ayant pas de voyelles accentuées, il est assez curieux que certains sites osent encore proposer des textes français avec ce codage, comme *Du côté de chez Swan* du Projet Gutenberg de livres électroniques : http://www.ibiblio.org/gutenberg/etext01/7swan11.txt.

11. Il y a toujours un certain malentendu au sujet de certains attributs typographiques. Il est de coutume de conserver dans un texte brut les guillemets. Or ceux-ci peuvent indiquer une citation qui ailleurs, notamment chez l'auteur qui souligne, sera mise en italique. Or on ne conserve pas ce dernier marquage dans un texte brut…

12. Sur le concept de balises, voir (Role, 1996 ; p. 11-22).

13. Signalons au passage la difficulté de faire des recherches efficaces quand « on » remplace non seulement les voyelles accentuées par celles non accentuées et les majuscules par des minuscules. Nous avons eu le plus grand mal à trouver dans une *Encyclopédie* électronique de Diderot si on y parlait de Dürer à cause du bruit causé par le verbe « durer » !

ou bien paramétrés du type interrogation de base de données). Ces deux derniers modes permettent bien sûr de faire figurer des images dans le texte[14].

« Original » (Ici *N.D. de Paris,* de V. Hugo, Gallica)	— Jacques Coppenole. — Vos qualités ? — Chaussetier, à l'enseigne des *Trois Chaînettes,* à Gand.
Source (ici Latin-1)	— Jacques Copperole. — Vos qualités ? — Chaussetier, à l'enseigne des Trois Chaînettes, à Gand.
Restitution (ici RTF Word)	\par {\listtext...}Jacques Coppenole. \par {\listtext...}Vos qualit\u233\'8es\~? \par {\listtext...}Chaussetier \u224\'88 l\rquote enseigne {\i Trois Cha\u238\'94nettes, \u224\'88 Gand.
Structuré (TEI)	<sp who=Coppenole> Jacques <nom>Copperole</nom> </sp> <sp who=huissier> Vos qualités_? <sp who= Coppenole> <metier> Chaussetier</metier>, à l'enseigne des <boutique>Trois Chaînettes</boutique>, à <lieu>Gand</lieu>. </sp>

Figure 1. *Divers codages d'un même document*

2.3. *Mode mixte, texte et image*

Nous nous intéressons donc ici à la classe particulière de documents que sont les livres anciens numérisés pour les chercheurs – ce qui veut dire en général accessibles au moins en mode texte – et en particulier pour ceux qui ont aussi besoin du texte en mode image (pour étudier la mise en page, la typographie, etc.). Il existe de nombreux sites où cette dualité est bien présente, notamment pour les textes manuscrits[15], ce qui s'explique par la difficulté de lecture des originaux par des non

14. On peut assimiler le format PDF (qui est un format de visualisation et non d'édition) à ce mode puisqu'il permet d'afficher du texte avec les possibilités de recherche de mots ou des images (de texte) sans ces possibilités. C'est pourquoi les textes de Gallica scannés en mode image et affichés en PDF ne permettent pas cette recherche de mots.

15. Qu'il s'agisse de textes médiévaux comme *Le chevalier de la Charrette* ou d'œuvres du XIX^e^ siècle comme le *Journal* de Marie Daniel Bourrée de Corberon, dont les *url* sont http://www.mshs.univ-poitiers.fr/cescm/lancelot/ et http://egodoc.revues.org/corberon/, voire des manuscrits « modernes » comme ceux de Flaubert.

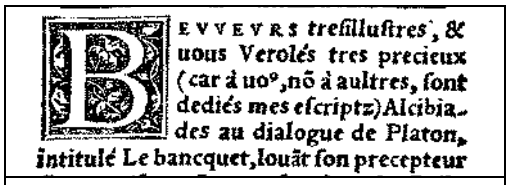
spécialistes. Elle est bien sûr à la base des vrais hypertextes[16] où l'on trouve des liens à la fois sur des manuscrits d'auteurs, des images d'œuvres éditées et diverses analyses. Cette dualité apparaît sans doute bien moins utile pour les livres imprimés (surtout ceux non illustrés !) du fait de la lisibilité des images de texte imprimé. Il existe toutefois des sites où l'on trouve(ra) ainsi des livres disponibles tant pour la forme que pour le fond, avec des liens hypertextuels de l'un à l'autre. Ces sites sont consacrés par exemple :

– à une période, comme les livres de la Renaissance du projet Debora[17],

– à un auteur, comme Cervantes[18],

– à un type d'ouvrages comme les dictionnaires[19] ou des textes scientifiques[20].

En ce qui concerne notre propos, étudier les textes d'un point de vue « typographique » autant que textuel, ce mode permet donc de mettre en relation un texte et sa représentation avec diverses granularités : page, paragraphe (comme ci-dessous en figure 2), signe (voir par exemple (Le Bourgeois, 2003 ; figure 2)), etc.


BEVVEVRS treſillustres, & uous Verolés tres precieux (car à uo⁹, nõ à aultres, ſont dediés mes eſcriptz) Alcibiades au dialogue de Platon, intitulé Le bancquet, louãt ſon precepteur

BEUVEURS tres illustres, et vous Verolés tres precieux (car à vous, non à aultres, sont dediés mes escriptz) Alcibiades au dialogue de Platon, intitulé Le bancquet, louant son precepteur…

Figure 2. *Prologue du Gargantua de Rabelais et translittération moderne*

Toutefois, à de rares exceptions, ceci n'est en général pas satisfaisant. Regardons cet extrait du *Gargantua* (figure 2) non comme une œuvre de Rabelais, mais de Dolet[21] : rien ne permet l'étude des caractères (abréviations latines, usage des

16. L'archétype de ces hypertextes étant HyperNietsche. Voir Paolo D'Iorio, *HyperNietsche*, Presses Universitaires de France, 2000 ; http://209.41.39.163/hypernietzsche/

17. http://rfv6.insa-lyon.fr/debora/

18. http://www.csdl.tamu.edu/cervantes/

19. http://www.lib.uchicago.edu/efts/ARTFL/projects/dicos/

20. http://www.colisciences.net

21. Imprimeur lyonnais, Etienne Dolet (1509-1546) a fait partie, avec les Estienne, Tory, etc., de ces humanistes qui ont forgé la langue française et l'ont fixée par leurs imprimés. La figure 2 est construite ici à l'aide d'une image extraite du site Gallica (notice : *Num. BNF de*

diacritiques, s long, ligatures, etc.), ni l'emploi des capitales, de l'italique (ici absent, mais dans l'édition de F. Juste on trouve « *Le Banquet* »), des espaces (avant ou après les ponctuations et parenthèses), la division des mots en fin de ligne (plus bas dans ce texte on verrait des divisions comme « perfe-ction »), sans parler de la lettrine à 5 points. Or il se trouve qu'aujourd'hui le codage de toutes ces informations peut aussi être fait, souvent automatiquement, mais qu'il en manque une certaine normalisation.

2.4. *Quelques apports technologiques liés aux caractères*

Depuis quelques années, les caractères ont fait l'objet de recherches ou de développements industriels les rendant plus manipulables même si souvent il demeure une certaine ambiguïté à leur sujet…

2.4.1. *Reconnaissance des caractères*

La reconnaissance des caractères *(OCR Optical Character Recognition)* a fait des progrès spectaculaires depuis quelques années, mais il reste encore à faire !

Caractères imprimés. A priori, on sait reconnaître les caractères imprimés (voir (Lefèvre, 1999) pour une synthèse) et on trouve donc sur le marché de nombreux produits commerciaux de qualité. Toutefois, ces derniers ne sont pas toujours capables de reconnaître des caractères « exotiques » ou anciens (du fait par exemple de leur encrage ou, ce qui revient au cas précédent, de leur forme non connue des reconnaisseurs ; voir par exemple figure 3). Signalons cependant que les recherches en cours utilisent des techniques nouvelles (voir par exemple (Allier, 2003) et (Le Bourgeois, 2003)) augurant d'une reconnaissance poussée.

Figure 3. *Il faut « apprendre » aux OCR que « ﬁ » est la ligature « si »*

---

l'éd. de A Lyon : chés E. Dolet, 1542. in 16[e]) et d'une translittération inspirée de celle de l'ABU (basée sur la dernière édition du *Gargantua* revue par Rabelais, celle de François Juste, Lyon, 1542).

Caractères manuscrits. La variabilité du tracé des scripteurs et la connectivité des lettres rendent difficile la reconnaissance de l'écriture manuscrite (voir (Crettez et Lorette, 1998) pour une synthèse). Toutefois de grands progrès sont ici en cours du fait du développement des téléphones portables et… de l'engouement des mises de documents d'archives sur le web (voir par exemple (Coüasnon, 2003)).

Caractères dactylographiés. A priori, ils sont assimilables aux caractères imprimés. Mais plus leur qualité baisse (à cause des rubans encreurs, pelures, etc., les caractères sont alors soit empattés et liés aux voisins ou, au contraire, formés de composants non connexes) moins ils sont reconnaissables. Paradoxalement, c'est sans doute là qu'il y a le plus d'études à faire, surtout si l'on sait la masse de documents archivés sous cette forme !

On est donc en droit de considérer que d'ici très peu on pourra reconnaître tout caractère imprimé. Reste à savoir ce qu'est un caractère, problème enfin abordé par Unicode récemment.

2.4.2. *Caractères et glyphes, Unicode et OpenType*

Les codages à 8 bits comme Latin-1 offrent un grand nombre de lettres accentuées mais pas toutes ni toutes en même temps. Par ailleurs, de nombreux organismes ont proposé des codages permettant de traiter divers autres alphabets (arabes, japonais, etc.). Depuis plus de dix ans déjà, deux organismes ont étudié un codage universel : l'ISO, organisme international de normalisation, et le consortium Unicode, un regroupement privé de constructeurs d'ordinateurs. Ces deux groupes ont fini par s'entendre et converger en produisant une norme internationale à 32 bits dite ISO/CEI-10646 et un standard propriétaire, Unicode[22], qui est presqu'un sous-ensemble à 20 bits de la norme. Il permet (ou permettra dans ses mises à jour) le codage de tous les caractères du monde, présents ou passés. Les principes de base d'Unicode comprennent les points suivants.

– Chaque caractère est défini par *un numéro* (code en base 16) et un nom auquel s'ajoute de façon non normative un exemple de représentation (glyphe privilégié).

– La combinaison de caractères et de signes diacritiques est possible et permet d'utiliser des caractères accentués qui ne seraient pas définis[23].

– Un principe d'unification[24] permet de ne pas faire de doublons pour les caractères chinois, japonais, coréens et vietnamiens. Pour les langues européennes,

22. On trouvera dans (Unicode, 2003) la définition de ce codage et dans (André et Hudrisier, 2002) divers articles sur l'histoire et les principes de ce codage mais aussi sur le concept de glyphe. On commence déjà à trouver des logiciels et des fontes qui utilisent ce codage (c'est par exemple le codage de base de produits comme Word depuis Windows-2000 et de Mac OSX).

23. Des médiévistes se définissent ainsi un caractère « æ̆ » par la combinaison de la ligature « æ » et de l'accent bref « ˘ ».

ça signifie que « Y » (code 0057) représente aussi bien le caractères français « i grec », que le caractère anglais « *wye* » ou que l'allemand « *Ypsilon* ». En revanche, « A » faisant partie des langues latines et de l'écriture grecque, Unicode distingue les deux (codes 0041 LETTRE MAJUSCULE LATINE A et 0391 LETTRE MAJUSCULE GRECQUE ALPHA).

– Mais le principe le plus important, du moins en ce qui nous concerne, est la distinction caractère-glyphe. Pour Unicode, un caractère est la plus petite unité distinctive d'une langue écrite, tandis qu'un glyphe en est une représentation possible (sur écran, papier, etc.). Par exemple, au caractère LETTRE MAJUSCULE LATINE R peuvent correspondre des glyphes variant selon le style du caractère « R, R, R, R, R, R », ses attributs typographiques (graisse, italique, etc.) « R, **R**, *R*, ***R*** », sa taille « R, R, R, R », etc. Ce principe conduit Unicode à considérer que les ligatures ne sont que des problèmes de « rendu » et donc à les ignorer. Toutefois, pour des raisons de compatibilité, on en trouve quelques-unes (« fi », « fl », etc.). La distinction glyphe-caractère n'est pas aussi simple qu'il y paraît et Ken Whistler, directeur technique d'Unicode, en rappelle les principaux points : « Unicode code les caractères d'une écriture et non d'un alphabet… On ne peut coder toutes les combinaisons de digrammes… Il faut garder à l'esprit les différents niveaux en présence : les phonèmes d'une langue, les graphèmes d'un système d'écriture, les lettres d'un alphabet et les caractères d'une écriture commune à plusieurs langues[25]. »

Unicode ne serait pas utilisable sans le concept de fonte qui, lui, va traiter des glyphes. Aux divers formats déjà existants (Type1, TrueType, etc.) est donc venu s'ajouter un nouveau format allant dans l'esprit d'Unicode : OpenType. Celui-ci permet par exemple de traiter le æ brève cité en note 23 ou de remplacer automatiquement[26] la séquence des lettres « f » et « i » par la ligature « fi ».

Partant de la translittération de la figure 2 (qui, écrite en latin-1, est donc écrite dans un sous-ensemble d'Unicode), on imagine qu'un « programme » OpenType pourrait produire un texte similaire à celui de l'image[27]. Restent toutefois certains problèmes comme celui de savoir si cette différence caractère-glyphe s'applique bien, au XVI^{e} siècle, aux caractères « u » et « v », « i » et « j », etc.

---

24. En contradiction parfois avec un autre principe (permettant à Unicode d'être compatible avec des codes plus anciens comme Ascii ou Latin-1 mais aussi avec des codes propriétaires comme les codes pages d'IBM ou des fontes du Mac) qui fait qu'il y a parfois des doublons comme « Å » qui est défini à la fois en 00C5 (LETTRE MAJUSCULE LATINE A ROND EN CHEF) et en 212B (SYMBOLE ANGSTRÖM).

25. Patrick Andries, « Entretien avec Ken Whistler », (André et Hudrisier, 2002 ; p. 329-351).

26. Mais de façon programmée, par le biais des tables CMPA, GLYPH, CFF, GSUB, GPOS… voir P. Andries, « Introduction à Unicode », (André et Hudrisier, 2002 ; p. 51-88).

27. Ce programme dirait que les glyphes de U et de V sont des V, que le s à l'intérieur d'un mot est toujours un s long (« ſ » et celui-ci est alors ligaturé avec le i ou le t suivant) et que celui final est notre s court, que « non » s'abrège en « nõ » et « us » final en « ^{9} », etc.

2.4.3. *Le chaînon manquant*

Pour pouvoir faire plus « avec l'électronique » que de façon traditionnelle, il faut donc que l'on puisse numériser (au sens de scanner) un document et conserver le détail des caractères effectivement imprimés, qu'il s'agisse de « caractères » ou de « glyphes » au sens d'Unicode.

Si, dans le cas de la figure 2, il n'y a *a priori* pas de problèmes (« si », « st », « õ », etc., sont bien des glyphes[28]), il n'en est pas de même en figure 4 : à côté des caractères usuels (appartenant au sous-ensemble « latin-1 » d'Unicode), Antoine de Baïf[29] utilise des « caractères » également dans Unicode mais inventoriés ailleurs, comme le « e », dit e-ogonek[30] mais aussi des caractères comme le « A avec trompe » (entre O et P, se prononçant « au ») qui sont des caractères que l'on espère voir ajoutés prochainement à Unicode.

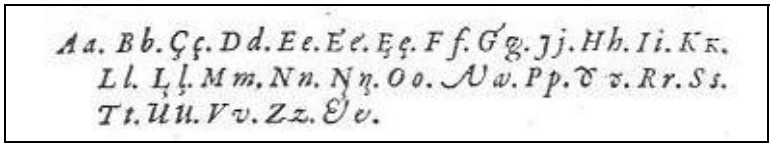

Aa. Bb. Çç. Dd. Ee. Ee. Ęę. Ff. Gg. Jj. Hh. Ii. Kk.
Ll. Ļļ. Mm. Nn. Ņņ. Oo. Ww. Pp. Vv. Rr. Ss.
Tt. Uu. Vv. Zz. Ee.

Figure 4. *Extrait des Etrénes de Poézie fransoeze an vers mezurés d'Antoine de Baïf (1574), numérisé par Gallica*

Enfin, si l'on désire que les textes sortis des OCR soient utilisables par toute la communauté scientifique, il convient d'utiliser le même codage pour les « caractères imprimés reconnus », qu'il s'agisse donc de purs « caractères Unicode » ou de « glyphes ». Il faut alors un consensus sur ce codage et donc sur l'inventaire des caractères à reconnaître. Tel est l'objet du projet Cassetin.

---

28. En revanche, l'abréviation « ^{9} » pour « us » ne peut être considérée comme le chiffre 9 en supérieur mais doit être codée comme un glyphe de « us », même si souvent cette abréviation est ainsi éditée, comme ici...

29. Comme les autres membres de la Pléiade, dont notamment du Bellay, Antoine de Baïf a cherché un mode d'écriture du français qui tienne d'avantage compte de la phonétique. De nombreux ouvrages ont ainsi été composés à cette époque.

30. Des langues d'Europe de l'Est comme le polonais. On l'appelle parfois e-ogonille car le diacritique a souvent la forme d'une cédille vers la gauche. Ce caractère a été utilisé pour « é » en France dans des textes latins imprimés dès le XV^{e} siècle et était encore attesté dans la casse de Fertel, *La science de l'imprimerie*, 1723.

3. Le projet Cassetin

Le projet Cassetin[31] a pour objet de faire un inventaire des principaux « caractères » (d'imprimerie), de leur associer un codage et de proposer au consortium Unicode une liste de « caractères » (au sens d'Unicode) qui n'en font pas encore partie.

Précisons d'abord le statut de ce projet : il n'a pas encore officiellement démarré. L'objet de ce papier est notamment de faire un appel à participation en vue d'un projet européen. La première application étant son utilisation comme codage des sorties d'OCR de livres, il nous paraît intéressant de rappeler d'abord ce qu'ont fait les historiens des textes manuscrits.

3.1. *Codage des manuscrits anciens*

Malgré certaines divisions dans le monde des spécialistes d'histoire des textes (approches diplomatiques, codicologiques, etc.), divers projets internationaux s'intéressent au codage des manuscrits anciens. Voici deux approches différentes.

Codage d'entités. Lancelot ou *Le Chevalier de la Charrette* est un texte célèbre de Chrétien de Troyes (XII^{e} siècle) ; il est en cours de numérisation par le département des Langues romanes de l'université de Princeton et le Centre d'études supérieures de civilisation médiévale de Poitiers[32]. Pour coder la transcription TEI des manuscrits, le projet a défini toute une série de *tags* (entités SGML) sous la forme « `&xxx;` » (voir figure 5).

Entités SGML/TEI	Exemple (glyphe)	Commentaire
&et2;		Caractère « et »
&xx-til;		Signe composite (ici p+tilde)

Figure 5. *Exemples d'entités SGML utilisées par le projet Charrette*

31. Cassetin peut être vu comme acronyme de « CAS(s)e Encoding Type INitiative ». Cassetin est le nom des cases de la casse où se rangeaient les caractères d'une police.

32. Voir : http://www.mshs.univ-poitiers.fr/cescm/lancelot/. Notons qu'il existe depuis longtemps une autre version du *Lancelot*, numérisée et translittérée par Guy Jaquesson et dont le gros intérêt est la vision synoptique (synchronisée et parallèle) des huit versions : http://homepage.mac.com/guyjacqu/sisyphe/bibliotheca/index.html

Complément à Unicode. Le projet *MUFI (Medieval Unicode Font Initiative)* a pour but le codage et l'affichage de caractères spéciaux de manuscrits médiévaux de langue latine[33]. Contrairement au projet Charrette, MUFI propose que ces entités (qu'il définit également) soient aussi de nouveaux caractères Unicode. MUFI étudie donc l'inventaire des caractères à ajouter à Unicode pour pouvoir manipuler les textes médiévaux (nordiques, portugais, etc., mais à vocation européenne).

	Type	Glyphe	Entité	Nom
1	Ecritures mélangées		`&tunc;`	LATIN LETTER UNCIAL T
2	Diacritiques précomposés		`&Avligac;`	LATIN CAPITAL LIGATURE AV WITH ACUTE
3	Ligatures		`&aalig;`	LATIN SMALL LIGATURE AA
4	Signes de ponctuation		`&pause;`	PUNCTUATION MARK PAUSE (DOT BELOW BREVE)
5	Abréviations au niveau de la ligne		`&xpm;`	ABBREVIATION SIGN "CHRISTUM"
6	Signes combinatoires		`&rabar;`	COMBINING ABBREVIATION MARK SUPRALINEAR "RA" (OMEGA SIGN) WITH BAR ABOVE
7	Abréviations précomposées		`&kbarasc;`	LATIN SMALL LETTER K WITH BAR THROUGH ASCENDER (ABBREVIATION SIGN "KALENDAS")
8	Caractères de métrique		`&ancgr;`	METRICAL SYMBOL ANCEPS WITH SECONDARY STRESS

Figure 6. *Quelques codes du projet MUFI*

Cet inventaire comprend un glyphe privilégié, un code Unicode (et, si ce caractère n'y existe pas, une proposition de codage dans la zone privée), un nom mais aussi une entité SGML (figure 6). A cet inventaire, encore en cours, sont liés

33. http://www.hit.uib.no/mufi/

deux autres projets ; l'un, Junicode[34], propose des fontes pour l'édition de ces textes médiévaux ; l'autre indique comment saisir divers caractères[35].

3.2. *Inventaire des types*

Le projet Cassetin a donc pour but de faire pour les livres un peu l'équivalent du projet MUFI et reprendra dans celui-ci tout ce qui peut être compatible avec l'imprimé. On considère donc comme candidats à cet inventaire les caractères devenus aujourd'hui des choix typographiques mais qui ont été la règle autrefois (rentrent bien sûr ici les caractères de La Pléiade, figure 4, les « s longs », les ligatures comme « ct » et « si », mais aussi des ligatures du début du XX^{e} siècle comme le « k barré » breton), les caractères tels que les petites capitales, les lettres supérieures, etc., qu'Unicode considère comme glyphes alors qu'ils ont un rôle orthotypographique certain[36], les guillemets, tirets et… espaces, les très nombreux signes divers (tels que les pieds de mouche « ¶ », les croix mortuaires françaises « † », etc., dont les manuels de typographie[37] et spécimens donnent des listes plus longues). On pourra aussi citer les ornements et vignettes pour lesquels il existe des sites spécialisés[38].

Cet inventaire se fait[39] :

– en inventoriant les caractères déjà existants dans Unicode, ce qui n'est pas toujours facile[40] compte tenu de leur ventilation « un peu partout » ;

– en étudiant les polices et casses disponibles – on étudie notamment les inventaires publiés par (Barber, 1969 ; Baudin, 1998 ; Bigmore, 1978 ; Méron, 2002) – et bien sûr les *spécimens* (catalogue de caractères des fonderies) dont (Dreyfus, 1963) ;

– en étudiant divers ouvrages, notamment ceux déjà numérisés pour voir l'utilité de notre inventaire ;

– enfin, s'il n'a guère été question ici que du français, il est bien évident que toutes les langues latines sont concernées.

---

34. Nom de travail d'un groupe de *Old English at the University of Virginia* définissant une fonte Unicode pour médiévistes : http://www.engl.virginia.edu/OE/junicode/junicode.html et … OE/Fonts.About.html

35. *The Menota Handbook* : http://helmer.hit.uib.no/menota/guidelines/

36. Voir les articles de Y. Haralambous et de O. Randier dans (André Hudrisier, 2002).

37. Voir les titres dans (Méron, 2002) ; citons par exemple le *Guide du compositeur* de Théotiste Lefèvre (1855) et le *Manuel de typographie* de Frey (1857).

38. Ainsi existe-t-il une banque des ornements d'imprimeur :
http://www.unil.ch/BCU/docs/collecti/res_prec/fr/todai_intro.html

39. Une version de travail se trouve à http://www.irisa.fr/faqtypo/BiViTy/cassetin.html. Voir aussi (André, 2003).

40. Il existe heureusement des pages comme http://www.eki.ee/letter/ qui sont très utiles !

Glyphe	Entité	Unicode	Notes	15	16	17	18	19	20
	&aog;	0105	a ogonek		+				
	&an;	0101	Abrev. lat. « an »	+	+				
	&Ainit;		A initial		+				
	&Ch;	00C7	Ch		+				
	&ligct;		Ligature ct	+	+	+	+	+	+
	&Ebref;		E bref		+				
	&Ekom;		E komun		+				
	&eog;	0119	e-ogonek	+	+	+	+		
	&eogaigu;	0119+02CA	e-ogonek aigu		+				
	&eograve;	0119+02CB	e-ogonek grave		+				
			Pléiade		+				
			Pléiade		+				
		0118	E ogonek		+				
	&Eu;	–	Eu Pléiade		+				
.....									
	s	0073	s final	+	+	+	+	+	+
	&st;	FB06	Ligature st	+	+	+	+	+	+
	&S;	017F	s long	+	+	+	+		
	&Si;		Ligature slong+i	+	+	+	+		
	&SSi;		Ligature slong+slong+i		+	+	+		
	&Sp;		Ligature slong+p		+	+	+		
	ß ou &Ss;	00DF	Ligature slong+s		+	+	+	+	+
	&SS;		Lig. slong+slong	+	+	+	+		
	&St;	FB05	Ligature slong+t	+	+	+	+		

Figure 7. *Quelques candidats à l'inventaire du projet Cassetin*

Assez paradoxalement, si on exclut les vignettes et ornements, il ne nous semble pas que cet inventaire doive contenir un très grand nombre de nouveaux caractères (on ne tient évidemment pas compte ici des variétés d'œil dues aux différences de style, de corps, de graisse, etc. : « fl » est pour nous un seul caractère, identique à « *fl* » ou « **fl** ») !

La figure 7 montre, d'une part, quelques lettres qui pourraient faire l'objet de cet inventaire et, d'autre part, des informations sur elles : nom d'entité proposé (ceux donnés ici sont provisoires), le code Unicode lorsqu'il est connu (un « – » précise que ce caractère n'est pas dans Unicode mais devrait y être). Les croix des dernières colonnes indiquent des occurrences attestées dans le siècle correspondant et pointeront sur des références aux sources.

4. Conclusion

Nous espérons avoir ainsi montré la nécessité de ne pas perde les informations que peuvent donner la numérisation de documents imprimés et le besoin d'une normalisation du codage des caractères ainsi lus.

5 Bibliographie

Allier B., Emptoz H., « Le traitement des images au service du document patrimonial », *Document numérique*, spécial *Numérisation et patrimoine*, vol. 7, n° 3-4, (ce numéro), 2003.

André J., « The Cassetin Project – Towards an inventory of ancient types and the related standardized encoding », *Proceedings of the EuroTeX conf.* (Haralambous Y. ed.), ENST-Brest, juin, 2003, p. 165-169 (à paraître dans *TugBoat*, vol. 23).

André J., Chabin M.-A. (sld), *Documents anciens*, numéro spécial de *Document numérique*, vol. 3, n° 1-2, 1999, 169 p.

André J., Hudrisier H. (sld), *Unicode, écriture du monde ?*, numéro spécial de *Document numérique*, vol. 6, n° 3-4, 2002, 364 p.

Barber G., *French Letteerpress Printing. A list of Fench printing manuals and other texts in french bearing on the technique of letterpress printing*, Occasional publication n° 5, Oxford Bibliographical Society, 1969.

Baudin F., *L'effet Gutenberg*, éditions du Cercle de la Librairie, Paris, 1994.

Bigmore F.C., Wyman C.W.H., *A Biliography with notes & illustrations*, Holland Press Ltd (London) and Oknoll B (USA), 1978.

Coüasnon B., Camillerapp J., « Accès par le contenu aux documents manuscrits d'archives numérisés », *Document numérique*, spécial *Numérisation et patrimoine*, vol. 7, n° 3-4 (ce numéro), 2003.

Crettez J.P., Lorette G., « Reconnaissance de l'écriture manuscrite », *Techniques de l'ingénieur*, H1358, 1998.

Dreyfus J., *Type specimen facsimiles […] between the sixteenth and eighteenth centuries*, Londres, Bowes & Bowes Putnam, 1963.

Le Bourgeois F., Emptoz H., Trinh H., « Compression et accessibilité aux images de documents numérisés », *Document numérique*, spécial *Numérisation et patrimoine*, vol. 7, n° 3-4 (ce numéro), 2003.

Lefèvre Ph., « Reconnaissance de l'imprimé », *Techniques de l'ingénieur*, H1348, 1999.

Méron J., *Orthotypographie – recherches bibliographiques*, Convention typographique, Paris, 2002.

Role F. (sld), *TEI : Text Encoding Initiative*, numéro spécial des *Cahiers GUTenberg*, n° 24, 1996, http://www.gutenberg.eu.org/publications/cahiers/50-cahiers24.html

The Unicode Consortium, *The Unicode Standard, Version4.0*, Addison-Wesley, Reading, 2003. Pour une version à jour, voir : http://www.unicode.org/reports/ et pour la version française, voir http://Pages-infinite.net/hapax/