The Cassetin Project — Towards an Inventory of Ancient Types and the Related Standardised Encoding

Jacques André
Irisa/Inria-Rennes
Campus de Beaulieu
F-35042 Rennes Cedex, France
Jacques.Andre@irisa.fr

Abstract

The Cassetin Project purposes are to inventory all of the types used in Europe for centuries, to standardize their encoding or naming and to propose this list as a Unicode addendum (or at least as a private area).

Résumé

Nous décrivons le projet Cassetin d'inventaire des types utilisés en typographie européenne en vue de la normalisation de leur codage ou de leur nommage, voire de leur intégration dans Unicode (au pire dans une zone privée).

Introduction

Character encodings such as Unicode [15] make a strong difference between characters (abstract linguistic entities) and glyphs (the rendition of characters) and so ignore the "types" effectively used when composing books. Typical examples are ligatures (such as "ct"), abbreviations or vanished characters (such as the old French "ç ę ŋ"). This absence makes it difficult to standardize OCR outputs and quite impossible to get genuine plain text from electronic editions of old books (especially Renaissance or even 18th century ones). Strangely, far more complicated texts, such as medieval manuscripts, are now electronically editable thanks to encoding projects.

We first show how encodings have been applied to medieval manuscripts, then describe the equivalent, and not favorable, situation for old books. We follow with a presentation of the Cassetin project: its aims are to inventory all of the European types, to name them and to propose this list as a Unicode candidate.

Electronic Editing of Manuscripts

More and more manuscripts are edited on the Web or on CDs: medieval charts, writers' drafts, registers of births, etc. Most of them are only available in image mode. Some are accompanied with the corresponding text, or rather with the corresponding "texts" when the edition is undertaken by humanities scholars. Indeed they are different views of the same text.

Let us take as an example a manuscript of Bernard

Fig. 1: Manuscript of De Ventadour (France, 12th century)

[...] chanter et vint conter et enseigner.
[...]to sing and came singing and teaching.
2-a: Modern translations

[...]chantar et venc cortez et enseingnatez.
2-b: Modern composition

[...] châ-
tar 7 venc cortef 7 ēfeingnatz.
2-c: Diplomatic edition

Fig. 2: Various editions of figure 1 (last two lines of text)

```
... ch&an;<br>tar &et7; venc corte&longz; &et7;
&en;&longs;eingn<unclear reason="stain">a</unclear>tz.<br>
```

Fig. 3: TEI encoding of figure 1

GLYPH	MUFI		ISO	UNICODE	
	ENTITY	NAME	ENTITY	CODE	NAME
	&con;	SIGN CON		0254	SMALL OPEN O
	✗	SIGN CROSS		271D	LATIN CROSS
	&ed;	SIGN ED	;	003D	SEMICOLON
	&est;	SIGN EST		F150	HOMOTHETIC
	&etslash;	SIGN ET WITH SLASH			

Fig. 4: Example of MUFI proposal for encoding Nordic medieval abbreviations

de Ventadour[1] and even let us consider just the last two lines of Figure 1. They can be encoded in various ways. In reverse order, we can first translate these lines into modern French or even into modern English (figure 2-a). A student in medieval philology would prefer a form closer to the genuine language, like figure 2-b, where each sign is composed with modern types, without respect to specific handwritten signs. A way[2] closer to the manuscript is to edit the text with types close to the handwritten signs (figure 2-c) and to respect the layout (hyphenation, etc.).

In the end, a researcher needs far more, such as ligatures or signs used, unclear letters, who wrote or annotated the original text and who has translated or encoded it. Many projects have been launched to access electronic versions of medieval manuscripts.[3] They generally use SGML-like tags to indicate either the structure of the document or the actual characters used. Figure 3 shows how our two lines could be encoded.[4] Tags like `&et7;` or `&an;` indicate special characters ("et" and "an") and are like the ones used by Project Charrette at Princeton University.[5] Other projects use other tags (e.g., `&abar;` instead of `&an;`). That is why some new projects, such as MUFI[6] (figure 4), try to inventory all of these signs, to give them names (used as entity names), to propose[7] them as Unicode characters and finally to propose a font offering all of these glyphs.

From now on, thanks to such projects, the various electronic editions of manuscripts should follow a standardized encoding, hence become portable and usable by any researcher.

1. A French troubadour (12th Century). Here is an example of a version of his *Quand vei la laudeta mover...* (When the skylark wings...), a song written in "Provencal" (an Old French dialect).

2. This is close to the so called "diplomatic" version, generally used for modern writers' drafts, that takes care as well of the position (orientation, length, etc.) of lines in pages.

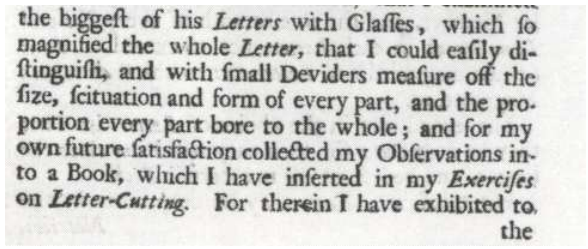
3. Such as Digital Scriptorium (`http://sunsite.berkeley.edu/Scriptorium/`), EAMMS (`http://www.csbsju.edu/hmml/eamms/`), or the TEI based project MASTER (`http://www.cta.dmu.ac.uk/projects/master/`).

4. That form is rich: it can be translated to forms 2.a to 2.c, while the opposite way is generally impossible.

5. The list of entities may be found at `http://www.mshs.univ-poitiers.fr/cescm/lancelot/keys.html`.

6. Medieval Unicode Font Initiative, `http://www.hit.uib.no/mufi/`.

7. See section "Unicode" below for the position of the Unicode Consortium regarding glyphs and characters.


the biggeſt of his *Letters* with Glaſſes, which ſo magnified the whole *Letter*, that I could eaſily diſtinguiſh, and with ſmall Deviders meaſure off the ſize, ſcituation and form of every part, and the proportion every part bore to the whole; and for my own future ſatisfaction collected my Obſervations into a Book, which I have inſerted in my *Exerciſes* on *Letter-Cutting*. For therein I have exhibited to the

the biggest of his *Letters* with glasses, which so magnified the whole *Letter*, that I could easily distinguish, and with small dividers measure off the size, situation and form of every part, and the proportion every part bore to the whole; and for my own future satisfaction collected my observations into a book, which I have inserted in my *Exercises* on *Letter-Cutting*. For therein I have exhibited to [...]

5-a: Modern transliteration

```
the bigge&st; of his <ital>Letters</ital> with
Gla&ss;es , which &longs;o<br>
 magni&fi;&ed the whole <ital>Letter</ital>,
that I could ea&longs;ily
di&hyphen;<br>&stlig;tingui&shlig;, [...]
[...]
&longs;atisfa&ctlig;ion collec&ctlig;ed  [...]
```

5-c: TEI-like encoding

FIG. 5: Extract from Moxon's *Mechanick Exercises* (1683), (volume 2 *Applied to the Art of Printing*) followed by a modern transliteration (5-a) and a TEI-like encoding (5-c)

Electronic Editions of Old Books

Probably because printed characters are thought to be more readable than handwritten ones, quite often digitized books are edited as images and/or as plain text encoded with modern characters. If some other types are used, no attempt is made to standardize their encoding.

Moxon's Mechanick Exercises Let us consider first the famous Moxon's book (1683, a pioneer in the matter of printers' manuals [4]) *Mechanick Exercises* [14]. Figure 5 shows an extract and the corresponding "plain text" version of it: it is a modern version, with modern words, no emphatic caps, no ligatures, etc. As for manuscripts, it is possible to encode this text (figure 5-c) to take care of specific characters (like ct), hyphenation (tag `&div;`), etc.

However, many people would like an edition (like figure 6) that is both more legible than TEI-encoding (figure 5-c) and more authentic than a traditional plain text.[8] Alas, if one looks in minute detail at these glyphs, one can see that if ligatures such as "fi," "long s + i," "long s + t" or "ct" are present, other ones are absent, such as

8. Furthermore, figure 5-a does not allow any research on the use of the long s or of old words (e.g., scituation).

the biggeſt of his...
magnified the whole...
diſtinguish, and with ſmall...
ſize, ſcituation...
...
own future ſatisfaſtion...

FIG. 6: Moxon text (figure 5) edited with convenient glyphs[9]

FIG. 7: Fertel lower case (1723)

"sh": it is quite impossible to find a font with all the usual (old) glyphs! It is not a matter of difficulty to design these characters, but simply that designers do not know them!

Note that today OCRs still have problems recognizing old types; however, current research allows one to think that very soon OCRs will be able to recognize Renaissance characters and their ligatures. What will be the output when recognizing the ligature "ct"? A code (which one? nnn_8, $xxxx_{16}$?) or a name? Will it be "ct" (if so, any study on ligatures will be impossible)? Or "`&ctlig;`"? Or "`&ct;`", or "`\ct`", or ...? Standardized naming or encoding is required!

Another example: Fertel's case Figure 7 exhibits the lower case used by Fertel [8, page 12] in the first quarter of the 18th century. Among other special characters (such as ligatures "si", "ssi" or "ct" and typically French characters like "ç" and "œ"), it offers an "ę" (left upper corner) that has nothing to do with the "e ogonek" used in eastern Europe. It is an echo of a former system to write French (by a grammarian and poet, Jean-Antoine de Baïf, 16th century) and this kind of "breve e" was used for less than a century.[10] Obviously, this character is not an isolated example and you can find many characters that are neither in today's encodings nor in font character

9. Here, as in the rest of these proceedings, the *HW Caslon* font is used, designed by Howard Jones.

10. 50 years after Fertel, Diderot's *Encyclopédie* shows the same case, however the "ę" is replaced by an "é" while the genuine box for "é" is empty!

sets. Yet, they are present in foundry specimens, books or even in grammars ...

Unicode, Characters and Glyphs

Unicode [15] makes a strong difference between characters (abstract linguistic entities) and glyphs (a possible physical stylistic representation or rendition of these entities). Very few clever papers give a good explanation of those concepts; let us cite here one by Ken Whistler, the technical director of Unicode [2] and one by a typographer, John Hudson [12]. On the other hand, there are also good papers that say that Unicode made the wrong choice and that characters and glyphs are not so easily different [10, 11]. We would like to add that "types" (with the usual typographic meaning) are neither characters, nor glyphs.[11]

An important point is that the Unicode principle that separates glyphs and characters has been historically violated by another one: Unicode is based on previous encoding systems (proprietary or international standards) where ligatures were present. If Unicode was clean, even the sign "&" should not be there! However we can be suspicious why "long s" and even "ligature st" have been very recently added and not "ligature ct"!

Imagine the dialog:
– "How could I describe[12] Fertel's case (figure 7) and its ę using Unicode?" I ask.
– Answer from Unicode specialist: "Use `latin small letter e with ogonek`, U+0119."
– "No, I say, Fertel's character is not that character, there is the same glyph resemblance as with `latin capital a` and `greek capital alpha`, but they are different characters and Unicode encodes them separately."
– "Why don't you encode this character as letter e and a combining diacritic ogonek?"
– "For it is not an ogonek, rather a kind of breve," I answer.
– "OK," he says, "your ę is a glyph of some `latin small letter with breve`."
I disagree, it's not the same breve as the one used by Fertel in another case: "ĕ", so it's not the same character ... And now, if you look at the alphabet given by the same Baïf, you can see an "`a with raising tail`" that is rather a nasal O (its place in the alphabet is just before the P letter). Let us restart the same dialog ...

Last point: Unicode knows old languages such as the Runes or Ogham. Why should it ignore old European languages and their writing used for centuries?

The Cassetin Project

Being involved in digitization projects such as Fournier's *Manuel typographique*,[13] I am continuously confronted with such problems of coding or naming old[14] characters. Discussions with many people involved in such tasks pushed me recently to undertake a project[15] to inventory these types and try to establish a standardized list of names or ... codes.

Its main aims are:

Inventory of types Prepare an inventory of all types used in texts[16] printed in European[17] languages.

Typical characters are

- Ligatures, such as the ones already quoted here (sh, si, st, ...) and many other ones (like the Hungarian gz ...).
- Vanished characters, such as the "ę," the tailed A, etc.
- Accented characters (like the old Spanish consonants).
- Abbreviations.
- Special characters such as verset and respons (these two are in Unicode, but many other special characters are not).
- Historical typographical characters[18] (that are not already in Unicode) such as raised letters.

This inventory is based on

- Previous studies, such as [3, 4, 5, 7], including Web pages such as Bolton's on cases [6].
- Specimens published by foundries.
- Ancient books.
- The MUFI project for manuscripts!

11. There are many stylistic variants of our "ę"! On the other hand, Unicode speaks about rendition of abstract characters. However, what about the other way: when scanning documents, printed characters exist before the corresponding "abstract" character, they are not only images of abstract characters, they are characters by themselves at an intermediary level between glyphs and linguistic entities.

12. Even if "[t]he Unicode Standard is explicitly *not* aimed at being a system for facsimile representation of text" [2], one may need to quote such a character. Actually, it is not only a Unicode problem!

13. Like Moxon's, a famous 18th century book on type-cutting and typefounding. See [9, 13] and `http://www.irisa.fr/faqtypo/BiViTy`.

14. Old means here before DTP! A typical example is the use, still current in 1950, of the abbreviation "crossed K" that represents the Breton "ker" occurring in many names.

15. Temporarily called CASSETIN: "cassetin" is the French name of case boxes. It can stand for "CASSE Type encodING" ... See also [1].

16. One problem not yet solved: should we consider all types, even the ones used outside of plain text, such as ornaments and rules? I do not think so, however the limits are not yet fixed!

17. This is again an unsolved question: Which languages do we consider? Latin ones? What about Cyrillic, Greek, Hebrew, Arabic, Syriac, etc.? Actually, today it is only a matter of specialists working in this project ...

18. We do not dare to speak about small caps!

Naming and Encoding Each identified type will enter a file with typical glyphs, usages, etc. and a name will be given. This name will be usable as an XML entity (&xxx;), a TEX name (\xxx), an (Adobe or OpenType) glyph name, etc.

Each type will as well be assigned a number: the Unicode number when it exists; if not, a new number that could be an entry in some Unicode *user private area*.

Note that these glyph/character names permit not only standardizing output from OCR, but as well a standardized way to type in these special signs for rendition.

Experimental Font As MUFI does, an existing font should be upgraded to offer all of these, say, glyphs: editing "facsimile" texts like those in figures 5 or 6 should be easier! Again, it's not only a matter of actual glyphs, but rather, to be standardised, a matter of table encoding. In that way, TEX LMs or Ω could be good candidates!

Calendar This project is still a private undertaking, however, I'd like to make it an international project[19] (with, e.g., European Union help).

A quick glance at the already published cases shows that the number of new characters to inventory is not very large and the list should not be as large! So this work should not last years ...

Conclusion

Glyphs or not, characters or not, types belong to a class that is not recognized by Unicode. Historians of books, of languages, etc. do need a standardization of their names, even of their encodings, in such a way that the increasing number of sites offering digitized books can be researched in a portable manner.

References

[1] André J. (2003), "Numérisation et codage des caractères de livres anciens", *Numérisation et patrimoine* (B. Coüasnon éd.), special issue of *Document numérique*, vol. 7, num. 3–4, 2003, pp. 127–142.

[2] Andries P. (2002), "Entretien avec Ken Whistler, directeur technique du consortium Unicode", *Document numérique*, vol. 6, 3-4, pp. 13–49. http://hapax.iquebec.com/hapax/

[3] Barber G. (1969), *French Letterpress Printing. A list of French printing manuals and other texts in French bearing on the technique of letterpress printing*, Occasional publication num. 5, Oxford Bibliographical Society.

[4] Baudin F. (1994), *L'effet Gutenberg*, Éditions du Cercle de la Librairie, Paris.

[5] Bigmore F.C. and Wyman C.W.H. (1978), *A Bibliography of Printing with Notes & Illustrations*, Holland Press Ltd (London) and Oak Knoll Books (USA).

[6] Bolton D. (2002), *Type cases*, The Alembic Press, http://members.aol.com/typecase/

[7] Dreyfus J. (1972), *Type specimen facsimiles*, London.

[8] Fertel M.D. (1723), *La science pratique de l'imprimerie* ..., Saint Omer. Facsimile by Libris editions, 1998.

[9] *Fournier le jeune* (1764 et 1766), *Manuel typographique utile aux gens de lettres et à ceux qui exercent les différentes parties de l'art et de l'imprimerie*, Paris, chez Barbou, 2 tomes.

[10] Haralambous Y. (2000), "Unicode, XML, TEI, Ω and Scholarly Documents", 16th International Unicode Conference, Amsterdam, 24 p.

[11] Haralambous Y. (2002), "Unicode et typographie : un amour impossible", *Document numérique*, vol. 6, n° 3-4, p. 105–137.

[12] Hudson J. (2002), "Unicode, from text to type", *Language Culture Type — International Type design in the Age of Unicode* (Berry ed.), ATyPI/Graphis, pp. 24–44.

[13] Mosley J. and Carter H. (1995), *The Manuel Typographique of Pierre-Simon Fournier le jeune*, 3 volumes, Darmstadt.

[14] Moxon J. (1683), *Mechanik Exercises or the Doctrine of Handy-works*, printed for Joseph Moxon on the Westside of Fleet-ditch, at the Sign of Atlas. Misc. eds. by Davis and Carter (Dover, 1978), De Vinne (1886), Johnson and Gibson (Oxford University Press, 1978), etc. Note that the second volume, *Applied to the Art of Printing*, is more difficult to find.

[15] The Unicode Consortium (2003), *The Unicode Standard, Version* 4.0, Addison-Wesley, ISBN 0-321-18578-1. See also http://www.unicode.org.

19. One goal of this paper is a call to participation! Don't hesitate to mail to the author.