

18 – Jeux de caractères, signes diacritiques, etc.

Pour ceux qui travaillent avec des formes classiques des langues européennes, les recommandations de la TEI concernant l'emploi des jeux de caractères sont simples. Pour un emploi local, il suffit d'utiliser tout jeu de caractères supporté par la machine et le logiciel. Au cas où le logiciel rend difficile la saisie directe des caractères spéciaux au clavier, il est possible de définir des conventions propres de saisie (par exemple, représenter les lettres accentuées en dactylographiant l'accent approprié aussitôt après la lettre, ou en utilisant des séquences spéciales qui n'ont que peu de chance d'apparaître dans le texte normal, tel que «aE» pour «ä»). Des fonctions de recherche et de remplacement globales peuvent être ensuite utilisées pour transformer ces raccourcis en des caractères corrects¹. Si l'on doit employer des écritures non latines et qu'il existe un système de translitération normalisé dans le domaine particulier (par exemple, pour le grec ancien, le code beta du *Thesaurus Linguae Graecae*), il faut utiliser cette norme. Toute translitération employée devrait être réversible (ce qui exclut un nombre surprenant de schémas employés communément dans l'écriture normale), et son utilité sera plus grande si elle ne nécessite aucune ligature spéciale ni lien ni signe diacritique (ce qui exclut un nombre surprenant des schémas restants...).

Pour l'échange de fichiers entre des systèmes, seules les références d'entité SGML sont à employer pour remplacer tout caractère ne figurant pas dans la liste de caractères ci-dessous (les caractères de cette liste sont ceux qui peuvent être échangés sans perte d'informations entre la plupart des systèmes):

```
a b c d e f g h i j k l m n o p q r s t u v w x y z
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
0 1 2 3 4 5 6 7 8 9
" % & ' ( ) * + , - . / : ; < = ? _ (space)
```

Cette liste exclut les caractères suivants

```
! # $ [ \ ] ^ _ ` { } | ~
```

qui, dans bien des cas et au grand mécontentement des utilisateurs non avertis, ne survivent pas aux transferts au-delà des frontières nationales ou à travers les réseaux longue distance².

1. C'est ce que nous faisons dans la version française de ce document où nous avons utilisé les caractères accentués é è œ À etc. au lieu de leur codage TEI.

2. Ces caractères font en fait partie de l'IVR (*International Reference Version*) du codage Ascii/Iso 646. Voir le *Cahier GUTenberg* n° 20 sur ces problèmes de codage.

Par contre, pour un simple transfert entre Mac et PC, ces caractères pourront peut-être être échangés sans dommage.

Afin d'assurer la transmission correcte à travers des réseaux hétérogènes, des références d'entité doivent être employées pour tous les caractères latins accentués et étendus, pour tous les caractères non latins, et enfin pour tous les symboles ne figurant pas sur un clavier d'ordinateur conventionnel.

Si on le désire, on peut employer ses propres noms d'entité SGML dans des fichiers conforme à la TEI, à condition de fournir des mentions standard d'entité SGML à leur place ; mais les noms standard, (quoique longs ou compliqués) ont l'avantage de la clarté ; ces noms sont parlants pour tout locuteur anglophone³ qui peut reconnaître qu'il s'agit d'un nom de caractère, souvent même sans recours à une liste. Notons que ce n'est pas le cas de beaucoup d'autres méthodes employées pour représenter des caractères accentués.

Les noms d'entité requis pour les caractères présentés ci-dessus comme peu sûrs, et pour les caractères accentués de certaines langues majeures de l'Europe occidentale, sont donnés ci-dessous. Les listes des jeux d'entité publics ainsi que leur contenu sont disponibles dans tout ouvrage de référence traitant de SGML : les noms donnés ci-dessous sont extraits des jeux d'entité publics ISO, sont largement employés et sont donc recommandés.

Lorsqu'un caractère ne paraît pas dans les jeux d'entité public, on peut désirer générer un nom, au moyen des mêmes conventions de nommage employées dans les jeux d'entité publics ISO, comme ici :

digrammes

créer les noms d'entité pour les digrammes en adjoignant la chaîne *lig* aux lettres formant le digramme ; si une forme capitalisée est nécessaire, les deux lettres sont rendues en majuscules (rappelons que la casse est habituellement significative dans des noms d'entité) ;
exemple : *aelig* (æ), *AElig* (Æ) *szlig* (ß) ;

signes diacritiques et accents

créer les noms d'entité pour des lettres accentuées dans la plupart des langues européennes occidentales en adjoignant une des chaînes suivantes à la lettre qui porte l'accent, celle-ci pouvant être en majuscules ou en minuscules ;

umlaut employer *uml* pour umlaut ou trema⁴ : par exemple *auml* (ä), *Auml* (Ä), *euml* (ë), *iuml* (ï), *ouml* (ö), *Ouml* (Ö), *uuml* (ü), *Uuml* (Ü) ;

3. Il est donc très important que le français soit aussi accessible « naturellement » grâce justement à ces mentions d'entités.

4. Notons toutefois que ces deux symboles ne sont pas, typographiquement parlant, équivalents : le *umlaut* allemand est plus bas, plus proche, de la voyelle que le trema français {NdT}.

-
- acute** employer *acute* pour l'accent tonique ou aigu : par exemple *aacute* (á), *eacute* (é), *Eacute* (É), *iacute* (í), *oacute* (ó), *uacute* (ú) ;
- grave** employer *grave* pour l'accent grave : par exemple *agrave* (à), *egrave* (è), *igrave* (î), *ograve* (ó), *ugrave* (ù) ;
- circumflex**
employer *circ* pour circonflexe : par exemple *acirc* (â), *ecirc* (ê), *Ecirc* (Ê), *icirc* (î), *ocirc* (ô), *ucirc* (û) ;
- tilde** employer *tilde* pour tilde : par exemple *atilde* (ã), *Atilde* (Ã), *ntilde* (ñ), *Ntilde* (Ñ), *otilde* (õ), *Otilde* (Õ) ;
- consonnes**
les noms d'entité suivants sont recommandés pour certaines consonnes spéciales utilisées dans les langues de l'Europe de l'ouest :
ccedil (ç), *Ccedil* (Ç),
eth (eth bas de casse : le d croisé «ð» anglo-saxon ou islandais), *ETH* majuscule : «Ð»,
thorn (thorn minuscule : «þ»), *THORN* (thorn majuscule : «Þ»),
szlig (ligature s-z allemande ou *esszett* : ß) ;
- signes de ponctuation**
les noms d'entité suivants sont recommandés pour certains signes de ponctuation communément rencontrés :
ldquo (*left double quotation mark* guillemet double gauche anglais : ««»),
rdquo (*right double quotation mark* guillemet double droit anglais : «»»),
mdash (*one-em dash* – tiret d'un cadratin «—»),
hellip (*horizontal ellipsis* – points de suspension horizontaux «...»),
rsquo (*right single quote* – signe anglais de citation droite «'») ;
voir également la liste des «caractères dangereux» juste ci-après et la figure 1 des caractères français ;
- caractères «dangereux»**
les caractères présentés ci-dessus (page 104) comme dangereux pour la transmission sur des réseaux internationaux académiques et publics peuvent être représentés par les entités suivantes :
excl (!), *num* (#), *dollar* (\$),
lsqb (*left square bracket* – crochet gauche [),
bsol (*back-slanted solidus* – barre de fraction inverse : \),
rsqb (*right square bracket* – crochet droite]),
circ (*circumflex* – circonflexe, ^),
lsquo (*left single quotation mark* – fin de citation gauche),
grave (accent grave), *lcub* (*left curly bracket* – accolade gauche, {),
rcub (*right curly bracket* – accolade droite, }),
verbar (*vertical bar* – barre verticale, |),
tilde (~).

En résumé, pour le français⁵, les codages utiles à connaître (mais rappelons le, le codeur ne devrait normalement pas s'en soucier) sont les suivants :

FIGURE 1 – Codage des caractères français

à	à	À	À
â	â	Â	Â
é	é	É	É
è	è	È	È
ê	ê	Ê	Ê
ë	&euuml;	Ë	Ë
î	î	Î	Î
ï	ï	Ï	Ï
ô	ô	Ô	Ô
ù	ù	Ù	Ù
û	û	Û	Û
ü	ü	Ü	Ü
ç	&ccdel;	Ç	&Ccdel;
æ	æ	Æ	Æ
œ	œ	Œ	Œ
«	&lDaq;	»	&rdaq;
–	–	—	—

5. D'après le *Lexique des règles typographiques en usage à l'Imprimerie nationale*, Imprimerie nationale, Paris, 1990 (p. 102). Notons toutefois que les caractères « œ Œ » et « æ Æ » ne sont pas des ligatures, facultatives, mais de vrais caractères : voir *Cahier GUTenberg* n° 22 à ce sujet [NdT].