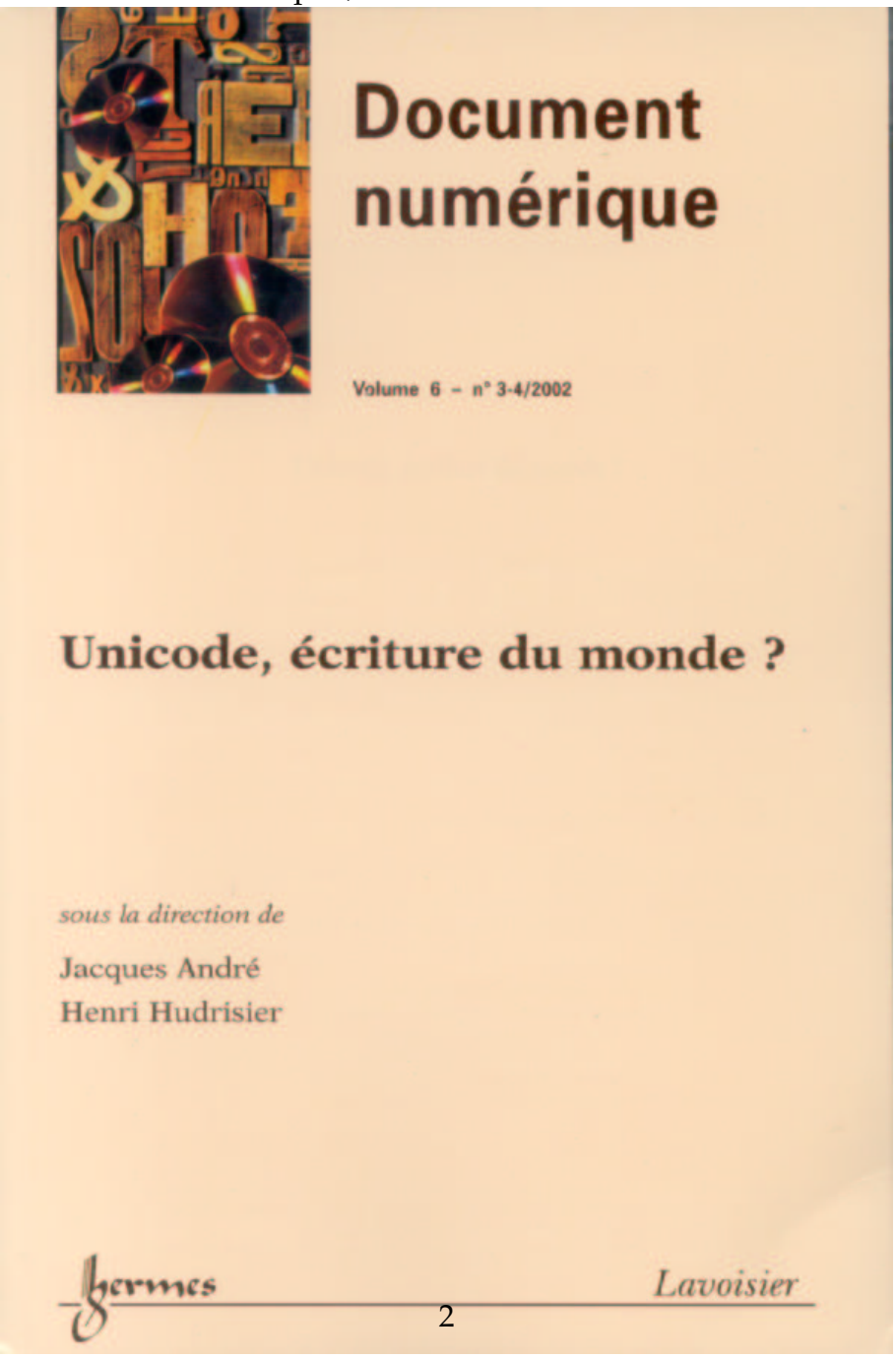
Jacques ANDRÉ
« Caractères, codage et normalisation – de Chappe à Unicode »,
Document numérique, Éditions Lavoisier+Hermès, ISBN 2-7462-0594-7, vol. 6, nº 3-4, 2002, p. 13-49.

Pour le sommaire complet, voir www.e-revues.Lavoisier.fr



Caractères, codage et normalisation

De Chappe à Unicode

Jacques André

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

RÉSUMÉ. *La transmission de l'information s'est faite d'abord de façon visuelle (fanions de la marine, télégraphe Chappe, Morse, etc.) avant d'être électrique (Télex) puis informatique. Nous présentons l'évolution des divers codages de caractères liés et les normes associées (IA1, Télex, BCD, Ascii, Latin-n et enfin Unicode). Cet aperçu historique nous permet de préciser à l'occasion divers concepts tels que codage, caractère, glyphe, norme, standard, etc.*

ABSTRACT. *Up to the 20th century, information was transmitted through visual mediums such as navy's flags, Chappe's then Morse's telegraph, etc. Then electricity was used (telex) and now computer networks. Here is shwn the evolution of corresponding character encodings and associated standards (IA1, Telex, BCD, ASCII, Latin-n, and now Unicode). Emerging concepts, such as encoding, characters, glyphes, (proprietary) standards, etc. are explained as they occur.*

MOTS-CLÉS : *normes, Unicode, caractère, codage, glyphe, Ascii, ISO, Latin-1, histoire.*

KEYWORDS: *Unicode, standards, characters, glyphes, ASCII, ISO, Latin-1, history.*

Document numérique. Volume 6 – n° 3-4/2002, pages 13 à 49

Unicode[1] est en train de devenir un gros succès mondial en matière de codage multilingue de toutes les écritures du monde. Mais certaines critiques ont d'ores et déjà été émises et ce numéro spécial de *Document numérique* va les approfondir. Si nombre d'entre elles sont fondées, nous avons l'impression que certaines relèvent d'ambiguïtés liées aux concepts de caractères, de codages, de normes, etc. et de ce que chacun d'entre nous attend d'un tel standard. Nous allons donc essayer de dégager, avec une vision historique, quelques-uns de ces malentendus qui existent en fait depuis des siècles[2] !

1. Quelques concepts

1.1. *Polysémie du mot « caractère »*

Le mot « caractère » a de nombreux sens[3].

– En linguistique des langues occidentales, un caractère correspond à l'unité minimale (caractère abstrait pour Unicode) ayant un sens et à laquelle correspond un graphème.

– On y emploie aussi ce mot caractère pour les « idéophonogrammes ».

– En typographie, c'est un synonyme de « type », la pièce de métal servant à imprimer.

– C'est aussi l'œil, la partie en relief sur le type qui, après encrage, donne une image (la trace imprimée) qu'on appelle aussi caractère et qu'en Unicode on va appeler glyphe (voir ci-dessous section 1.2).

– En typographie toujours, caractère a aussi le sens de « fonte[4] » et on dira par exemple « le Garamond est un caractère qui … ».

– En informatique, on parle aussi, par abus de langage, du « caractère 41 », pour le caractère dont le code est 41 ; etc.

---

1. Tout au long de cet article, sauf mention contraire, Unicode veut dire en fait ce qu'il y a de commun entre le standard propriétaire Unicode (voir section 1.4. pour la distinction entre standard et norme) et la norme ISO/CEI-10646 (voir section C de [UNI00]). Précisons à ce sujet que les références […] renvoient à la bibliographie « générale » à la fin de cet article tandis que certaines références spécifiques pourront être mises en bas de page. Enfin, signalons que les dernières pages de ce numéro spécial contiennent une liste d'abréviations et sigles.

2. Cet article est basé sur [AND96] et [AND01] et a fait l'objet d'un exposé lors d'une journée Mediadix sur Unicode (mai 2002, Université de Paris 10).

3. On trouvera dans http://iquebec.ifrance.com/hapax/glossaire.htm de nombreuses définitions de ce mot en relation avec Unicode. Jacques Anis a, longuement abordé ces problèmes : *L'écriture – théorie et descriptions*, De Boeck-Wesmael, Bruxelles, 1988.

4. La terminologie en la matière est plutôt vague entre fonte, police, caractère, etc. Police de toutes façons a signifié initialement « inventaire du nombre de sortes » (par exemple 5000 a, 2000 b, etc.) contenues dans une casse. C'est le même mot (d'origine italienne) que l'on retrouve dans « police d'assurance ». Il ne faut pas croire que les Anglo-saxons aient une terminologie plus rigoureuse (voir par exemple *fo(u)nt, typeface, character*, etc.).

1.2. *Caractères et glyphes*

On vient d'y faire allusion, la notion de caractère en linguistique rassemble au moins deux entités différentes : le caractère abstrait et le graphème physique. De même en typographie distingue-t-on l'image imprimée (par exemple « M ») de ce qu'elle représente (« M majuscule[5] »).

1.2.1. *Définitions*

Bien qu'implicite dans toutes les normes d'échanges de caractères, cette distinction a été fermement prononcée par Unicode qui distingue explicitement « glyphe » et « caractère ». Voici des définitions possibles, sans aucune dépendance ni relation d'ordre.

Caractère	unité d'information abstraite utilisée pour coder des éléments de texte
Glyphe	forme géométrique utilisée pour présenter graphiquement des morceaux de texte

On trouvera dans ce numéro spécial de nombreux détails sur cette distinction [AND02, HAR02, RAN02]. Donnons-en ici quelques explications préliminaires.

Ces définitions permettent donc de distinguer le caractère abstrait « M majuscule » de la quasi-infinité de glyphes que sont ses diverses formes imprimées (par exemple M *Times-Italique* corps 10 : « *M* », M *Zapf-Chancery romain* corps 12 : « $\mathcal{M}$ » ou M *Courier gras italique* corps 8 : « 𝙈 »). Bien sûr, cette distinction s'applique aussi à des caractères autres que les lettres (tels que § [½ € ♪ ⇒ ⋟ etc.) et n'a aucune raison d'être limitée aux langues alphabétiques européennes ! En fait cette notion de glyphe correspond à celle d'« œil » en typographie française. L'encyclopédie *La chose imprimé* [DRE77] donne par exemple comme première définition à œil : « Quelle que soit l'origine d'une composition (chaude ou froide), l'œil des caractères est ce que l'on voit sur le papier. L'œil d'un A ou d'un a, d'un B ou d'un b, etc. est le signe imprimé permettant d'identifier[6] chacune de ces lettres respectivement en tant que A, a, B, b, etc. ». Mais le néologisme[7] américain « glyphe » commence à être très employé aussi le gardons nous ici. D'autant qu'il a quand même l'avantage de supprimer la polysémie du mot œil !

---

5. Ici, nous ne donnons volontairement aucun nom de caractère d'Unicode !

6. Cette identification est loin d'être triviale ! « *What is the a-ness of an a?* » [qu'est-ce qui fait qu'un a est un a ?], c'est un peu sur ce thème qu'ont longuement discuté le mathématicien typographe Donald Knuth (The concept of a meta-font, *Visible Language*, XVI, 1,1982, p. 3-27 ; traduit en français : « Le concept de Metafonte », *Communication et langage*, n° 55, 1983, p. 40-53) et le philosophe Douglas Hofstadter (*Ma thémagie,* InterEditions, 1988).

7. Bien sûr on trouve la racine grecque « glyphe » dans hiéroglyphe (avec le sens de gravure) et dans les « glyphes mayas » (en relief).

1.2.2. *Relations glyphes-caractères*

À un caractère correspond souvent un glyphe. Mais…
– Il y a des caractères sans glyphes (typiquement les caractères de commande[8]).
– Un caractère peut être représenté par plusieurs glyphes ; ainsi le caractère « a accent circonflexe » peut-il être composé à l'aide du glyphe « a » et de celui « ^ ». C'est aussi le cas des caractères composites utilisés, par exemple, pour les formules mathématiques où il existe un seul caractère « intégrale », mais selon le corps de l'intégrale (l'expression qu'elle contient) on utilisera le glyphe « ∫ » ou plusieurs glyphes en nombre variable : une crosse supérieure « ⌠», plusieurs barres verticales « ⎮ » et une crosse inférieure « ⌡ ».
– Un glyphe peut représenter plusieurs caractères ; typiquement la ligature « fl » peut être un glyphe pour la suite des deux caractères « f » « l ».
– La relation glyphe-caractère n'est pas biunivoque ; par exemple le glyphe « A » peut correspondre (de façon ambiguë sans contexte) au caractère « lettre latine A majuscule » aussi bien qu'à « lettre grecque ALPHA majuscule » ; de même à la vision rapide[9] du seul glyphe « o » ne peut-on décider s'il s'agit de la « lettre minuscule latine o », du symbole « degré », du « rond en chef » (le petit o au dessus de « Å »), d'un « o supérieur » comme dans l'abréviation n°, d'une « puce creuse » ou « boulet blanc » , voire du « chiffre zéro[10] ».
– Cette relation caractère-glyphe peut dépendre de la direction d'écriture. Au caractère abstrait « parenthèse ouvrante » correspond en français le glyphe « parenthèse gauche (» tandis qu'en arabe c'est « parenthèse droite) ».
– La distinction caractère-glyphe est très utile pour les caractères contextuels simples (lettres initiales, médiales ou finales en arabe, grec, français ancien, etc.) ou complexes (devanagari).

Cette distinction caractères-glyphes conduit à deux principes pour la création d'un jeu de caractères normé :
1. Même si des candidats au codage sont visuellement identiques (comme le A majuscule latin et le alpha majuscule) et pourraient de ce fait être représentés par un même glyphe, ils doivent quand même être codés séparément pour avoir une correspondance biunivoque entre majuscules et minuscules dans un alphabet donné et pour garantir une invariance aller-retour des données avec les normes existantes.
2. Les variations de forme (des glyphes multiples) exigées par une présentation de qualité supérieure d'un texte *ne doivent pas* être codées comme des caractères séparés si leurs significations sont identiques.

---

8. Ces caractères sont souvent dits « de contrôle », mais à tort car *to control* est un faux-ami signifiant « commander ». Voir ci-dessous section 4.2.1.
9. En fait, un œil exercé distingue le o supérieur º du signe degré °, ce dernier étant rond.
10. Cette ambiguïté s'applique aussi au chiffre 0 et à la lettre O, à tel point que la façon de noter ces signes sur les bordereaux de perforation a fait couler beaucoup d'encre dans les années 1960, l'école la plus forte préconisant l'emploi de Ø pour la lettre O et de O ou 0 pour le chiffre.

Enfin, les limites de cette distinction glyphe-caractère sont parfois difficiles à discerner aussi le codage d'Unicode parait parfois incohérent [HAR02, RAN02] surtout lorsque d'autres principes d'Unicode entrent en lice.

1.3. *Textes*

On peut dire, en première approximation, que des standards comme Unicode servent à coder les textes.

1.3.1. *Visions diverses du concept de texte*

Mais il y a beaucoup de façons de voir le même texte. Prenons comme exemple, en cette année de bicentenaire, un extrait des *Misérables* de Victor Hugo[11]. Pour l'auteur, ce texte est le résultat d'un manuscrit qui a évolué dans le temps (figure 1).

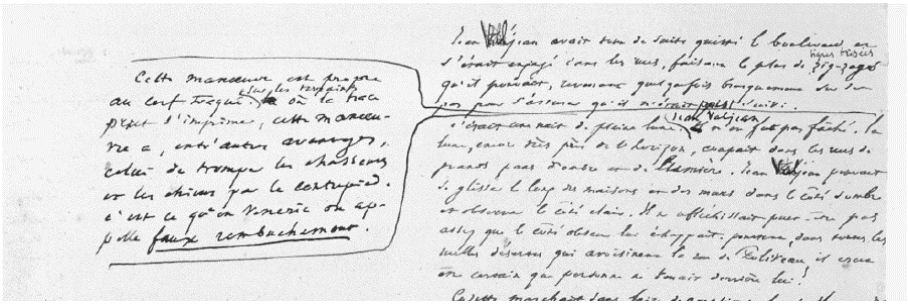


Figure 1. Manuscrit de Victor Hugo

Jean Valjean avait tout de suite quitté le boulevard et s'était engagé dans les rues, faisant le plus de lignes brisées qu'il pouvait, revenant quelquefois sur ses pas pour s'assurer qu'il n'était point suivi.

Cette manœuvre est propre au cerf traqué. Sur les terrains où la trace peut s'imprimer, cette manœuvre a, entre autres avantages, celui de tromper les chasseurs et les chiens par le contre-pied. C'est ce qu'en vénerie on appelle *faux rembuchement*.

Figure 2. *Le même texte de Victor Hugo qu'en figure 1, imprimé*

Le texte imprimé (figure 2) n'a bien sûr plus la même allure (typo, mise en page, etc.), mais le contenu reste le même (à quelques détails d'édition près, comme le mot zig-zags du manuscrit qui n'apparaît plus dans le texte imprimé). Toutefois, on note

11. Volume 2 *(Cosette),* chapitre « Les zigzags de la stratégie », page 259 de l'édition de 1881 (Hetzel-Quantin). Les images relatives à ce texte sont issues du site de la BNF ; le manuscrit est extrait de *Brouillons d'écrivains* (sous la direction de Marie-Odile Germain et de Danièle Thibault), Bibliothèque nationale de France, 2001, p. 66.

quelques différences importantes : les divisions (traits d'union) dans le texte imprimé (« li-gnes » et « chas-seurs ») ne sont pas présentes dans le manuscrit et pourraient ne pas être présentes (ou être différentes) avec une autre justification ; de même les mots « faux rembuchement » sont soulignés dans le manuscrit (en bas dans l'édition marginale, figure 1) mais en italique dans le texte imprimé.

Sur cet exemple très simple, tout se passe comme s'il y avait trois choses :
1. Une sorte de texte abstrait[12] qui, ici, serait celui de la figure 3.
2. Des textes « visuels » (le manuscrit, les textes imprimés ou affichés sur écran).

```
Cette manœuvre est propre au cerf traqué.
Sur les terrains où la trace peut s'imprimer, cette
manœuvre a, entre autres avantages, celui de tromper les
chasseurs et les chiens par le contre-pied. C'est ce
qu'en vénerie on appelle faux rembuchement.
```

Figure 3. *Texte « abstrait »*

```
\pard\plain\widctlpar\adjustright\fs20\lang1036\cgrid
Cette man\'9cuvre est propre au cerf traqu\'e9.
Sur les terrains o\'f9 la trace peut s'imprimer,
cette man\'9cuvre a, entre autres avantages, celui de
tromper les chasseurs et les chiens par le contre-pied.
C'est ce qu'en v\'e9nerie on appelle
{\i faux rembuchement}.\par
```

Figure 4. *Codage RTF/Word pour le texte de la figure 2*

```
<P> Cette man&oelig;uvre est propre au cerf traqu&eacute;. Sur
les terrains o&ugrave; la trace peut s'imprimer, cette
man&oelig;uvre a,
entre autres avantages, celui de tromper les chasseurs et les
chiens par le contre-pied. C'est ce qu'en v&eacute;nerie on
appelle <EM>faux rembuchement</EM>.</P>
```

Figure 5. *Codage HTML du texte de la figure 2*

3. Des « textes codés[13] » décrivant les propriétés graphiques du texte visuel. La figure-2 pourrait ainsi être décrite par le texte Word/RTF de la figure 4 ou par celui HTML de la figure 5.

12. En typographie, on parlait de « texte au kilomètre », en informatique de « texte source » parfois (par abus de langage, on y reviendra en 4.5) de « texte Ascii ». Unicode parle de « texte en clair » ou texte pur *(plain text)*.
13. Ou formatés, ou balisés, ou marqués, ou … ; Unicode parle de texte « enrichi » *(fancy text)*.

Enfin, le texte envoyé à une imprimante PostScript (par exemple par Word ou comme ici par LaTeX) pourra être celui de la figure 6 où on reconnaît le texte entre parenthèses (certains caractères sont codés) le reste étant des commandes de mise en page (`515 623 y` par exemple étant un positionnement dans la page correspondant à un changement de ligne, `Fb` le passage en romain et `Fa` en italique). La division `chas-seur` est implicitement commandée par la présence du signe division « - » et par un changement de ligne ; un crénage est prévu de part et d'autre du « v » de avantage, etc.

```
639 523 a Fb(Cette)22 b(man\234uvre)c
(est)k(propre)d(au)h(cerf)g(traqu\351.)g
(Sur)g(les)515 623 y(terrains)25 b
(o\371)g(la)h(trace)f
(peut)g(s'imprimer)m(,)e(cette)j(man\234uvre)
515 722y(a,)j(entre)f(autres)h(a)n(v)n(antages,)
f(celui)g(de)h(tromper)f(les)h(chas-)515 822 y
(seurs)c(et)h(les)g(chiens)g(par)f(le)h
(contre-pied.) c (C'est)27 b(ce)e(qu'en)
515 922 y(v\351nerie)19 b(on)h
(appelle)f Fa(faux)h(r)m(emb)n(uc)o(hement)p Fb(.)
```

Figure 6. *Texte PostScript permettant d'imprimer la figure 2*

Dans ces codages, on retrouve le texte pur (avec des codages spécifiques pour certains caractères) mais dans RTF on trouve quelques commandes typographiques (choix des fontes, `\fs20`, choix de l'italique, `<i>`, etc.) alors que HTML se voulant indépendant du formatage, les deux mots « faux rembuchement » sont simplement mis en exergue (`<EM>`) : le navigateur décidera de souligner, mettre en italique, en gras, etc. En revanche, ce qui est envoyé finalement à l'imprimante (figure 6) est complètement déterminé (où couper, choix des fontes, etc.).

Mais un texte abstrait peut avoir bien d'autres interprétations que celles graphiques. Considérons cette œuvre de Victor Hugo en entier.
– Pour une bibliothèque, la notice sera par exemple celle de la figure 7.
– Pour un libraire ou un bibliophile, elle peut se résumer par la couverture de l'ouvrage (figure 8).
– Si on doit citer cet ouvrage dans une bibliographie, on écrira :

Hugo Victor, *Les misérables*, tome 2 *(Cosette)*,
J. Hetzel et A. Quantin éds., Paris, 1881.

– Un philologue, lui, sera sans doute enclin à utiliser un codage comme celui de la TEI[14] tandis qu'un spécialiste de critique génétique en fera une « version diplomatique[15] ».

14. *Text Encoding Initiative*, langage de balisage donnant des informations sur des bouts de textes non pas de façon graphique (gras, corps, etc.) mais plus sémantique (auteur, titre, date, lieu, etc.). Voir, en français, *le Cahier Gutenberg* 24 à :
http://www.gutenberg.eu.org/article50.html

– Un lexicologue enfin sera amené par exemple à trier les occurrences des mots par ordre alphabétique en respectant les usages de chaque langue (en suédois, par exemple, Å vient après Z).

Auteur	Hugo, Victor
Titre	Oeuvres complètes de Victor Hugo. Roman. VI, Les misérables, 2
Titre d'ensemble	Oeuvres complètes de Victor Hugo
Publication	Num. BNF de l'éd. de Paris : J. Hetzel, A. Quantin, 1881
Description	485 p.
Autre(s) titre(s)	[Les] misérables. 2. Cosette / Victor Hugo
Domaine	Littérature française
Identifiant	N037495

Figure 7. *Notice de la BNF*

Cette fois, les différences sont, a priori, plus grandes ou ne relèvent pas que de la simple différence des concepts caractère/glyphe. Selon le cas, « Hugo » s'écrit avec une majuscule et des minuscules (la notice de la BNF), des petites capitales (la référence bibliographique) ou des capitales de titrage (la couverture). Pour les uns, on écrit le nom de l'auteur puis son prénom, pour les autres c'est le contraire. L'alphabet n'est pas le même d'une vue à l'autre : la BNF écrit « Oeuvres[16] » tandis que la couverture montre bien « ŒUVRE ». De même, le titre de la couverture est en capitales, celui de la référence bibliographique en italique tandis que la notice propose une forme adaptée sans doute à la recherche (indexation) : « [Les] misérables », etc.

Enfin, si on avait pris comme exemple une œuvre écrite en hébreu., la tradition bibliothécaire veut que l'on compose les notices non pas dans la langue d'origine, mais en translittérant en caractères latins selon des normes précises[17].

15. Voir sur les études des manuscrits d'écrivain : Almuth Grésillon, *Éléments de critique génétique – lire les manuscrits modernes*, Presses universitaires de France, 1994 et Pierre-Marc de Biasi*, La génétique des textes*, Nathan, 2000.

16. Ne respectant pas en cela la norme Z39.47 où œ existe (on aurait pu écrire OEuvres) : http://www/niso.org/standards/standard_detail.cfm?std_id=472

17. Par exemple NF-ISO 233 pour la translittération des caractères arabes, NF-ISO 259 pour les caractères hébreux, NF-ISO 3062 pour la romanisation du japonais (kana), etc.

Figure 8. Couverture(partie supérieure) de l'édition de 1881

1.3.2.*Coder quoi ?*

Ces quelques exemples montrent donc qu'il y a plusieurs façons de voir un texte, celle de l'auteur et celle du typographe n'étant pas les seules ! Disons dès à présent que ce que prétend coder Unicode c'est ce texte abstrait, mais ni la mise en page, ni la structure hiérarchique ou graphique d'un texte ! En fait Unicode, comme tout codage de caractère, n'est pas concevable seul, c'est un codage utilisé par un autre format (par exemple Word, XML, des métadonnées, PostScript, etc.). C'est peut-être ça que ne comprennent pas toujours ceux qui critiquent Unicode !

1.4. *Normes et standards*

1.4.1. *Concepts de norme et de standard*

Le concept de norme est très ancien et correspond souvent à des besoins industriels. Il y a quelques années, l'AFNOR avait montré l'utilité des normes par une carte de vœux où l'on voyait un Père Noël incapable de mettre un cadeau dans une cheminée : les deux n'étaient pas à la même norme point de vue dimensions !

Restons dans le contexte des caractères : le 28 février 1723, à la demande de la Chambre syndicale de la librairie, le Régent signe une ordonnance réglementant les dimensions physiques des caractères d'imprimerie et notamment la « hauteur en papier[18] » :

Veut Sa Majesté que six mois après la publication du présent règlement, tous les Caractères, Vignettes, Réglets et autres ornements de fonte servant à l'imprimerie, depuis le Gros-Canon jusqu'à la Nompareille, soient fondus d'une même hauteur en papier fixée à dix lignes géométriques [...][19]

18. Il s'agit de la hauteur du « type », le parallélépipède en plomb. Cette hauteur variait d'une fonderie à l'autre. Un caractère plus petit que les autres risquait fort de ne pas être imprimé tandis qu'un plus grand risquait de crever le papier !

19. La ligne valait 1/12 de pouce. Texte cité par A. Frey, *Nouveau manuel complet de typographie*, Manuels Roret, Paris 1857 ; édition fac-similé, Léonce Laget, 1979.

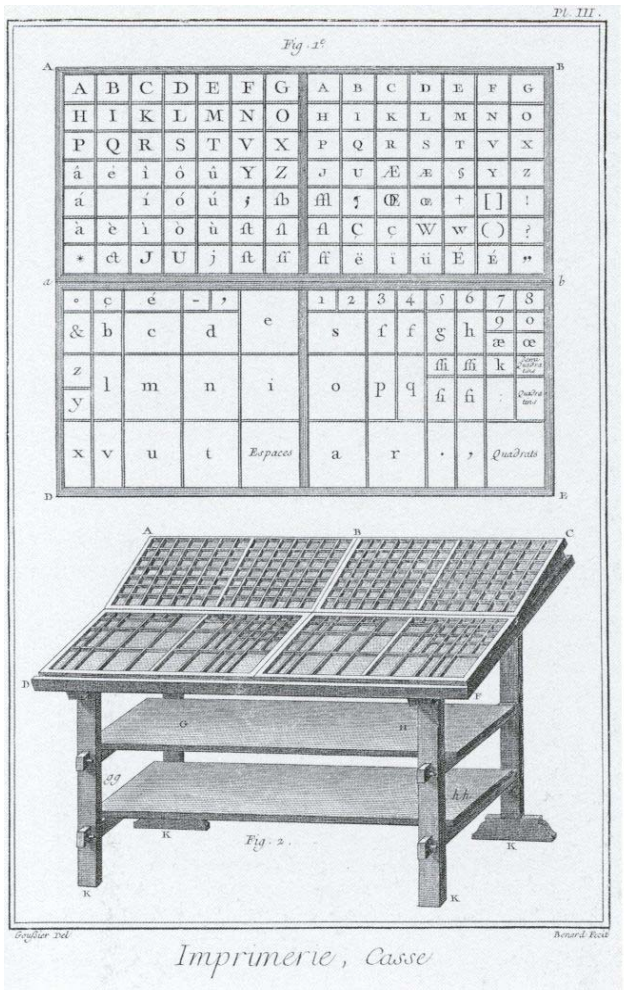


Figure 9. *La position des caractères dans une casse est normée*

Cette hauteur est toujours en vigueur avec quasiment cette même valeur et permet d'utiliser des caractères venant de n'importe quelle fonderie sur n'importe quelle presse. Toujours dans ce secteur, les caractères étaient rangés dans des « casses » où la répartition était toujours la même de façon que les typistes trouvent toujours, d'un atelier à l'autre, le même caractère au même endroit quelle que soit la fonte utilisée. Mais, cette disposition a pu varier d'une époque à l'autre (la figure 9 montre une casse française antérieure à celle dite parisienne) et surtout d'un pays à l'autre[20].

20. Tout comme aujourd'hui les claviers de machines à écrire et d'ordinateurs ont des configurations (par exemple AZERTY ou QWERTY) dépendant de la langue ou des pays (figure 12).

Depuis quelque temps, on utilise en français deux termes[21] :

1. Les **normes** sont des règles approuvées par des instances officielles ; elles offrent une certaine garantie de stabilité et de pérennité. Exemples : les normes ISO-8859 et ISO-10646.
2. Les **standards** sont définis par des groupes privés, en général industriels ou commerciaux (par exemple, IBM et son EBCDIC ou Adobe et son PostScript) mais aussi collégiaux (par exemple, les consortiums Unicode et W3C, voir 1.4.4.).

En anglais, le même mot *(standard)* est utilisé dans les deux cas, quoiqu'on qualifie souvent les seconds de *proprietary*.

1.4.2. *Organismes de normalisation et de standardisation*

Les normes sont définies soit par divers organismes nationaux (par exemple en France l'AFNOR[22], en Allemagne le DIN, etc.), soit par des organismes regroupant géographiquement divers pays (par exemple le CEN au niveau de l'Europe), soit des organismes internationaux (l'ISO, voir ci-dessous, mais aussi l'UIT Union Internationale des Télécommunications) regroupant au niveau mondial tous ces organismes. Par ailleurs ces organismes s'appuient sur des organismes sectoriels, par exemple l'ECMA (*European Computer Manufacturers Association*) qui a beaucoup travaillé sur le codage des caractères vers 1980. À ces derniers on peut ajouter des groupes comme l'IETF *(Internet Engineering Task Force)* qui produit des RFC *(Request For Comment)* dont MIME et le W3C *(World Wide Web Consortium)* et ses « recommandations » comme XML.

Tous ces organismes ont leurs propres règles de fonctionnement mais finalement suivent à peu près le même processus de définition d'une nouvelle norme que l'ISO.

1.4.3. *L'ISO et son fonctionnement*

L'ISO est l'Organisation mondiale de normalisation[23]. Créée en 1947, située à Genève, elle regroupe plus de 130 pays soit « membres » (la France, par exemple, y est représentée par l'AFNOR[24]), soit « correspondants » ou « abonnés » (comme

---

21. Sur la normalisation en général et sur celle des caractères et documents en particulier, voir [CHA99], [HER00] et [MAR90].

22. Les organismes xxx cités ici ont une page web dont l'url est en général http://www.xxx.org/. Voir toutefois certaines variantes (comme pour l'AFNOR et le DIN) dans la liste des sigles à la fin de ce numéro.

23. Contrairement à ce qu'on croit souvent, ISO n'est pas le sigle de *International Standards Organization*, mais le nom choisi, basé sur le grec *isos* (égal), pour désigner cet organisme qui a trois langues officielles (anglais, français et russe) et donc trois noms : l'Organisation internationale de normalisation, *International Organization for Standardization* et Международная организация по стандартизации.
Voir : http://www.iso.org/iso/fr/aboutiso/introduction/whatisISO.html

24. Contrairement à d'autres pays francophones, dont notamment le Canada, la France est assez peu active en matière de normalisation de caractères et ses rares représentants sont plus des ingénieurs que des spécialistes linguistes ou typographes ! On explique ceci par le fait

certains pays en voie de développement). Ce sont ces représentants qui votent l'approbation finale des normes, mais tout le travail est fait dans le cadre de comités techniques (TC) et de sous-comités (SC).

L'élaboration d'une norme prend de nombreuses années et suit le processus suivant (chaque phase se terminant par un vote) :

1. Expression du besoin (NP : *new proposal*) ;
2. Spécifications, avec publication d'un CD *(Comittee Draft)* et de DIS *(Draft Inter-national Standard)* ;
3. Approbation, publication IS *(International standard)*, traductions, etc.
4. Réexamen (tous les 5 ans).

Notons que les normes sont en général payantes (et souvent relativement chères), et ne sont donc pas toujours faciles à trouver (notamment sur le web !), contrairement aux standards. Il est ainsi facile de suivre le développement du standard Unicode et beaucoup moins celui de la norme ISO-10646 qui en est pourtant un sur-ensemble !

1.4.4. *Autres organismes de standardisation*

Les organismes de « standardisation » eux sont privés et il peut s'agir soit d'une entreprise comme Adobe, soit de groupes collégiaux d'utilisateurs ou d'entreprises. Citons ici deux importants organismes de standardisation :

– Le **consortium Unicode**, composé essentiellement de compagnies (telles que Apple, IBM, Microsoft, Sun, Xerox, etc.), est donc en charge du standard Unicode. On trouvera l'historique et le fonctionnement du consortium dans l'interview de Ken Whistler à la fin de ce numéro (voir aussi [AND03] et [Uni00]).

– Le **W3C**, *World Wide Web Consortium*, produit un certain nombre de recommandations à valeurs normatives pour le web (telles que les définitions de XML, XSL, SVG, etc.). Le W3C est formé d'instituts publics (comme l'INRIA, l'Université de Keio, le MIT) et privés (actuellement il y a près de 500 membres). Le fonctionnement du W3C est assez proche de celui de l'ISO[25].

2. Premiers codages de transmission de textes

Dès l'aube de l'humanité, les hommes ont appris à communiquer par signes, par la voix, puis par écrit[26]. Mais ces modes de transmission de l'information se heurtent à

---

que souvent normalisation rime avec bénévolat et que les Français ont quelque difficulté à parler anglais (langue de travail *de facto* à l'ISO), ce qui est une mauvaise excuse car c'est aussi le cas des Asiatiques, pourtant fort actifs.

25. Voir à ce sujet http://www.w3.org/Consortium/

26. On trouvera une histoire récente de l'écriture sous la direction de Anne-Marie Christin dans *Histoire de l'écriture – de l'idéogramme au multimédia*, Flammarion, 2001.

plusieurs problèmes dont les principaux sont liés à des problèmes de distance, de vitesse de transmission, mais aussi de sécurité (avec les deux sens de confidentialité et de fiabilité). Pas étonnant donc que ce soient les gouvernements ou les militaires qui aient le plus œuvré dans cette voie des transmissions ! Mais une troisième force, celle du commerce, née aux États-Unis au XIX^{e} siècle, a aussi joué un rôle non nul dans cette course à la transmission de l'information. Voici quelques jalons de cette histoire qui nous permettent de mieux cerner le concept d'échange de caractères !

2.1. *Les fanions et sémaphores de la marine*

Depuis très longtemps, les marins avaient (avant la TSF, mais c'est toujours en usage) l'habitude de communiquer à l'aide de fanions qui peuvent être vus de loin.

Le principe est le suivant :

1. À chaque lettre correspond un fanion[27] (par exemple à Q correspond un fanion tout jaune, à N un damier noir et blanc et à L un damier noir et jaune. Un fanion est en quelque sorte un glyphe !

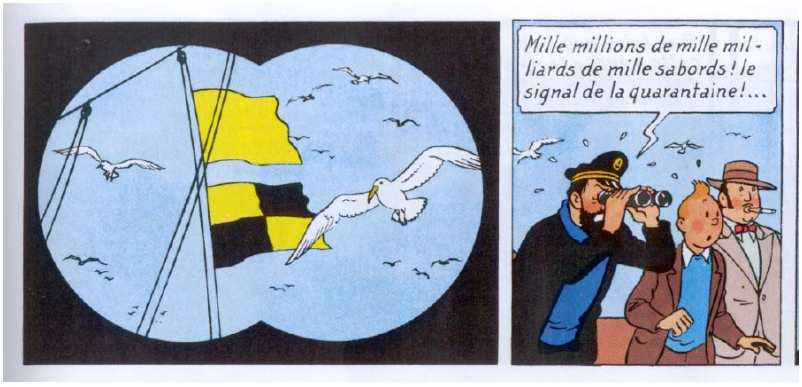


Figure 10. *Extrait du* Temple du soleil *de Hergé*

2. À chaque fanion, correspond un nom, aujourd'hui ceux de la radio : alpha pour A, bravo pour B, charlie pour C, etc. À chaque caractère d'Unicode va aussi être attaché un nom…
3. L'écriture se fait verticalement, du haut des haubans vers le bas (bel exemple d'écriture verticale dans un contexte occidental).
4. Les messages sont codés : ainsi N signifie « je suis en détresse » tandis que « QL » veut dire que le bateau est en quarantaine (figure 10).

27. Inventaire complet dans http://www.geocities.com/ecp_champlain/drapeaux.htm

Ce codage ne permet toutefois pas d'émettre rapidement des textes longs. Les marines de guerre ont alors conçu un autre alphabet visuel : un marin se met au pied d'une cheminée du bateau et donne à ses bras (prolongés de fanions pour être visibles) des positions correspondant à une lettre ou un chiffre. Ainsi (figure 11) le bras droit à l'horizontal signifie B et à la verticale D. Mais ce codage utilise deux nouveaux concepts :

1. La même position de bras (par exemple droit à l'horizontal) peut avoir deux sens : B ou 2 ; en effet, une position de bras signifie « ce qui suit est une série de chiffres » et une autre « ce qui suit est une série de lettres ». On va retrouver cette économie de codes avec les tableaux des machines à écrire, le code du Télex. Ceci correspond au concept d'extension, sur lequel nous reviendrons en 6.1.

2. Certaines positions ont aussi une autre signification. Ainsi celle qui sert à C en mode alphabétique ou à 3 en mode numérique signifie aussi « aperçu » en réponse à une position du sémaphoriste de l'autre bateau voulant dire « je vais émettre un message ». On va retrouver ces codes dans le Morse (p. ex. le code QRM) ou dans les codages comme Télex, Ascii, etc. (voir ci-dessous 4.2.1).

Figure 11. *Sémaphores de la marine*

2.2. Le télégraphe de Chappe

Utilisé de 1793 jusqu'en 1870, le télégraphe de Chappe utilisait un principe[28] un peu équivalent à celui des sémaphores de la marine :
1. En haut d'une tour, deux bras articulés (en 4 morceaux, figure 12) pouvant prendre une centaine de positions.
2. Chaque position est numérotée, par exemple ⌈=1 ⌊ =3 ⌉=9 ⌋=41.
3. Chaque lettre ou chiffre a un numéro, par exemple A=11.
4. Ici encore, on trouve des codes spéciaux (dits signaux réglementaires) pour les problèmes de transmission (par exemple « je suis en attente » ou « grande urgence »).
5. Un code secret (vocabulaire) est utilisé pour la confidentialité des messages[29] ; à un moment donné, les codes 63+18 peuvent signifier Paris et 89+16 troupes.

28. Voir plus de détails dans par exemple http://www.ec-lyon.fr/tourisme/Chappe/ et dans http://vvv.it.kth.se/docs/early_net/
29. Chappe profita de l'expérience en matière de code secret d'un sien oncle du Corps diplomatique.

Notons que ce codage ignorait les lettres accentuées. Ce système a tellement bien fonctionné que l'armée française a pris un gros retard dans l'utilisation de la télégraphie avec puis sans fil[30]…

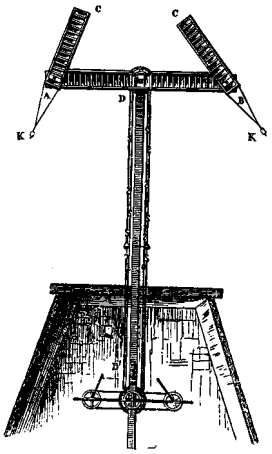


Figure 12. *Télégraphe de Chappe*

3. 1850-1950 : un siècle d'apports technologiques

Dans la seconde moitié du XIX^e^ siècle et au début du XX^e^ ont lieu diverses inventions qui vont profondément modifier la transmission de l'information.

3.1. *Le télégraphe et le Morse, premier alphabet international*

Le télégraphe électrique a été inventé dès le XVIII^e^ siècle[31] mais fonctionnait au départ avec 26 fils puis avec seulement 5. Vers 1840, aux USA, Samuel Morse réussit à n'utiliser qu'un seul fil grâce à un codage original. Une première version de ce codage a été utilisée dès 1844 (sous le nom *de American code*) puis ce codage a été stabilisé comme *International Morse Code* peu après et est devenu le premier alphabet international (ITA1 *International Telegraph Alphabet # 1*). Les principes du Morse sont les suivants :

1. Le codage est ternaire : trois états nommés bref, long et silence.

30. Un peu comme aujourd'hui où Internet n'est pas complètement adopté en France à cause du Minitel, son précurseur trop bien implanté !

31. Au départ, le premier prototype (construit en Écosse en 1753) utilisait 26 fils (un par lettre) et de l'électricité statique. À la fin du XVIII^e^ siècle on n'utilisait plus que 5 fils. Voir [BER81] (p. 61).

2. Un silence court sépare les signes et un long les mots.
3. Chaque état peut avoir des représentations (des glyphes !) variées : visuelles[32] (point, trait, espace ; un, deux ou aucun bras levé à l'horizontale ; un éclat lumineux bref, long ou le noir ; etc.), sonores (un son bref (ti), un son long (ta) ou un silence) voire électriques (d'où l'utilisation pour la télégraphie).
4. Divers caractères sont codés ; Morse propose (figure 13) :
– les chiffres ou lettres (sans distinction majuscule/minuscule), codés avec un nombre variable d'états ; par exemple 2 pour A « • ▬ » et 4 pour B « ▬ • • • » ;
– des caractères propres à certains pays qui ont pu être ajoutés (comme le CH espagnol ou le Å nordique), mais pourquoi ceux-là et seulement ceux-là ?
– des signes de ponctuation[33] et des commandes.
5. Certaines combinaisons de lettres servaient pour des fonctions spéciales (par exemple QRM signalait l'annonce de la transmission d'un message).

classe	signe	code	classe	signe	code
lettres	A	•–	spéciales	Ä	•–•–
	B	–•••		Á	•– –•–
	C	–•–•		Å	•– –•–
	…			CH	– – – –
chiffres	0	– – – – –		É	••–••
	1	•– – – –		Ñ	– –•– –
	2	••– – – –		Ö	– – –•
	…			Ü	••– –
ponctuations	point	•–•–•–	commande	Attente	•–•–
	virgule	– –••– –		Attaque	–•–•–•–
	apostrophe	•– – – –•		Répétez	•–••–•
	deux-points	– – –•••		Final	•–•–•–•
	…			Erreur	••••••••

Figure 13. *Extrait du code Morse [BER81, p. 102]*

3.2. *Machines à écrire*

Bien que divers prototypes aient existé avant, la première machine à écrire remonte à 1870 (Sholes et Remington). Dès cette époque, on trouve, au niveau du clavier (figure 14) des caractéristiques qui vont rester pratiquement des invariants jusqu'à nos ordinateurs.

32. Morse était professeur d'art graphique !
33. Notons l'absence de codage pour l'espace inter-mot.

– Les caractères sont alignés sur plusieurs lignes, une de chiffres (notons que les chiffres 0 et 1 sont absents : on tapait la lettre O et la lettre I), trois de lettres[34], les signes de ponctuation et autres étant sur le côté.

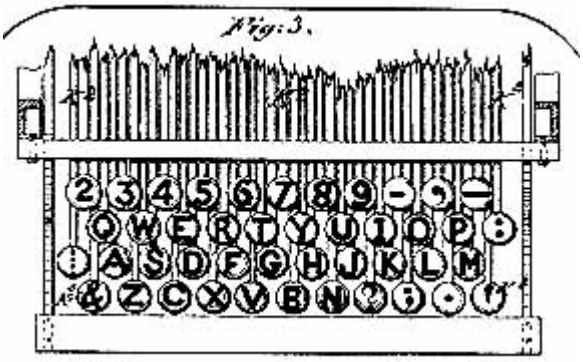


Figure 14. *Le clavier de la première machine à écrire*[35]

– Le second modèle, 1878, offrit des lettres minuscules avec le même principe qu'aujourd'hui : une touche commande le passage d'une série à l'autre (en fait le basculement de la corbeille portant les barres minuscules ou majuscules). On retrouvera ce principe dans le codage du télex !

– Il est intéressant de signaler la présence de l'esperluette « & », caractère d'origine latino-médiévale mais très en usage dans les noms d'entreprises américaines : la machine à écrire[36] a commencé par une vocation commerciale et il ne faut pas s'étonner que l'on retrouve de nombreux caractères commerciaux sur nos claviers d'ordinateurs et dans les codages de caractères !

Depuis cette époque, la machine à écrire a certes évolué sur le plan technique (vitesse de frappe, entraînement du papier par roues à rochets, boules, ordinateurs, etc.) mais relativement peu dans son ergonomie et le codage sous-jacent : à une touche (ou à une combinaison de touches) correspond un caractère ! Citons quelques points importants néanmoins :

34. En fait, le tout premier modèle proposait l'ordre alphabétique. Cet ordre QWERTY a été choisi pour optimiser certains voisinages tels que "we" et pour que les barres ne se coincent pas. Cet ordre a même fait l'objet d'un brevet (en 1878). Signalons que le clavier Dvorak (voir http://web.mit.edu/jcb/www/Dvorak/ ; il est surtout répandu en Amérique du Nord) propose une combinaison plus optimale des touches en matière de déplacement des doigts.

35. Extrait de http://home.earthlink.net/~dcrehr/whyqwert.html

36. Et en fait l'écriture humaine : rappelons que les premières traces d'écriture sont des relevés comptables (voir Jack Goody, *La raison graphique*, Les éditions de minuit, 1977).

– Invention de la touche *escape* (SUP ou MAJ) et des touches mortes permettant de composer un caractère à partir de plusieurs (la séquence « esc ^ a » par exemple permet d'obtenir « â » . Ce que permet encore Unicode !

– Plus grand nombre de caractères : ainsi dès la fin du XIX[e] siècle voit-on sur les claviers américains les caractères # @ (que l'on vient de découvrir en France il y a quelques années !) & $ etc.

– Adaptation du clavier à des langues non anglaises. Le clavier flamand est du type AEUZF, l'anglais QWERTY et le français du type AZERTY (figure 15).

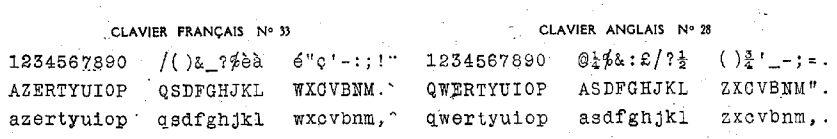


Figure 15. *Claviers de machines Corona (*Papyrus*, n°15, juin 1921)*

Divers autres modèles de claviers ont été proposés (outre l'américain Dvorak cité plus haut, mentionnons le français Neuville[37]) mais sans succès universel ! Bien sûr ces claviers ont été adaptés aux langues alphabétiques non latines (hébreux, arabe, cyrillique, etc.) mais pas vraiment aux idéogrammes des langues orientales !

3.3. Machines à composer

Dans ce même souci d'industrialisation et de gain de temps, sont apparues dans le de l'imprimerie[38] diverses inventions[39] qui ont aussi fortement marqué, de façon implicite, les problèmes de codage.

– La Linotype, inventée par Mergenthaler en 1886, permet de composer des lignes de caractères à partir d'un clavier : le linotypiste tape le texte, des matrices portant l'empreinte des caractères sont automatiquement assemblées, un système d'espaces en coins biseautés permet de justifier les lignes, le plomb est alors coulé à chaud et une « ligne bloc » est ainsi obtenue, qu'il suffit ensuite de mettre dans une galée. Ces machines ont servi dans la presse (où rapidité prime sur qualité) jusque vers 1980.

– La Monotype (Lanston, 1899) repose sur le même principe de composition à chaud, mais distingue très nettement deux tâches : celles de la saisie et celle du moulage et s'adresse beaucoup plus aux travaux de labeur.

37. ISO 9995. Voir http://std.dkuug.dk/JTC1/wg5/32. Mais il ne semble pas avoir pris.
38. On classe souvent [DRE77] les travaux d'imprimerie en trois : la Presse, le Labeur (livres, catalogues, formulaires administratifs, etc.) et Travaux de ville (plaquettes, factures, cartes de visite, etc.).
39. On trouvera dans [DRE77], [PHI87] et [MAR91] de nombreux détails sur cette histoire technique de l'imprimerie au sens large.

Mais au-delà des problèmes techniques, quelques concepts émergent de ces inventions.

– La saisie se fait par clavier. Même si, contrairement à ceux des machines à écrire, le nombre de touches est beaucoup plus grand (environ 300 pour un clavier de Monotype, sans tenir compte des positions spéciales, contre une cinquantaine pour une machine à écrire), le nombre de caractères est limité contrairement aux casses d'autrefois où il était toujours possible d'ajouter des casseaux[40].

– Même si la banalisation des usages américains a pu être, entre autres, une raison expliquant la perte des accents sur les capitales, voire sur les bas de casse, il faut reconnaître que la société Monotype (notamment avec les photocomposeuses) a tout fait pour s'adapter au marché européen (voire mondial) en offrant de très grands jeux de caractères, notamment pour la gestion des diacritiques pour 21 langues européennes[41].

– La communication entre les deux organes, clavier et fondeuse, de la Monotype se faisait à l'aide de rubans perforés à … 31 canaux ! En fait on y codait déjà un mélange d'ordres de composition et de données (le texte à composer). Mais ces rubans, pas plus que ceux du télégraphe automatique, ne semblent pas avoir été conservés à des fins de stockage. C'est qu'en effet ce mélange de commandes de composition, de transmission et de texte pur est très difficile à utiliser sur un autre matériel ou avec une autre mise en page que ceux spécifiquement prévus initialement.

– Ce que proposent les matrices des Linotype, Monotype, etc. ce sont des caractères au même sens qu'en composition manuelle. On y trouve par exemples des ligatures comme « fi ». Mais aussi toutes les déclinaisons graphiques d'un caractère (petites capitales, gras, etc.). C'est au typiste de choisir d'utiliser la ligature « fi » là où l'auteur n'a sans doute fait qu'écrire «fi », traduisant ainsi en terme de rendu (graphique) le concept de texte abstrait du manuscrit. De même pour tous les attributs typographiques !

3.4. Rubans et cartes perforés

L'invention de Morse a eu le succès que l'on sait. Mais assez vite, on s'est rendu compte que le télégraphe serait mieux utilisé si la transmission se faisait en deux phases : 1) saisie du texte codé et 2) envoi automatique de ce message. C'est ainsi qu'utilisant la technique des rubans perforés de Jacquard connus depuis le début du

40. Petite boîte contenant des caractères spécifiques à un travail. Ça existait aussi pour les Linotypes, mais nécessitait beaucoup d'opérations manuelles !

41. *The Monotype Recorder*, Vol. 42, n° 4, *Languages of the World* (cité par [PHIL68, chapitre 19 : *Character Sets*]). Le typographe Stanley Morrison a beaucoup œuvré pour la production de fontes de très haute qualité dont le dessin est aujourd'hui à la base de beaucoup de fontes vectorielles du catalogue d'Adobe par exemple.

XIX[e] siècle et celle des machines à écrire, apparaît le *Wheatstone Automatic Telegraph* où les textes à télégraphier sont d'abord saisis sur un ruban perforé à deux canaux (tiret et point). Au début du siècle, le Français Baudot modifia ce code en 5 canaux ce qui devint l'ancêtre du système TTS et du télex (voir figure 16 et ci-après 3.5 et 3.7).

Si, au XIX[e] siècle, les rubans perforés sont de plus en plus employés pour **transmettre des messages** ou commander des machines, c'est quand même Hollerith qui inventa la carte perforée comme support pour **le stockage et le traitement** de l'information (de données). Ses premières cartes utilisées pour le recensement des États-Unis ne comprenaient que des chiffres répartis sur 45 colonnes. En 1931, la société fondée par Hollerith, après divers regroupements, devint IBM *(International Business Machines)* et définit le standard de carte perforée à 80 colonnes (figure 17).

5	●					●		●		●	●	●					●	●	●
4	●	●	●	●	●			●		●	●			●			●	●	●
3				●	●		●		●		●		●			●	●		●
F	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
2		●	●		●					●			●	●			●		
1					●					●	●		●		●		●		

	A	R	R	I	V	E		A			1	3	.	4	5			A	U

Figure 16. *Un ruban de télex et sa transcription*

Chaque carte comprend donc 80 colonnes. Dans chacune on peut faire 1, 2 ou 3 trous dans les lignes appelées 12 ou 11 (dans la partie supérieure) et 0 à 9. Un trou unique dans la ligne 0 d'une colonne signifie 0 (col. 61 de la figure 14), un double trou en lignes 12 et 1 indique un A (col. 2). Des codes à trois trous permettaient de coder parenthèses, ponctuation, etc. Notons l'absence de caractères accentués, même dans les codages de la compagnie Bull qui, d'origine norvégienne, s'est installée à Paris en 1931 et a été le principal concurrent d'IBM pendant des années[42].

42. Voir au sujet des cartes perforées : Robert Ligonnière, *Préhistoire et histoire des ordinateurs*, Laffont 1987 ; et Lars Heide, The role of patents and standards in shaping the punched card systems of the Bull Company from 1918 to 1952, *Actes du 5[e] colloque Histoire de l'informatique*, Cepadues éditions, 1998, p. 167-180.

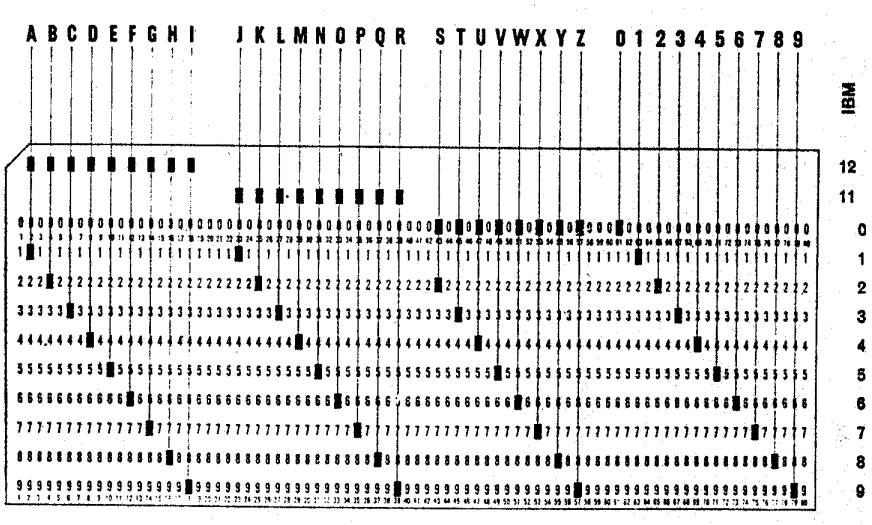


Figure 17. *Carte IBM 80 colonnes*

3.5. *Le télex*

Le réseau mondial du télex date des années 1920 et permettait de connecter entre elles des stations formées d'un clavier, d'un lecteur/perforateur de ruban, d'une imprimante[43] et bien sûr d'un dispositif de transmission. Les rubans (et la transmission) étaient basés sur un système[44] binaire à 5 moments qui est devenu *l'International (Telegraph) Alphabet 2*.

Un ruban de télex était formé de 6 pistes dont une, appelée F (figure 16), perforée tout du long (pour l'avancement du ruban sur des picots). Les cinq autres positions[45] (numérotées 1 à 5) peuvent être perforées ou non permettant d'avoir 32 codes. Mais, comme pour les machines à écrire, deux codes (en l'occurrence ceux numérotés 29 et 30) permettent de dire quelque chose comme « ce qui suit correspond à la colonne table haute » ou « ce qui suit correspond à la colonne table basse » (voir ci-dessous la section 6.1). Donc finalement on peut coder ainsi 2×(32–6)+6=58 caractères.

43. De mauvaise qualité (système à aiguilles) mais munie d'un ruban encreur de couleur permettant de garantir l'authenticité d'un texte, 'une des raisons du succès du télex.
44. Il s'agit du code Baudot redéfini par Murray ; aux USA on parle de code Baudot et en Grande Bretagne de code Murray ! http://www.wikipedia.org/wiki/Baudot_code
45. Ces positions sont parfois appelées canaux ou moments. Voir section 3.7.

L'alphabet comprend relativement peu de signes : les 26 lettres de l'alphabet latin[46] et quelques signes de ponctuation (il n'y a même pas le point virgule). En revanche, on voit déjà apparaître des signes de composition (p. ex. les code 28, *line feed*, interligne, et 27, retour chariot, qui permettent de mettre en page les lignes) mais aussi le code 10 (appel) qui met en marche une sonnerie chez le correspondant pour qu'il sache qu'il va recevoir un message et soit prêt à y répondre. Notons que le télex est encore en usage aujourd'hui !

N°	Table haute	Table basse	Code	N°	Table	Table	Code
1	A	-	11000	17	O	1	11101
2	B	?	10011	18	R	4	01010
3	C	:	01110	19	S	'	10100
4	D	*	10010	20	T	5	00001
5	E	3	10000	21	U	7	11100
6	F	(réservé)	10110	22	V	=	01111
7	G	(réservé)	01011	23	W	2	11001
8	H	(réservé)	00101	24	X	/	10111
9	I	8	01100	25	Y	6	10101
10	J	Appel	11010	26	Z	+	10001
11	K	(	11110	27	← Retour		00010
12	L	)	01001	28	≡ Interligne		01000
13	M	.	00111	29	↑ Inversion		11111
14	N	,	00110	30	↓ Inversion		11011
15	O	9	00011	31	Espace		00100
16	P	0	01101	32	(réservé)		00000

Figure 18. *Alphabet international IA2 (télex)*

3.6. *Le système TTS*

Combinant à la fois le télégraphe et le télex, avec des rubans perforés à 6 canaux, ce système a été très conçu vers 1930 mais pratiquement utilisé que depuis 1950, notamment entre divers sites d'un même organe de presse ou d'une imprimerie.

46. Mais pas de minuscules ni, bien sûr, de lettres accentuées ! Notons que le caractères « espace » a un code, 4, qui n'est pas, contrairement au Morse, un caractère sans perforation ! On retrouvera dans Ascii, Latin-1 et Unicode cette même façon de coder l'absence de caractère ou plutôt la présence d'un blanc qui occupe de la place (Unicode allant donc plus loin en proposant des « espaces sans chasse »).

canaux	binaire	Base 10	Pos. Basse	Pos. haute
.. ...	00000	0	(réservé)	
.. ..•	00001	1	T	5
.. .•.	00010	2	Retour chariot	
.. .••	00011	3	O	9
.. •..	00100	4	Espace	
.. •.•	00101	5	H	(réservé)
.. ••.	00110	6	N	.
.. •••	00111	7	M	,
.• ...	01000	8	Interligne	
.• ..•	01001	9	L	)

Figure 19. *Relation canaux/codes binaires et décimaux pour le télex*

Nom du codage	nombre de bits ou moments	nombre théorique de caractères
Télex	5	32
BDC	6	64
Ascii	7	128
ISO- Latin 1	8	256
Unicode versions 1 et 2	16	65536
Unicode versions > 3.0	20	~ 1 million
ISO/ CEI 10646	32	> 2 milliards

Figure 20. *Relation nombre de moments / nombre de caractères de codages*

3.7. *Moment, canaux, bits…*

Faisons à présent le point sur ce concept que nous avons utilisé plusieurs fois sous le nom de moment, canal, taille de codage, etc. Reprenons le code du télex (figure 18) et classons le cette fois non par ordre alphabétique mais selon les codes (figure 19).

En colonne de gauche, on indique les canaux du ruban perforé (voir figure 16). La colonne suivante représente la même chose, mais avec des 0 et 1 à la place des points pour l'absence de perforation et des boulets pour des perforations. Ceci correspond à des nombres en base 2, dont l'équivalent est donné en base 10 colonne suivante.

Le nombre de canaux (ou pistes) s'appelle aussi nombre de *moments*[47]. Les chiffres 0 et 1 s'appellent des *bits*. On peut donc dire que le télex est un codage à 5

47. Ce mot vient du latin *momentum* mais avec le sens de pression d'un poids. C'est ce même mot que l'on retrouve dans des expressions de la physique comme moment d'un vecteur, moment d'inertie ou moment magnétique (d'où son emploi ici).

moments ou à 5 bits. Chacune des 5 positions binaires pouvant prendre la valeur 0 ou 1, on peut avoir 2^5, soit $2 \times 2 \times 2 \times 2 \times 2$, donc 32 caractères différents. De façon plus générale, si un codage est fait avec n positions binaires (ou moments, ou canaux !), alors ce codage permet de traiter 2^n caractères.

C'est notamment en jouant sur le nombre de ces canaux (ou sur le nombre de bits) que les diverses normes de codage des caractères ont pu augmenter leurs répertoires.

4. Codages à 7 bits : Ascii

Jusqu'à ce jour, la seule norme de codage universellement utilisée aura été l'Ascii. Ce codage a vu le jour aux USA vers 1965 et a fourni pendant plus de trois décennies le seul codage non ambigu à 7-bits. Ses contenu et nom (ISO 646) actuels datent de 1983.

4.1. *Avant l'Ascii*

Dans les années 1940 apparaissent les ordinateurs qui remplacent les machines comptables mais en gardent notamment les cartes perforées et leurs codes. Mais, très vite, la plus grande confusion règne en matière de codage des informations de façon interne (ce qui n'était pas très gênant) mais aussi externe, rendant difficile toute communication d'un ordinateur à l'autre. Ceci était même vrai pour les diverses machines d'un constructeur. Le géant de l'époque, IBM, est ainsi amené à définir peu après 1955 le codage BCD.

– Le codage BCD *(Binary Coded Decimal)* est un code à 6 bits (donc 64 caractères), basé sur celui des cartes perforées. Il ne comprenait que les lettres majuscules, les chiffres et ponctuations et quelques commandes comme CR, *carriage return* (figure 21). Il a donc été abondamment utilisé sur les premiers ordinateurs d'IBM et a servi de base à son successeur EBCDIC (voir ci-dessous section 5.2.1) et aux normes ISO/R646-1967 et Afnor NF Z62-010.

	0	1	2	3	4	5	6	7
00	SP	(	0	8	NUL	H	P	X
10	HT	)	1	9	A	I	Q	Y
20	LF	*	2	:	B	J	R	Z
30	VT	+	3	;	C	K	S	[
40	FF	,	4	$	D	L	T	
50	CR	-	5	=	E	M	U	]
60	SO	.	6	&	F	N	V	ESC
70	SI	/	7	'	G	O	W	DEL

Figure 21. *Le codage BCD d'IBM*

– Un autre constructeur américain, Control Data Corporation, a défini un codage à 6 bits en remplaça les caractères de commande par d'autres symboles (tels que < et >) ; il est resté en usage très longtemps au Canada sous le nom d'Ascii banbang.

Mais IBM n'était pas le seul constructeur d'ordinateurs et vers 1960 on notait encore une très grande variétés de codages (figure 22). IBM et les autres constructeurs (dont Univac, Burrough, Honeywell, etc.) se sont regroupés pour définir un code commun vers 1960. Il a d'abord été adopté par l'ISO et le CCITT en 1963 sous le nom de IA5 *(International Alphabet # 5)* puis a été publié en 1967 par l'organisme de normalisation américain American Standard Association sous le nom d'Ascii *(American Standard Code for Information Interchange).*

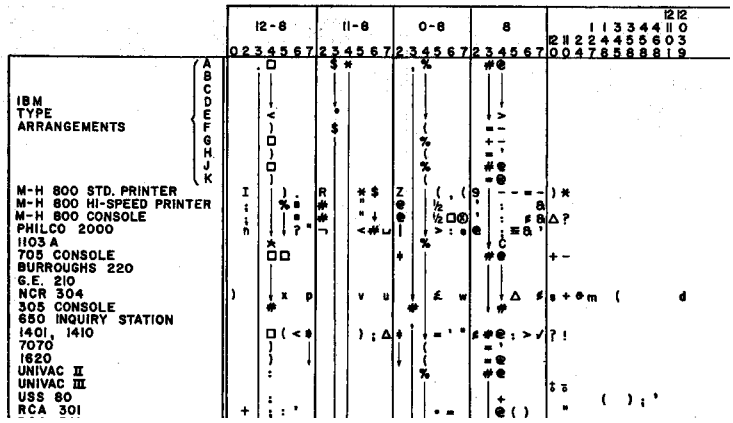


Figure 22. *Étude des codages de caractères existants en 1960* (Communications of the Association of Computer Machinery*)*

4.2. *Principes de l'Ascii*

Le principe de l'Ascii est une structure à 7 moments (7 bits) permettant donc le codage de 2^7 soit 128 caractères (figure 23). Ce codage ne comprend en fait que 95 caractères « imprimables[48] » codés en positions 33_{10} à 126_{10}, les autres étant des codes de commande.

4.2.1. *Caractères de commande*

L'Ascii contient donc 33 caractères « de commande » : les 32 premiers (numérotés[49] de 0 à $001F_{16}$) et le dernier ($007F_{16}$). Ces caractères étaient en fait des

48. Unicode dirait « ayant un glyphe ». Notons que l'espace, de code 33_{10} est généralement classé dans les commandes. Nous préférons le ranger ici parmi les caractères imprimables, même si son glyphe est blanc !

49. Pour unifier les notations de ce codage et des suivants, nous donnons désormais les numéros de code en hexadécimal.

caractères de commande pour périphériques tels que écrans, perforateurs de ruban ou Télétypes. Il y a par exemple :

CR (Carriage Return) pour « retour chariot » et *LF (line feed)* pour fin de ligne[50] ;

– *Bell* (sonnerie) pour activer la sonnerie d'un télex, etc. ;
– *DEL (delete)* dont le code $7F_{16}$ (soit 1111111 en binaire) permettait de supprimer un code erroné en trouant toutes ses positions dans un ruban perforé ;
– *ESC (escape)* et *BS (backspace)* permettant d'écrire la séquence E ESC ' BS pour commander l'affichage sur écran d'un É, mais ceci ne valait que pour quelques accents, ne permettait aucun tri alphabétique et aucune utilisation en dehors des écrans et n'a donc été que très peu utilisé ;
– *SO (shift out)* et *SI (shift in)* qui, combinés avec ESC, permettaient de faire des extensions (voir ci-après section 6.1) ; notons qu'ils existaient déjà dans BCD.

Ces caractères étaient donc très liés à une technologie aujourd'hui périmée (rubans notamment) et ne sont pratiquement plus utilisés de façon normale. Les cases correspondantes étant donc inutiles, nombre de standards propriétaires les ont utilisées pour y mettre d'autres caractères (voir section 6.4).

	000	001	002	003	004	005	006	007
0	NUL	DLE	SP	0	@	P	`	p
1	STX	DC1	!	1	A	Q	a	q
2	SOT	DC2	"	2	B	R	b	r
3	ETX	DC3	#	3	C	S	c	s
4	EOT	DC4	$	4	D	T	d	t
5	ENQ	NAK	%	5	E	U	e	u
6	ACK	SYN	&	6	F	V	f	v
7	BEL	ETB	'	7	G	W	g	w
8	BS	CAN	(	8	H	X	h	x
9	HT	EM	)	9	I	Y	i	y
A	LF	SUB	*	:	J	Z	j	z
B	VT	ESC	+	;	K	[	k	{
C	FF	FS	,	<	L	\	l	|
D	CR	GS	-	=	M	]	m	}
E	SO	RS	.	>	N	^	n	~
F	SI	US	/	?	O	_	o	DEL

Figure 23. *Le codage Ascii dans sa version finale de 1983 (ISO 646). Les codes sont donnés en hexadécimal (par exemple R a pour code 0051_{16})*

50. Ces deux commandes étaient complémentaires (LF remplissait une ligne en cours et CR passait à une autre ligne) mais les technologies ont évolué et une certaine confusion existe depuis, ce qui explique certains comportements variants d'un logiciel à l'autre.

4.2.2. Les caractères graphiques d'Ascii

Les 95 caractères sont les caractères dits « graphiques » car on peut les afficher sur un écran ou les imprimer. Ces 95 caractères sont eux-mêmes répartis en 3 groupes (voir figure 23) :

1. 83 caractères obligatoires :
   - l'espace,
   - 52 lettres : A-Z et a-z,
   - 10 chiffres : 0-9,
   - 20 signes de ponctuation ou autres : ! " % & '( ) * + , - . / : ; < = > ? _
2. Deux caractères « au choix »[51] # ou £ et $ ou ¤ (symbole monétaire international dont le glyphe représente une pièce d'or où brillent quatre rayons de soleil).
3. Dix positions réservées à des caractères d'usage national.

La norme Ascii comprenait donc à l'origine :

- des variantes nationales (parfois plusieurs pour un même pays – c'est le cas de la France) ; voir figure 24 ;
- une version internationale de référence, IRV, où les positions optionnelles sont affectées d'un caractère précis.

Version de référence (IRV)	#	¤	@	[	\	]	^	`	{	|	}	~
Allemagne (DIN66003)	#	$	§	Ä	Ö	Ü	^	`	ä	ö	ü	ß
Belgique	#	$	à	°	ç	§	^	`	é	ij	è	~
Espagne	#	$	¨	¡	Ñ	Ç	¿	`	'	ñ	ç	"
France (NF Z62010/1982)	£	$	à	°	ç	§	^	µ	é	ù	è	"
Grande Bretagne	£	$	@	[	\	]	^	`	{	|	}	~
Suisse romande			à		ç				é	ù	è	~
USA (norme US-Ascii)	#	$	@	[	\	]	^	`	{	|	}	~

Figure 24. *Caractères optionnels de la version de référence IRV de l'Ascii et quelques variantes nationales.*

4.3. *D'Ascii à ISO 646*

Plusieurs raisons ont amené à préciser ce codage.

51. Ce choix a été demandé par divers états, dont l'URSS et la Grande Bretagne, lors de la Guerre froide, afin de contrer l'hégémonie du dollar ! On verra plus bas que, lors de la Péristroïka, le dollar a repris sa place aux dépens du symbole monétaire international qui va réapparaître dans Latin-1 puis y être remplacé par le symbole euro en Latin-9.

– La version internationale de référence n'était pas la version américaine Ascii (d'ailleurs appelée, à l'époque, US-Ascii) : IRV contenait le symbole ¤ tandis que US-Ascii utilisait le dollar. On ne trouve donc le symbole ¤ sur pratiquement aucun matériel informatique (sauf au Canada et sur certains claviers très fidèles à Latin-1, comme ceux des SUN). En revanche, les caractères @ et #, utilisés en comptabilité américaine et alors complètement inconnus en France (bien que d'origine latine), sont systématiquement sur tous nos claviers d'ordinateurs. Les informaticiens américains se sont mis à utiliser nombre des caractères optionnels (#, @, les accolades, etc.) dans leurs programmes ce qui a donné un poids anormalement fort à la version américaine Ascii.
– Il y avait une grande incohérence d'un pays francophone à l'autre à tel point d'ailleurs que la France a abandonné sa norme Z62010 au profit de l'Ascii en 1983.

C'est pourquoi, en 1988, la norme ISO 646 a pris exactement le codage Ascii de la figure 23. Signalons que bien que diffusée par l'AFNOR, cette norme n'a pas été traduite en français ! Cette norme a aussi été publiée par le CCITT comme Recommandation V.3 et mise à jour en 1984 sous le nom de T.50 [MAR90].

4.4. *Pérennité d'Ascii*

Depuis, d'autres normes ont été définies mais, compatibilité oblige, cette norme sert de base à toutes les autres normes et en particulier à ISO-Latin-1 et par là à Unicode. Par ailleurs, comme elle est suffisante pour la majorité des Américains, cette norme Ascii reste très importante ! En revanche, ça a donné la mauvaise habitude d'envoyer des *mails* français sans lettres accentuées comme si on ne pouvait pas le faire !

Mais la grande force d'Ascii aura été d'avoir servi de base à des codages plus complexes pour coder notamment tous les codages à 8 bits comme ISO-Latin-1. Le principe est simplement de coder un caractère avec une succession de caractères Ascii, le premier en général étant alors un caractère réservé qui doit donc lui aussi être codé. Ainsi peut-on écrire du HTML avec le seul Ascii : le caractère & doit alors être remplacé par la séquence & ce qui permet donc d'utiliser ce symbole comme début du codage, plutôt du nommage, d'autres caractères. La figure 25 en donne quelques exemples pour des applications comme MIME[52]. Ceci concerne aussi bien des lettres accentuées, des signes de Latin-1. TeX va plus loin et offre, outre des caractères aussi accessibles par HTML (comme δ) de nombreux glyphes, y compris typographiques (comme le tiret cadratin — qui n'existe de façon normative que dans Unicode).

52. MIME *Quoted printable* permet d'utiliser dans les courriers électroniques supposés travailler en 7 bits tous les caractères accentués français codés en 8 bits. Voir http://www.ietf.org. On peut dire que UTF-8 permet de même de passer en 8 bits tous les codes 16 (voire 20) bits d'Unicode.

Caractère ⇒	é	Å	δ	—	±
MIME	=E9	=C5			=B1
HTML	é	Å	δ		±
TeX	\'e	\AA	δ	---	$\pm$

Figure 25. *Utilisation d'Ascii pour nommer des caractères de codages de plus de 8 bits*

4.5. *Remarques sur Ascii/ISO 646*

– Il s'agit donc d'une norme d'échange. Mais, par abus de langage et sans doute par méconnaissance des principes de codage, certains informaticiens ont tendance à utiliser le mot Ascii avec le sens de « non formaté », voire de « texte source ». Le codage Unicode a, depuis, introduit les concepts de texte enrichi *(fancy text)* et de texte brut *(plain text)* (voir section 1.3).
– Ascii est un codage à 7 bits. L'expression Ascii-8bits est un abus de langage à bannir d'autant plus que selon les uns il signifie EBCDIC, selon d'autres ISO-Latin1, voire des codages propriétaires comme ceux d'Apple ou de Windows dont il existe un très grand nombre de variantes (voir RFC1345).
– Ascii a eu d'abord comme vocation l'échange de programmes informatiques et de données techniques ou comptables. C'est ce qui explique le fait qu'il n'y ait pas de glyphes typographiques (par exemple le blanc insécable ou le symbole paragraphe qui eux font partie de ISO-8859) ou que les caractères soient polyvalents voire ambigus (par exemple le signe « - » dont on ne sait s'il s'agit du signe moins « – », du trait d'union ou du signe division en typographie « - » ou du tiret typographique « — »).
– Bien qu'Unicode dise être compatible à 100% avec Ascii, ceci n'est pas vrai : le caractère de code 0027_{16} « ’ » que l'on retrouve en Latin-1 sous le nom de APOSTROPHE existe bien dans Unicode (même code, même nom), mais avec la mention « le caractère recommandé pour l'apostrophe est 2019 ’ ». Ce qui veut dire qu'un texte écrit en Ascii doit être recodé pour être utilisable en Unicode[53].

Nous nous sommes un peu étendu sur ce codage car il est encore très employé mais aussi pour montrer que même sur un codage aussi simple, il peut y avoir beaucoup d'interprétations, de variances d'un constructeur à l'autre, d'un utilisateur à l'autre, etc. Toutes choses que l'on va naturellement retrouver avec Unicode !

53. C'est l'emploi de ce nouveau code qui explique que l'on voit des « ? » dans certains textes (notamment dans le courrier électronique) à la place d'apostrophes ou au contraire des apostrophes remplacées par « ' ».

5. Codages à 8 bits : ISO 8859

L'anglais étant pratiquement la seule langue utilisable avec l'Ascii, de nombreux organismes ont bien sûr tenté de définir des normes plus riches. La plus célèbre est la norme, ou plutôt la série de normes ISO 8859, dont l'une est plus connue sous le nom de Latin 1.

	000	001	002	003	004	005	006	007	008	009	00A	00B	00C	00D	00E	00F
0	NUL	DLE	PAD		sp	&	-						é	è	ç	0
1	SOH	DC1	HOP	PU1			\		a	j	#		A	J		1
2	STX	DC2	EPH	SYN					b	k	s		B	K	S	2
3	ETX	DC3	NBH	STS					c	l	t		C	L	T	3
4	ST	OSC	IND	CCH					d	m	u		D	M	U	4
5	HT	NEL	LF	MW					e	n	v		E	N	V	5
6	SSA	BS	ETB	SPA					f	o	w		F	O	W	6
7	DEL	ESA	ESC	EOT					g	p	x		G	P	X	7
8	EPA	CAN	HTS	SOS					h	q	y		H	Q	Y	8
9	RI	EM	HTJ	SGC				`	i	r	z		I	R	Z	9
A	SS2	PU2	VTS	SCI	°	§	ù	:								
B	VT	SS3	PDL	CSI	.	$	,	£								
C	FF	IS4	PLU	DC4	<	*	%	à								
D	CR	IS3	ENQ	NAK	(	)	_	'								
E	SO	IS2	ACK	PM	+	;	>	=								
F	SI	IS1	BEL	SUB	!	^	?	"								

Figure 26. *Le codage EBCDIC d'IBM (version FR)*

5.1. *Premières tentatives officielles*

Les organismes de normalisation ont très tôt essayé de coder d'autres caractères que ceux Ascii :
– 1973 : norme ISO 2022 reprenant et étendant le concept des caractères d'extension d'Ascii (ESC, SO et SI). Ce sera le point de départ des normes permettant de coder le japonais par exemple.
– Dès 1978, études « officielles » pour un codage des langues latines sur 8 bits.
– En 1978, norme UKPO pour le *VIEWDATA character set* pour symboles et lettres accentuées (pour le videotex).
– En 1979 norme ISO 4873 « jeux de caractères codés à 8 éléments pour l'échange d'information ».

5.2. *Standards des constructeurs*

En parallèle, les constructeurs définissent de leur côté des standards pour leurs propres besoins. Nous en retiendrons deux ici.

5.2.1. *EBCDIC d'IBM*

EBCDIC, abréviation de *Extended Binary-Coded Decimal Interchange Code*, est le codage propriétaire qu'IBM a substitué à BCD en 1964 pour ses ordinateurs de la série 360. EBCDIC est défini sur 8 bits mais ne propose en gros que l'équivalent de l'Ascii. Toutefois, il existe des variantes pour de nombreuses langues (57 jeux nationaux). La figure 26 montre la présence de lettres accentuées françaises, le grand nombre de caractères de commande et … la place perdue !

Ce codage a été suffisamment important pour qu'Unicode définisse officiellement une transformation d'EBCDIC vers Unicode (UTF-EBCDIC).

5.2.2. *VT200 de DEC*

De son côté le constructeur américain DEC (qui a tenu une bonne part du marché mondial grâce à ses machines orientées réseau et à son Vax) a défini un codage 8 bits pour ses terminaux VT-200 : *Multinational Character Set*[54] qui servira de base pour Latin 1. Notons que les œ et Œ étaient dans le codage de VT200 et ont été remplacés par × et ÷ dans Latin-1…

5.3. *Les normes ISO 8859*

Se basant sur ces « standards *de facto* » la norme la plus importante pour les langues européennes a été définie par l'ISO et est connue sous le nom d'ISO/CEI 8859-n (avec, actuellement, n de 1 à 16) qui est une extension à 8 bits de l'Ascii. Le seul fait de passer de 7 à 8 bits permettait de doubler le nombre de caractères, donc de passer à 256 caractères (moins les fameux caractères de commande !). Comme les langues en usage en Europe utilisent plus de 256 caractères différents, il a été décidé de regrouper ceux-ci par affinités ... commerciales. C'est ainsi qu'il y a ISO Latin-1 pour la zone occidentale, Latin-2 pour la zone orientale, etc. Pour des raisons politico-économiques, un codage spécial (Latin-5) a dû être ajouté pour la Turquie et ses partenaires ! Par ailleurs, depuis quelques années, de nouveaux codages sont proposés pour satisfaire la qualité « linguistique » de certains alphabets : c'est ainsi que le codage Latin-9 corrige les manques de Latin-1 pour le français (où « Œ », « œ » et « Ÿ » étaient absents) et que Latin-8 permet d'écrire l'ancienne orthographe du gaélique irlandais.

Toutes ces normes 8859-n ont trois parties :

1. les 128 premières positions sont rigoureusement identiques à l'Ascii,
2. les 32 autres sont de nouvelles commandes, mais identiques dans tous les codages 8859-n,
3. les 96 dernières positions sont spécifiques au codage 8859-n.

54. http://czyborra.com/charsets/iso8859.html.

On trouvera l'ensemble de ces codages et les langues traitées correspondantes dans l'article de Sylvie Baste dans ce numéro[55]. Ici, on voit les codages liés au français.

5.3.1. *ISO-Latin1*

Les caractères spécifiques au codage Latin-1 sont montrés figure 27. On y remarque que les « ligatures » Œ et œ ainsi que la capitale Ÿ en sont absentes (voir à ce sujet [AND96]). Mais à part cette erreur, Latin-1 permet de coder tous les caractères français et d'Europe occidentale ; c'est pourquoi elle a été adoptée par de très nombreux produits (ou d'autres normes comme HTML). Le répertoire[56] de Latin-1 comprend donc (les possesseurs d'ordinateurs n'en sont pas toujours conscients !) :

- tous les caractères d'Ascii,
- des capitales et minuscules avec signes diacritiques (comme À Ç É Ï Û, à ç é ï û, mais aussi Í ou ò, etc.),
- des lettres propres à certaines langues (Ø Đ Ñ å ß, etc.),
- des symboles monétaires (¢ £ ¥ ¤)
- des symboles scientifiques (± ÷ × µ, etc.)
- des signes typographiques (espace insécable, ° § ª, etc.)
- des signes spécifiques à certaines langues(« » ¿ ¡, etc.),
- divers symboles (© ® ¦, etc.).

	008	009	00A	00B	00C	00D	00E	00F
0	XXX	DCS	NBSP	°	À	Đ	à	ð
1	XXX	PU1	¡	±	Á	Ñ	á	ñ
2	BPH	PU2	¢	²	Â	Ò	â	ò
3	NBH	STS	£	³	Ã	Ó	ã	ó
4	IND	CCH	¤	´	Ä	Ô	ä	ô
5	NEL	MW	¥	µ	Å	Õ	å	õ
6	SSA	SPA	¦	¶	Æ	Ö	æ	ö
7	ESA	EPA	§	·	Ç	×	ç	÷
8	HTS	SOS	¨	¸	È	Ø	è	ø
9	HTJ	XXX	©	¹	É	Ù	é	ù
A	VTS	SCI	ª	º	Ê	Ú	ê	ú
B	PLD	CSI	«	»	Ë	Û	ë	û
C	PLU	ST	¬	¼	Ì	Ü	ì	ü
D	RI	OSC	-	½	Í	Ý	í	ý
E	SS2	PM	®	¾	Î	Þ	î	þ
F	SS3	APC	¯	¿	Ï	ß	ï	ÿ

Figure 27. *Codage ISO8859-1 (Latin1), seconde partie (la première partie est équivalente au codage Ascii de la figure 23)*

55. On les trouve aussi à http://babel.alis.com/codage/iso8859/jeuxiso.htm.

56. Le choix des caractères non alphabétique est assez curieux. Outre des caractères présents à la suite d'erreur (p.ex. une mauvaise photocopie de | ayant « créé » la barre percée ¦), on peut se demander le pourquoi de certains autres (tels que ¼ ¬ ÷ µ etc.). La réponse semble politique !

Pour leurs saisie sur clavier, ces caractères sont souvent directement associés à une touche ou à une combinaison de plusieurs touches ; par exemple, sur les SUN/Unix, « © » se tape « composer C O » . Latin-1 est suffisamment important pour que Adobe ait créé un *ISOLATIN1 Encoding vector* (voir section 7) depuis longtemps, ce qui fait que tous ces caractères sont toujours présents dans toutes les fontes. Enfin, Latin-1 correspond aux premiers codes d'Unicode !

5.3.2. *ISO-Latin9*

ISO Latin-1 présente quelques lacunes pour les langues d'Europe occidentale. Outre les Œ, œ et Ÿ manquants, on a regretté l'absence de caractères carons. Par ailleurs, la création du signe euro nécessitait la définition d'un code pour ce nouveau glyphe. Il n'était pas question de modifier ISO Latin-1 : une nouvelle variante ISO8859 a donc été mise en chantier et après de nombreuses années de discussions (elle était connue sous le nom de Latin-0), elles a été adoptée en 1999 sous le nom d'ISO Latin-9, ou de ISO8859-16. Pour adopter de nouveaux symboles, il a fallu remplacer certains de Latin-1. La figure 28 montre les différences entre ces deux normes.

Code hexadécimal	Glyphe Latin-1	Remplacé par le caractère de Latin-9	Glyphe Latin-9
00A4	¤	SYMBOL EURO	€
00A6	¦	LETTRE MAJUSCULE LATINE S CARON	Š
00A8	¨	LETTRE MINUSCULE LATINE S CARON	š
00B4	´	LETTRE MAJUSCULE LATINE Z CARON	Ž
00B8	¸	LETTRE MINUSCULE LATINE Z CARON	ž
00BC	¼	DIGRAMME SOUDÉ MAJUSCULE LATIN OE	Œ
00BD	½	DIGRAMME SOUDÉ MINUSCULE LATIN OE	œ
00BE	¾	LETTRE MAJUSCULE LATINE Y TRÉMA	Ÿ

Figure 28. *Comparaison Latin-1 / Latin-9*

La norme ISO Latin-9 devrait donc être encore plus utilisée que ISO Latin-1 car elle offre l'intégrité des caractères français et le symbole euro. Si effectivement de nombreux constructeurs ont donné leur accord pour suivre cette norme, il se trouve qu'ils ont aussi donné leur accord pour Unicode et il est probable que ISO Latin-9, ayant tardé à sortir, ne soit pas vraiment adoptée !

5.3.3. *Multi-linguisme et ISO 8859*

Si chaque norme 8859 permet en général de couvrir plusieurs langues[57], il n'est pas possible (sauf en codant explicitement les changements de codage employés dans un texte) de mélanger deux normes 8859, par exemple d'écrire une phrase contenant du grec (ISO 8859-7) et de l'arabe (ISO 8859-6).

6. Autres codages à 8 bits

6.1. *Normes d'extensions*

On a vu que les machines à écrire (section 3.2) puis le télex (3.5 permettaient de faire correspondre deux caractères par touche grâce à un mécanisme (table haute ou basse pour la machine à écrire, codes d'inversion lettre/chiffre pour le télex). De même les codages BCD puis Ascii et donc Latin-1 comprenaient trois caractères de commande (ESC, SO et SI) qui permettaient d'étendre les jeux de caractères. Dès 1973, la première version de la norme ISO-2022 a explicité l'emploi de ces trois codes. Cette norme a été redéfinie en 1985 et permet de manipuler des jeux de 8 bits. Marti [MAR90, pages 247-255] donne les détails de fonctionnement de ces extensions qui sont finalement plutôt complexes. Cette norme a servi notamment à la définition des caractères mosaïques et à celle de la norme ISO-4873 d'où est issue la norme britannique de codages 8 bits BS-6006. Le même principe d'extension est utilisé par la norme japonaise JIS X 0201 (voir 6.3).

Unicode a pu se passer de ce concept, grâce à l'emploi de codage à 16 voire 20 bits.

6.2. *Alphabet phonétique international*

L'alphabet phonétique international API[58] a été défini par l'Association Internationale de Phonétique en 1993 et révisé en 1996. Il ne s'agit toutefois pas d'un codage dans le sens où nous l'entendons ici, chaque « fonte phonétique » utilisant son propre codage. Mais, ces caractères font maintenant partie d'Unicode (rangée 02) ce qui en fait désormais un répertoire *de facto*.

6.3. *Langues non latines*

La norme ISO-8859 couvre de nombreuses langues dont notamment l'arabe, l'hébreu, le cyrillique, le thaï et le grec. Mais souvent ces langues font l'objet d'autres codages, soit pour des raisons politiques ou historiques, soit pour des problèmes d'exhaustivité. Par ailleurs, les langues extrême-orientales nécessitent des

---

[57] Ainsi Latin-1 couvre les langues suivantes : l'albanais, l'allemand, l'anglais, le catalan, le danois, l'espagnol, le féroïen, le finnois, le français, le galicien, l'irlandais, l'islandais, l'italien, le néerlandais, le norvégien, le portugais et le suédois.

58. Ou *IPA (International Phonetic Alphabet)* ; http://www.arts.gla.ac.uk/IPA/ipa.html

milliers de codes qui ne peuvent tenir sur 8 bits ! La figure 29 donne quelques-uns de ces codages multi-octets[59]. Par ailleurs, de nombreuses RFC[60] ont été définies pour les langues concernées mais Unicode risque de périmer tous ces codages spécifiques

Langue	codages
Arabe	ASMO,ISO-9036,ISO-11822. (pour les bibliographies)
Chinois	CNS GB-2312-80 : 6723 'idéophonogrammes chinois CN GB-7590-87 et 12345-90 : jeux complémentaires BIG5 (Da wou) : standard couvrant environ 14 000 carctères
Coréen	KS X 1001:1992 (8 224 caractères dont 2 856 hanja)
Japonais	JIS X 0201 et 0202 : Katakana et Kanji phonétique. JIS 0212:1990 : 6067 caractères énumérés

Figure 29. *Codages multi-octets pour quelques langues non latines*

6.4. *Codages propriétaires*

Divers constructeurs ou logiciels utilisent des codes « propriétaires ». Si ceci pouvait rester anecdotique il y a encore peu (de nombreux programmes de conversions existant), l'utilisation massive de l'Internet depuis des PC fait que la moindre déviance par rapport à une norme devient exaspérante dès que l'on utilise un ordinateur d'une autre marque ou un programme diffèrent (en incluant notamment sous le nom de programme les outils de lecture des textes sur le web). Voici quelques exemples de codages propriétaires, c'est-à-dire propres à un constructeur et incompatibles avec d'autres codages !

6.4.1. *Apple*

En général, les Macintosh du monde occidental utilisent le *Standard Roman Character Set* (à 8 bits) dont les 128 premiers codes sont identiques à l'Ascii mais dont les suivants diffèrent partiellement d'ISO Latin-1. Un codage à 2 octets est disponible pour les langues orientales.

MacOS par ailleurs récupère les places des caractères de commande pour y mettre quelques caractères manquants de Latin-1 (comme œ) mais aussi des accents flottants (voir section 7).

6.4.2. *IBM*

IBM a défini, entre autres, EBCDIC (voir ci-dessus 5.2.1) et pour ses PC la notion de *code page* qui n'est jamais que l'adaptation à un pays donné du codage voulu.

59. On trouvera sur le site officiel de l'ISO/CEI *JTC 1/SC 2 - Coded Character Sets* un grand nombre de ces codages : http://www.dkuug.dk/.
60. *Request For Co*mments de l'IETF. Voir http://www.rfc-editor.org/.

6.4.3. *Microsoft*

Microsoft a défini principalement deux codages :
– codepage 850 pour le système DOS. Codage à 8 bits, dont la première partie est l'Ascii. Dans la second table on trouve diverses lettres accentuées, ligatures, mais aussi tout une panoplie de symboles graphiques (comme ╠ ▒ ►) qu'Unicode a repris dans sa rangée 25.
– codepage 1252 pour Windows avant 2000. Il était identique à ISO Latin-1 (mais a évolué, incluant par exemple le symbole euro qui n'est pas dans Latin-1) à cette différence près que Windows utilise les positions 0 à 5 (caractères de commande d'Ascii) pour y mettre quelques signes diacritiques et les positions 128-159 (caractères de commande de Latin-1) pour y mettre certains caractères tels que divers guillemets et apostrophes, les croix † et ‡, le caractère ‰, les tirets — et –, nos ligatures Œ et œ, etc.

Actuellement, Windows utilise un sous-ensemble d'Unicode et affiche même les noms des caractères selon la norme française.

6.4.4. *TeX*

Le monde de T_EX et LaT_EX à défini un codage, dit de Cork et utilisable par défaut avec le paquetage t1enc, qui est en fait un codage Latin-1 où les caractères de commande ont été remplacés dans la zone 0_{10}-31_{10} par des caractères supplémentaires, comme des accents flottants, guillemets et ligatures (fi, ffi,ffl, etc.) et dans la zone 128_{10}-159_{10} par des caractères de *Extended Latin* d'Unicode, comme đ ou ğ. Notons que ces derniers glyphes sont utilisés soit automatiquement (le compilateur TeX remplaçant la séquence « fi » par la ligature « fi ») soit par nom (la séquence « \^i » créant « î » et celle « \dd » nommant « đ » . D'autres paquetages permettent d'appeler d'autre codages, notamment pour les langues cyrilliques et orientales, voire pour tout Unicode (Omega).

7. Normes de glyphes

Puisque les glyphes peuvent être en nombre infini, il n'y en a bien sûr pas de norme. Toutefois de nombreux standards propriétaires essayent de modéliser les casses des fontes du temps du plomb, dont Type-1 (PostScript), TrueType et maintenant OpenType. Ils ont en commun de considérer qu'une fonte numérique est une base de données contenant des procédures de tracé de caractères qui sont appelées par nom et qui produisent le dessin d'un caractère en fonction d'un certain contexte (corps, espace graphique, couleur, etc.). Ces bases sont indépendantes des codages utilisés, un vecteur de codage *(encoding vector)* permettant d'associer au code du caractère un nom de glyphe (si ce vecteur est celui de Latin-1, on associera au caractère de code 0041_{16} le nom LATIN_A_CAPITAL). Le même jeu de caractères (par exemple Times-Roman) contient ainsi beaucoup plus de caractères que les 256 que l'on peut voir d'un coup.

8. Enfin vint Unicode

Le reste de ce numéro est consacré à Unicode et nous renvoyons donc le lecteur aux divers articles, notamment à celui d'introduction par Patrick Andries et à l'entretien de Ken Whistler qui raconte l'histoire d'Unicode.

9. Bibliographie

[AND01] André J. 2001, « Codage des caractères », *Techniques de l'ingénieur*, H7008, p. 1-18+F1-F9.

[AND02] Andries P. 2002, « Introduction à Unicode et à l'ISO 10646 »,, *Document numérique*, vol. 6, n° 3-4, 2002 (ce numéro).

[AND03] Andries P. 2003, *Traduction et annotation du standard Unicode 3.0*, Longueuil (à paraître ; voir http://iquebec.ifrance.com/hapax/).

[AND96] André J. 1996, « IsoLatin-1, une norme de codage de caractères européens ? trois caractères français en sont absent ! », *Cahier GUTenberg*, n° 25, p. 65-77.

[AND98] André J. 1998, « Iso-Latin 9, euro et typographie française », *Document numérique*, vol. 2, n° 2, p. 231-240.

[BER81] Bertho C 1981., *Télégraphes et téléphones – de Valmy au microprocesseur*, Le livre de poche.

[CHA99] Chartron G. et Noyer J.M. (eds.) 1999, « Normes et documents numériques: quels changements ? », *Solaris*, nº 6, http://www.info.unicaen.fr/bnum/jelec/Solaris/d06.

[DRE77] Dreyfus J. et Richaudeau F. (sous la direction de) 1977, *La chose imprimée*, éditions Retz-CEPL, Paris. Mise à jour par Marc Combier et Yvette Pesez*, Encyclopédie de la chose imprimée - du papier @ l'écran*; Retz, 1999.

[HAR02] Haralambous Y. 2002, « Unicode et typographie : un amour impossible », *Document numérique*, vol. 6, n° 3-4 (ce numéro).

[HER00] Hernandez A., « Normalisation et standardisation dans les nouvelles technologies », *Techniques de l'ingénieur*, H5018.

[MAR90] Marti B. et coauteurs 1990, *Télématique - techniques, normes, services*, Dunod,.

[MAR91] Marshall A. 1991, Ruptures et continuités dans un changement de systèmes techniques – le remplacement du plomb par la lumière dans la composition typographique, Thèse, Grenoble.

[PHI68] Phillips A. 1968, *Computer peripheral and typesetting*, Her Majesty's Stationery Office, Londres.

[RAND02] Randier O. 2002, « Unicode : tentations et limites – l'avis d'un typographe », *Document numérique*, vol. 6, n° 3-4 (ce numéro).

[UNI00] Unicode Consortium 2000, The Unicode Standard, Version 3.0, Addison-Wesley, Reading.