Petit précis de codage des caractères

Sans caractères, pas d'écriture, sans écriture, pas de livres, et sans livres, pas de bibliothèques ! On pourrait ainsi résumer l'importance, pour le professionnel des bibliothèques et de la documentation, de la gestion informatisée des jeux de caractères, dans un contexte où, après l'informatisation des catalogues, c'est à la numérisation des fonds qu'il faut faire face.

Yves Desrichard

Université de Montpellier
yves.desrichard@univ-montp1.fr

Élaborée par l'Occident, la technique informatique s'est peu à peu imposée au monde en exportant ses modes de représentation, au premier rang desquels se trouve l'alphabet latin. Nous sommes tellement habitués à utiliser cet alphabet que peu de gens se rendent compte que des milliards d'individus de par le monde utilisent couramment des systèmes d'écriture très différents.

L'usage des normes de translittération, qui était jusqu'alors la règle, est désormais complété par la possibilité de saisir directement dans leur écriture d'origine les descriptifs des documents produits avec d'autres écritures que l'alphabet latin, et les documents eux-mêmes en mode texte (autrement appelé « mode caractère »).

Qu'est-ce qu'une écriture ?

Pour Le Petit Robert, l'écriture est une « *représentation de la parole et de la pensée par des signes* », ce qu'on pourra trouver singulièrement général : après tout, un tableau peut être un ensemble de signes, mais il faut être critique d'art pour y voir une écriture. On préférera la définition proposée dans le magnifique ouvrage publié comme support d'une exposition organisée par la Bibliothèque nationale de France : « *Représentation visuelle du langage par un système de signes graphiques* [1] ». L'expression a le mérite de ne pas limiter les écritures, comme on le fait trop souvent, aux systèmes à base de lettres, voire de pictogrammes. Et elle introduit une notion fondamentale, celle de « signe graphique », et de sa matérialisation sur des supports très variés, pierre, terre séchée ou cuite, papyrus, papier, microfilm et, désormais, support optique ou magnétique.

Les différents types d'écriture

Schématiquement, on distingue trois grands « types » d'écriture. Pour autant, la plupart des écritures sont des combinaisons de ces trois types.

– *Systèmes idéographiques :* dans ces systèmes, chaque signe représente un objet (pictogramme) ou une idée (idéogramme). Pour prendre en compte la nécessité de transcrire de plus en plus de mots différents, ces écritures ont évolué dans un sens phonétique : on utilise les idéogrammes de mots existants pour exprimer les sons (et non plus la signification) du mot nouveau. Les systèmes idéographiques comprennent, par la force des choses, un très grand nombre de signes, parfois des dizaines de milliers.

– *Systèmes syllabiques :* dans les systèmes syllabiques, chaque signe

* Merci à Mireille Pénichon de sa relecture.

Note de l'auteur : ce texte est extrait d'un manuscrit intitulé « Bibliothèques et écritures, d'ASCII à Unicode », à paraître.

1. *L'aventure des écritures*, Bibliothèque nationale de France, 1997-1999, 3 vol.

Titulaire d'un DESS option informatique documentaire, ***Yves Desrichard*** *est conservateur à la bibliothèque interuniversitaire de Montpellier. Il a auparavant travaillé à Médiadix, à l'Abes, à la Bibliothèque interuniversitaire scientifique de Jussieu, au Centre national de la cinématographie et à la DLL. Il est l'auteur en 2001 de* Julien Duvivier, cinquante ans de noirs destins *et en 2003 de* Henri Decoin, un artisan du cinéma populaire *(Durante/Bibliothèque du film). Il vient de publier une nouvelle édition de* Administration et bibliothèques *(Éd. du Cercle de la librairie).*

représente un son. Une écriture syllabique comprend en moyenne 80 à 120 signes.

– *Systèmes alphabétiques :* un alphabet est une collection de signes graphiques qui représentent des sons vocaux dans une langue ou un groupe de langues données. En d'autres termes, chaque signe représente un son décomposé, et plusieurs signes (lettres ou caractères) sont nécessaires pour représenter un son. Les principaux alphabets comprennent une trentaine de signes au maximum.

Pour beaucoup, les écritures à alphabets sont l'aboutissement de l'évolution de l'écriture : en se dégageant de l'écriture du pictogramme ou de l'idéogramme, les civilisations indo-européennes apparaissent comme les civilisations du concept, au contraire des civilisations symboliques ou emblématiques de l'Extrême-Orient. Jean-Jacques Rousseau écrivait, en 1778, que « *ces trois manières d'écrire répondent exactement aux trois divers états sous lesquels on peut considérer les hommes rassemblés en nation. La peinture des objets convient aux peuples sauvages ; les signes des mots et des propositions aux peuples barbares et l'alphabet aux peuples policés*[2] ». On pourra ne pas être d'accord !

Caractéristiques des écritures

Les graphèmes (signes graphiques) sont l'unité fondamentale d'une écriture donnée : selon le type d'écriture, le graphème se réalise visuellement et phonétiquement de diverses matières, qu'on peut résumer comme suit :

Alphabets : un graphème = une lettre.

Syllabaires : un graphème = une syllabe.

Idéogrammes : un graphème = un caractère = une idée, un mot, un composé idéo-phonétique, etc.

Sens : les écritures peuvent se lire de gauche à droite ou l'inverse, de haut en bas ou l'inverse, ou de gauche à droite puis de droite à gauche (boustrophédon), etc. Le sens de la lecture peut être unique, ou mélangé. La ligne droite, horizontale ou verticale, reste la norme (sauf dans certaines représentations calligraphiques).

Présence de diacritiques : Le Petit Robert définit le diacritique comme un « *signe graphique (point, accent, cédille) portant sur une lettre ou un signe phonétique, et destiné à en modifier la valeur ou à empêcher la confusion entre homographes* ». Si, dans l'alphabet latin, les signes diacritiques sont le plus souvent situés au-dessus de la lettre qu'ils servent à qualifier, on peut aussi en trouver en dessous, ou à côté. Qui plus est, pour certaines langues, on peut avoir plusieurs diacritiques en même temps au-dessus ou au-dessous du caractère.

Ligatures et autres caractéristiques : la ligature consiste, lors de l'impression de certains caractères, à les lier entre eux. C'est vrai par exemple de « fi », de « ffi » ou de « ct ». Les ligatures ne peuvent pas être considérées comme des caractères à part entière.

Caractéristiques de transcription : la transcription des langues est un point tout particulièrement épineux, car elle rend l'utilisation des jeux de caractères encore plus complexe. Dans certaines langues, on n'écrit pas les voyelles, ou on les écrit pour donner un sens spécifique au texte transcrit (une valeur poétique dans le cas de l'hébreu par exemple). Dans d'autres écritures, le sens d'un idéogramme varie suivant la langue qu'il retranscrit (par exemple en chinois, en japonais et en coréen).

Enfin, certaines écritures comprennent des majuscules et des minuscules (comme l'alphabet latin) tandis que d'autres – dites « monocamérales » – ignorent cette notion, comme l'alphabet arabe.

En se dégageant de l'écriture du pictogramme ou de l'idéogramme, les civilisations indo-européennes apparaissent comme les civilisations du concept, au contraire des civilisations symboliques ou emblématiques de l'Extrême-Orient

Les caractères

Qu'est-ce qu'un caractère ?

Il serait tentant de considérer que « graphème » et « caractère » sont des termes équivalents. Le terme de « graphème » renvoyant plutôt à la linguistique, n'a que peu de pertinence dans le contexte informatique abordé ici, et on lui préférera le terme de « caractère » – sans en faire pour autant des synonymes. Cependant, définir ce qu'est précisément un caractère et le distinguer de sa manifestation graphique (le glyphe que nous verrons plus loin) n'est pas facile. C'est la fameuse histoire du chien : quand on dit « chien », chacun voit dans son esprit un chien différent, mais un chien tout de même.

2. Jean-Jacques Rousseau, *Essai sur l'origine des langues*, 1778.

Pour les caractères, la situation est un peu moins compliquée : lorsqu'on parle du « e accentué », tout le monde comprend de quoi il s'agit, mais on oublie sans doute qu'il peut y avoir des représentations très différentes du « e accentué » : de police et de corps, en italique, en gras, souligné, etc. Disons que le « caractère », c'est tout élément d'une écriture, et restons-en là : ce qui nous intéresse, ce sont ses manifestations sur les innombrables supports et dans les encore plus innombrables documents que doit gérer une bibliothèque ou un centre de documentation ! Cette manifestation passe toujours par le glyphe.

Glyphes et caractères

Faute d'avoir précisément défini ce qu'était un caractère, définissons précisément ce qu'est un glyphe : un glyphe est une forme utilisée pour représenter un caractère. Le glyphe est la manifestation du caractère. Un caractère peut correspondre à plusieurs glyphes… mais un même glyphe peut aussi correspondre à plusieurs caractères différents.

Le codage des caractères

Petite histoire

Comme son nom l'indique, le codage des caractères est basé sur la mise en correspondance d'un code (informatique) et d'un caractère. Les premières applications télégraphiques de transmission à distance de caractères d'écriture utilisèrent des codages conventionnels, dont s'inspireront les premiers concepteurs de jeux de caractères informatisés. Parmi eux, le plus célèbre est sans conteste l'alphabet morse, inventé par Samuel Finley Breese Morse dans les années 1830, et désormais abandonné. On peut aussi citer le Télex, service de dactylographie à distance qui, lui aussi, utilisait un « jeu de caractères » pour la transmission télégraphique des informations.

Mais c'est l'arrivée de l'informatique, puis de la micro-informatique, qui vont bouleverser les pratiques et multiplier les besoins. Comme dans beaucoup de domaines, l'influence américaine est fondamentale, puisque la plupart des grands constructeurs informatiques sont aux États-Unis.

Quelques rappels informatiques

Même si, informatiquement, le codage de caractères relève d'une démarche plutôt simple, il n'est sans doute pas inutile de rappeler quelques principes informatiques de base pour mieux comprendre.

Vers la fin des années 30, Claude Shannon démontra qu'à l'aide d'interrupteurs fermés pour « vrai » et ouverts pour « faux », on pouvait effectuer des opérations logiques en associant le chiffre « 1 » à « vrai » et « 0 » à « faux ». Ce langage (c'en est un, des plus rudimentaires) est appelé langage binaire, et c'est lui qui continue à être utilisé dans la majorité des ordinateurs. La plus petite unité d'information manipulable est donc « vrai/faux », ou « 1/0 ». On l'appelle « bit », pour « *binary digit* » ; 8 bits font 1 octet, soit 8 positions, chacune différente, de 0 et 1.

Avec 2 bits, on peut avoir quatre « états » différents : 00, 01, 10, 11. Avec 3 bits, huit états différents : 000, 001, 010, 011, 100, 101, 110, 111. Avec 8 bits, on aura 256 possibilités.

On peut définir un jeu de caractères comme la combinaison entre un répertoire de caractères et les codages correspondants

Principes

Un caractère sera « exprimé » par une série de « 0 » ou de « 1 », à raison de « 0 » ou « 1 » pour chacun des bits qui le définissent. Ainsi, en supposant (voir plus loin) des caractères définis à l'aide de 8 bits, soit un octet, la suite : 0111 0001 va définir la lettre « q » dans le codage ASCII sur 8 bits. C'est ce que l'on appelle le codage binaire du caractère. La correspondance entre « 0111 0001 » et « q » est *purement conventionnelle*, et suppose pour être traitée d'être comprise comme telle par l'ordinateur et le logiciel, pour lequel une tout autre correspondance peut avoir été définie. C'est la source de bien des problèmes…

Aujourd'hui et surtout demain, les jeux de caractères seront définis sur 16 ou 32 bits. Unicode utilise au départ 16 bits pour coder « ses » caractères. Il est inclus dans la norme ISO 10646 qui, elle, fonctionnera théoriquement sur des machines à 32 bits, ce qui représente… jusqu'à 2 milliards de combinaisons !

Jeux de caractères

Définition

On peut définir un jeu de caractères comme la combinaison entre un répertoire de caractères et les codages correspondants. Un répertoire de caractères, c'est une liste comportant un glyphe pour chaque caractère (s'il est représentable bien sûr, ce qui exclut par exemple les caractères de « contrôle »), un nom pour chaque caractère (représentable ou non) et éventuellement certaines caractéristiques d'utilisation de ce jeux de caractères par rapport au répertoire. Le codage définit la valeur binaire de ce caractère, avec un nombre de bits qui varie désormais de 8 à 32.

De nombreuses confusions viennent de ce que l'on peut utiliser le même répertoire ou (pire) presque le même répertoire, mais avec dif-

férents codages. Et que le nombre de combinaisons binaires possibles étant forcément limité, on est souvent obligé d'utiliser le même code binaire pour coder des caractères différents dans des jeux de caractères différents. Il faut donc apporter une grande attention, quand on se soucie des jeux de caractères gérés par le matériel ou le logiciel dont on souhaite faire l'acquisition, à disposer tout à la fois du tableau de codage ET du répertoire correspondant du jeu de caractères géré par ce matériel ou par ce logiciel.

Historique

Le premier jeu de caractères au sens où nous l'entendons ici fut élaboré pour le Télex, déjà cité. Basé sur 6 bits, il permettait donc le codage de 64 caractères : ceci explique que « télex » reste synonyme de concision de l'expression ! La première norme internationale, créée en 1967, fut la norme ASCII (*American Standard Coded Information Interchange*). Basée sur 7 bits (128 caractères), elle a survécu aux incroyables bouleversements de l'informatique moderne et contemporaine et, même s'il n'existe plus désormais de machines ou de systèmes qui fonctionnent sur 7 bits, elle reste encore très utilisée, dans la mesure où la majorité des jeux de caractères existants (notamment Unicode) utilise, pour les caractères qui étaient inclus dans cette norme, les mêmes combinaisons binaires.

À la suite de la norme ASCII, des dizaines de jeux de caractères, nationaux, internationaux, ou même propriétaires furent élaborés, d'où une monstrueuse cacophonie qui a, d'une part, gêné le développement informatique d'écritures trop complexes ou correspondant à des jeux de caractères trop peu utilisés et, d'autre part, favorisé le développement d'une langue qui avait le double mérite d'être celle des leaders informatiques mondiaux et de pouvoir s'écrire en utilisant uniquement les caractères définis par la norme ASCII : l'anglais.

Les codages binaires des caractères en minuscules de l'alphabet latin dans le codage ASCII (sur 8 bits)

a	01100001
b	01100010
c	01100011
d	01100100
e	01100101
f	01100110
g	01100111
h	01101000
i	01101001
j	01101010
k	01101011
l	01101100
m	01101101
n	01101110
o	01101111
p	01110000
q	01110001
r	01110010
s	01110011
t	01110100
u	01110101
v	01110110
w	01110111
x	01111000
y	01111001
z	01111010

Aujourd'hui, la situation s'est améliorée. De nombreux jeux de caractères permettent de coder des écritures très différentes d'une manière satisfaisante, et certains jeux, quoique basés sur l'ASCII, ont pu assurer la présence de langues écrites à l'aide d'autres alphabets que l'alphabet latin de base. Parmi les plus répandues, on peut citer la norme ISO-Latin-1, bien connue des utilisateurs du web, et Unicode.

Norme ASCII

Créée en 1967, la norme américaine ASCII, transformée ultérieurement en norme internationale ISO 646, est restée la seule norme de jeu de caractères non ambiguë pendant près de vingt ans. Elle correspond essentiellement au codage d'une langue, l'anglais, et se limite, pour ce qui est des lettres, aux 26 lettres de l'alphabet latin, en minuscules et en majuscules (cf. tableau).

Normes ISO 8859-n

Pour permettre de transcrire non seulement les langues utilisant l'alphabet latin, mais aussi les langues correspondant à d'autres alphabets, l'association internationale de normalisation (ISO) élabora dans les années 80 un ensemble de normes, regroupées sous la dénomination ISO 8859-n. Ces normes ont été conçues pour traiter l'ensemble des langues à alphabets comprenant des caractères accentués : l'écriture latine, mais aussi le cyrillique, l'arabe, l'hébreu, le grec…

Les normes ISO codent les caractères sur 8 bits. Elles ne permettent donc de coder, pour chacune d'entre elles, que 256 caractères. Mais 128 de ces caractères sont déjà « réservés », puisqu'ils correspondent aux positions et caractères de la norme ASCII « d'origine ».

En fait, seule la norme ISO 8859-1, dite ISO-Latin-1, s'est véritablement imposée, notamment sur internet, puisque ce jeu de caractères était le seul accepté dans la première mouture d'HTML. Il permet de représenter la plupart des langues de l'Europe occidentale : 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.

Il a permis de développer sur internet notamment l'écriture d'autres langues que l'anglais, mais dans une perspective qui reste, on le constate, très occidentale.

Unicode

Le consortium

Le consortium Unicode[3] a été créé en 1989 par de grandes sociétés informatiques comme Adobe, Apple, IBM, Microsoft, Sun, Xerox… Unicode partait du principe qu'avec 16 bits (= 2 octets), on peut coder 65 536 caractères, soit l'essentiel des langues écrites. Le système est conçu pour évoluer jusqu'à la prise en compte de caractères définis par 32 bits (comme la norme ISO 10646 évoquée ci-après) mais, actuellement, la « limite » se situe entre 20 et 21 bits. Unicode a déjà défini 245 000 codes différents. Sur ce nombre, on peut considérer qu'environ 100 000 sont utilisés de manière commune pour coder les écritures les plus courantes – sachant que 70 000 codes sont utilisés pour des signes idéographiques.

Les principes

Unicode s'appuie sur un certain nombre de principes, qui définissent son contenu, son utilisation, et conditionnent son évolution. Contrairement aux jeux de caractères qui l'ont précédé, et profitant de leurs lacunes et de leurs manques, ce n'est plus un « simple » jeu de caractères, et on ne peut le réduire ni à un tableau ni même à un répertoire. Parmi les principaux éléments, on peut distinguer les suivants.

Pour Unicode, un caractère est « *the smallest component of written language that has semantic value* » : la plus petite composante d'un langage écrit ayant une valeur sémantique. Les caractères sont essentiellement des unités abstraites, dont les glyphes ne sont qu'une représentation visuelle. Unicode gère des écritures et non des langues : pour Unicode, un « u » est un « u », qu'il soit utilisé en allemand, en anglais ou en français (ce qui implique pourtant qu'il se prononcera différemment). De la même manière, un idéogramme *han* aura le même code, qu'il soit utilisé en chinois, en japonais ou en coréen.

Chaque caractère est défini par un ensemble de critères : ce principe explique qu'on ne puisse définir Unicode uniquement en tant que jeu de caractères. Si chaque caractère codé a bien un nom, un glyphe, un codage binaire, Unicode recense aussi un grand nombre d'informations supplémentaires sur chaque caractère, jusqu'à 50 différentes (par exemple, s'il s'agit d'une majuscule ou d'une minuscule, ou sur le sens de l'écriture ou sur les écritures qui utilisent ce caractère, etc.).

Unicode et UTF

UTF signifie « *Unicode Transformation Format* ». Les algorithmes UTF sont des « ruses » pour coder les caractères d'Unicode, qui sont sur deux, trois, bientôt quatre octets, avec un seul octet (UTF-8) ou deux octets (UTF-16), ou plutôt en utilisant des octets successivement pour un seul caractère. Là où Unicode utilise obligatoirement plusieurs octets pour chaque caractère codé, UTF ne les utilise qu'en tant que de besoin : cela permet par exemple que les caractères ASCII soient codés en UTF-8 sur un seul octet. Les avantages sont nombreux : économie en termes de stockage, possibilité d'utiliser certains logiciels originellement prévus pour traiter des caractères d'un seul octet, facilités de transmission des données avec des protocoles qui utilisent surtout des codes à 8 bits, etc. Il faut bien noter que ces UTF sont aussi définis pour ISO 10646 (voir ci-après), et non seulement pour Unicode. Cependant, puisque Unicode est pour l'instant le seul standard de codage inclus dans ISO 10646, il est majoritairement concerné par ces algorithmes de conversion. Le plus connu est UTF-8, dans lequel les codages sont transformés en codages sur 8 bits, de 1 à 6 octets étant nécessaires pour un caractère. Pour les codages correspondant au répertoire ASCII, ce sont d'ailleurs les mêmes codages sur 1 octet que ceux du jeu « standard », là encore pour faciliter la transition.

ISO 10646

Comme son nom l'indique, et contrairement à Unicode, ISO 10646 est une norme ISO. Au départ, les deux initiatives se sont développées parallèlement. Puis l'organisation internationale et le consortium ont décidé de collaborer, sans que pour autant il faille considérer qu'Unicode et ISO 10646 sont la même chose, ni même, comme on pourrait le penser un peu rapidement, qu'Unicode n'est qu'un simple sous-ensemble d'ISO 10646. ISO 10646 est basé sur l'utilisation de codes à 32 bits : cette norme permet donc en théorie le codage de plus de deux milliards de caractères différents.

> Si le passage généralisé à Unicode résoudra bien des problèmes en matière de gestion de polices et de fontes, il pourra aussi en poser

Polices et fontes

La représentation informatique des glyphes renvoie aux notions, issues de l'imprimerie, de « polices » et de « fontes », notions qui, dans l'univers « immatériel », prennent un sens très différent. De la même manière que la calligraphie a conditionné la typographie, celle-ci, à son tour, a nourri l'élaboration des glyphes utilisés dans les applications informatiques.

3. www.unicode.org

Les *polices* définissent les formes de caractères, c'est-à-dire un ensemble de glyphes pour un répertoire (un jeu de caractères) donné possédant les mêmes caractéristiques. C'est tout l'objet de la typographie. Il est dommage que cet art, si important en matière de livres imprimés, ne soit pas encore convenablement exercé dans l'univers des sites web, et autres bases de données en texte intégral.

Les *fontes*, sous-ensembles des polices, définissent des assortiments de caractères dans une taille particulière (exprimée généralement en points) avec des attributs particuliers (gras, italique, souligné, etc.). Ainsi, à chaque police peuvent correspondre différentes fontes et, théoriquement, il y a autant de fontes que de combinaisons possibles de tailles et d'attributs.

Dans la pratique, et comme les modes de constitution des caractères varient suivant la taille et la nature de l'attribut, cette distinction entre polices et fontes n'a plus, dans le monde informatique, grand sens et, pour être logique, il vaudrait mieux distinguer les ensembles de glyphes suivant leur mode d'élaboration informatique.

Ces modes sont essentiellement de trois types : « bitmap », « vectoriel » et « en traits ». La place manque pour en détailler le mode de fonctionnement et les usages.

La gestion informatique des polices et des fontes est complexe, car celles-ci sont liées au système d'exploitation, voire au type de matériel utilisé et, bien sûr, au jeu de caractères impliqué.

Si le passage généralisé à Unicode résoudra bien des problèmes en matière de gestion de polices et de fontes, il pourra aussi en poser. Il est en effet difficile de disposer de polices pour TOUS les caractères disponibles dans Unicode. C'est d'abord la taille des polices qui est en cause : UNE police pour 50 000 glyphes occupe déjà presque 25 mégaoctets de mémoire ! Sur des disques durs de plusieurs gigaoctets de capacité, leur stockage n'est cependant plus vraiment un problème. C'est l'utilisation de telles polices qui risque de ralentir considérablement le temps de traitement des données, un peu comme quand on manipule des images d'un « poids » important.

La typographie est, le plus souvent, l'âme d'un texte, ce qui permet de l'articuler, de lui donner sens par la forme autant que par le contenu

L'âme du texte

Qu'il soit permis de conclure ce texte en insistant, au-delà de la nécessité de gérer des écritures différentes, sur l'importance de la typographie. Cette technique n'est en effet ni un divertissement d'esthète, ni une lubie d'imprimeur. La typographie est, le plus souvent, l'âme d'un texte, ce qui permet de l'articuler, de lui donner sens par la forme autant que par le contenu. Bien utilisée, elle aide autant le rédacteur à exprimer ses idées par ses écrits que le lecteur à mieux les comprendre.

Qu'on imagine la lecture à haute voix d'un texte sur un ton monocorde, sans intonation, sans intention ni expression, et on aura une idée de ce que représente un texte sans l'usage de la typographie – le retour à « l'âge de pierre » de l'information. La typographie, dans sa complexité, dans sa diversité, reste une des façons inventées par l'homme pour mieux s'exprimer, individuellement ou collectivement, et il appartient aux professionnels des bibliothèques, peut-être plus qu'à d'autres, d'assurer la transmission de ce précieux capital.

Février 2007