CODAGE INFORMATIQUE DES SYSTÈMES D'ÉCRITURE

Version 2.0.1

Gianni Vacca - giannieanna@infonie.fr

Introduction

Je vous propose une petite doc sans prétention car je me suis aperçu qu'un nombre préoccupant d'internautes ignoraient tout du codage informatique des systèmes d'écriture. Il s'agira uniquement d'une présentation générale basée sur mon intérêt extra-professionnel pour tout ce qui a trait à la linguistique ; vous trouverez donc ici un aperçu des diverses normes, et non un exposé technique.

L'informatique ayant été conçue par des Américains, dont la langue utilise uniquement l'alphabet latin sans aucun accent et, qui plus est, par des radins (cf. le bogue de l'an 2000) qui ont voulu minimiser le nombre de bits employés, nous nous retrouvons aujourd'hui avec une foultitude de normes et de systèmes de codage divers pour les différents systèmes d'écriture utilisés au monde. En gros, dès qu'on sort du domaine anglophone, nous sommes perdus dans une véritable Babel du codage, qui empoisonne la vie de tous ceux qui transfèrent des fichiers d'un système d'exploitation à un autre, ou qui veulent utiliser des systèmes d'écriture différents au sein d'un même document.

Je vais essayer de faire une brève présentation des systèmes les plus utilisés (notamment sur Windows, ou pour le codage dans les fichiers **HTML**) en partant de mon expérience purement empirique.
Notre tour d'horizon partira de l'**ascii** sur sept bits pour arriver à la norme **UCS-4** : du moins complet au plus complet.

1. ascii sur sept bits ou us-ascii ou ISO 646 ou ANSI X3.4

Cette norme code l'alphabet latin sur sept bits (de x20 à x7E). C'est la base à partir de laquelle toutes les autres normes ont été conçues, ce qui signifie que les caractères codés de 32 à 126, quelle que soit la norme, seront gardés tels quels en cas de changement de codage (ex. : transfert d'un fichier de Mac à PC : les caractères accentués sont affichés de façon incorrecte, les caractères non-accentués le sont de façon correcte).

20	21 !	22 "	23 #	24 $	25 %	26 &	27 '	28 (	29)	2A *	2B +	2C ,	2D -	2E .	2F /
30 0	31 1	32 2	33 3	34 4	35 5	36 6	37 7	38 8	39 9	3A :	3B ;	3C <	3D =	3E >	3F ?
40 @	41 A	42 B	43 C	44 D	45 E	46 F	47 G	48 H	49 I	4A J	4B K	4C L	4D M	4E N	4F O
50 P	51 Q	52 R	53 S	54 T	55 U	56 V	57 W	58 X	59 Y	5A Z	5B [	5C \	5D]	5E ^	5F _
60 `	61 a	62 b	63 c	64 d	65 e	66 f	67 g	68 h	69 i	6A j	6B k	6C l	6D m	6E n	6F o
70 p	71 q	72 r	73 s	74 t	75 u	76 v	77 w	78 x	79 y	7A z	7B {	7C |	7D }	7E ~	

Évidemment, cette norme (1968) ne satisfit pas les pays plus civilisés tels que la France, l'Allemagne ou l'Italie car elle ne permettait pas de coder les caractères accentués dont nous autres Européens sommes si friands. La suite logique de cette norme fut donc d'utiliser le 8ème bit laissé libre (au départ prévu pour coder le surlignage) pour des extensions nationales. Ce furent les diverses normes ISO 8859-*N*., où N vaut entre 1 et 15.

2. les normes ISO 8859-N

Ces normes datent de la moitié des années quatre-vingts et sont encore de très loin les plus utilisées actuellement pour les langues européennes ou en général pour les langues qui s'écrivent avec un système alphabétique (cyrillique, grec, hébreu, arabe, thaï) par opposition aux langues qui s'écrivent avec un système d'idéogrammes (**CJK** pour les intimes : Chinese Japanese Korean).
Elles ont l'avantage de n'utiliser que 8 bits (en ajoutant le "bloc" xA1 à xFF à l'us-ascii) pour un codage satisfaisant de tous les systèmes alphabetiques ; ce sont les normes "de base" utilisées pour le HTML (cf. plus bas).

2.1 - la norme ISO 8859-1 ou Latin 1 ou Europe Occidentale

Cette norme code tous les alphabets d'Europe Occidentale (à l'exception notable du gallois) en utilisant le 8ème bit pour les caractères accentués. C'est celle que nous utilisons quotidiennement sur notre PC (bureautique, e-mail, HTML, **polices True Type** cf. plus bas) et aussi sur les stations utilisant X11 et les DEC.

A0	A1 ¡	A2 ¢	A3 £	A4 ¤	A5 ¥	A6 ¦	A7 §	A8 ¨	A9 ©	AA ª	AB «	AC ¬	AD -	AE ®	AF ¯
B0 °	B1 ±	B2 ²	B3 ³	B4 ´	B5 µ	B6 ¶	B7 ·	B8 ¸	B9 ¹	BA º	BB »	BC ¼	BD ½	BE ¾	BF ¿
C0 À	C1 Á	C2 Â	C3 Ã	C4 Ä	C5 Å	C6 Æ	C7 Ç	C8 È	C9 É	CA Ê	CB Ë	CC Ì	CD Í	CE Î	CF Ï
D0 Ð	D1 Ñ	D2 Ò	D3 Ó	D4 Ô	D5 Õ	D6 Ö	D7 ×	D8 Ø	D9 Ù	DA Ú	DB Û	DC Ü	DD Ý	DE Þ	DF ß
E0 à	E1 á	E2 â	E3 ã	E4 ä	E5 å	E6 æ	E7 ç	E8 è	E9 é	EA ê	EB ë	EC ì	ED í	EE î	EF ï
F0 ð	F1 ñ	F2 ò	F3 ó	F4 ô	F5 õ	F6 ö	F7 ÷	F8 ø	F9 ù	FA ú	FB û	FC ü	FD ý	FE þ	FF ÿ

Ceux qui suivent auront remarqué que le "bloc" x80 à x9F reste inutilisé. Sur votre PC, en allant dans le menu *Accessoires*, lancez la *Table de caractères*. Choisissez le sous-ensemble *Latin-1* ; une ligne reste, en effet, vide. Choisissez maintenant le sous-ensemble *Caractères Windows*, la ligne en question se remplit de caractères étendus supplémentaires. C'est que Microsoft, qui "n'aime pas gâcher", apparemment, a utilisé ce vide entre deux normes pour coder quelques caractères supplémentaires, dont l'Euro (€) en x80. Voir plus bas le paragraphe sur les **jeux étendus Windows**.

2.2 - la norme ISO 8859-2 ou Latin 2 ou Europe Centrale ou Europe Orientale

Cette norme code tous les alphabets d'Europe Centrale (à l'exception de ceux qui utilisent l'alphabet cyrillique) en utilisant le 8ème bit pour les caractères accentués. C'est celle qu'utilisent normalement les PC vendus en Europe Centrale.

A0	A1 Ą	A2 ˘	A3 Ł	A4 ¤	A5 Ľ	A6 Ś	A7 §	A8 ¨	A9 Š	AA Ş	AB Ť	AC Ź	AD -	AE Ž	AF Ż
B0 °	B1 ą	B2 ˛	B3 ł	B4 ´	B5 ľ	B6 ś	B7 ˇ	B8 ¸	B9 š	BA ş	BB ť	BC ź	BD ˝	BE ž	BF ż
C0 Ŕ	C1 Á	C2 Â	C3 Ă	C4 Ä	C5 Ĺ	C6 Ć	C7 Ç	C8 Č	C9 É	CA Ę	CB Ë	CC Ě	CD Í	CE Î	CF Ď
D0 Đ	D1 Ń	D2 Ň	D3 Ó	D4 Ô	D5 Ő	D6 Ö	D7 ×	D8 Ř	D9 Ů	DA Ú	DB Ű	DC Ü	DD Ý	DE Ţ	DF ß
E0 ŕ	E1 á	E2 â	E3 ă	E4 ä	E5 ĺ	E6 ć	E7 ç	E8 č	E9 é	EA ę	EB ë	EC ě	ED í	EE î	EF ď
F0 đ	F1 ń	F2 ň	F3 ó	F4 ô	F5 ő	F6 ö	F7 ÷	F8 ř	F9 ů	FA ú	FB ű	FC ü	FD ý	FE ţ	FF ˙

C'est pour cette raison que, si vous créez un site web perso qui accueille les internautes par un sympathique

Bonjour à tous, ici on fait la fête !

codé par exemple
```
<H1 align=center>Bonjour à tous, ici on fait la fête !</H1>
```
vos amis tchèques liront un mystérieux

Bonjour ŕ tous, ici on fait la fęte !

compréhensible, certes, mais intriguant. Nous verrons plus loin comment éviter ce genre d'inconvénient.

2.3 - la norme ISO 8859-3 ou Latin3 ou Europe Méridionale

Cette norme code le maltais, le turc et, accessoirement, l'esperanto. Son nom est un peu tiré par les cheveux, vu que le portugais, l'espagnol, l'italien ou l'albanais sont codés par la norme ISO 8859-1, et que le grec et le bulgare sont codés par d'autres normes encore. J'imagine que c'était par un souci de symétrie avec "Europe Occidentale" et "Europe Orientale".

A0	A1 Ħ	A2 ˘	A3 £	A4 ¤		A6 Ĥ	A7 §	A8 ¨	A9 İ	AA Ş	AB Ğ	AC Ĵ	AD -		AF Ż
B0 °	B1 ħ	B2 ²	B3 ³	B4 ´	B5 µ	B6 ĥ	B7 ·	B8 ¸	B9 ı	BA ş	BB ğ	BC ĵ	BD ½		BF ż
C0 À	C1 Á	C2 Â		C4 Ä	C5 Ċ	C6 Ĉ	C7 Ç	C8 È	C9 É	CA Ê	CB Ë	CC Ì	CD Í	CE Î	CF Ï
	D1 Ñ	D2 Ò	D3 Ó	D4 Ô	D5 Ġ	D6 Ö	D7 ×	D8 Ĝ	D9 Ù	DA Ú	DB Û	DC Ü	DD Ŭ	DE Ŝ	DF ß
E0 à	E1 á	E2 â		E4 ä	E5 ċ	E6 ĉ	E7 ç	E8 è	E9 é	EA ê	EB ë	EC ì	ED í	EE î	EF ï
	F1 ñ	F2 ò	F3 ó	F4 ô	F5 ġ	F6 ö	F7 ÷	F8 ĝ	F9 ù	FA ú	FB û	FC ü	FD ŭ	FE ŝ	FF ˙

2.4 - la norme ISO 8859-4 ou Latin 4 ou Europe Septentrionale

Cette norme code les langues des Pays Baltes, le lapon, et le groënlandais. Même remarque que ci-dessus, le nom me paraît franchement tiré par les cheveux vu que les langues scandinaves sont du ressort de la norme ISO 8859-1.

A0	A1 Ą	A2 ĸ	A3 Ŗ	A4 ¤	A5 Ĩ	A6 Ļ	A7 §	A8 ¨	A9 Š	AA Ē	AB Ģ	AC Ŧ	AD -	AE Ž	AF ¯
B0 °	B1 ą	B2 ˛	B3 ŗ	B4 ´	B5 ĩ	B6 ļ	B7 ˇ	B8 ¸	B9 š	BA ē	BB ģ	BC ŧ	BD Ŋ	BE ž	BF ŋ
C0 Ā	C1 Á	C2 Â	C3 Ã	C4 Ä	C5 Å	C6 Æ	C7 Į	C8 Č	C9 É	CA Ę	CB Ë	CC Ė	CD Í	CE Î	CF Ī
D0 Đ	D1 Ņ	D2 Ō	D3 Ķ	D4 Ô	D5 Õ	D6 Ö	D7 ×	D8 Ø	D9 Ų	DA Ú	DB Û	DC Ü	DD Ũ	DE Ū	DF ß
E0 ā	E1 á	E2 â	E3 ã	E4 ä	E5 å	E6 æ	E7 į	E8 č	E9 é	EA ę	EB ë	EC ė	ED í	EE î	EF ī
F0 đ	F1 ņ	F2 ō	F3 ķ	F4 ô	F5 õ	F6 ö	F7 ÷	F8 ø	F9 ų	FA ú	FB û	FC ü	FD ũ	FE ū	FF ˙

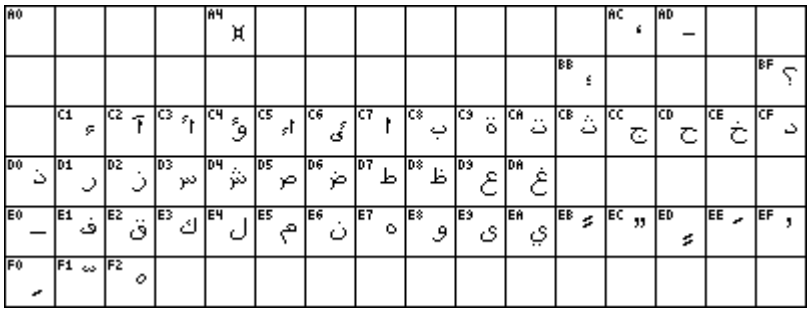
2.5 - la norme ISO 8859-5 ou Cyrillique

Cette norme code les langues européennes utilisant l'alphabet cyrillique telles que le russe, le bulgare ou le serbe. C'est donc celle qu'utilisent normalement les PC vendus en ex-Union Soviétique, par exemple.

Malheureusement, la Guerre Froide a encore frappé ici, et les Soviétiques avaient créé leurs propres normes **KOI** incompatibles avec la série ISO 8859-N avant que les statues de Lénine ne commencent à chuter un peu partout.
Sur internet, les normes ISO 8859-5 / KOI-8 / WinCyrillic (cf. plus bas) se répartissent de façon assez équitable les sites en langue russe, ce qui fait que même un paramétrage correct de votre fureteur n'empêchera pas que deux pages russes sur trois apparaissent comme du charabia incompréhensible.

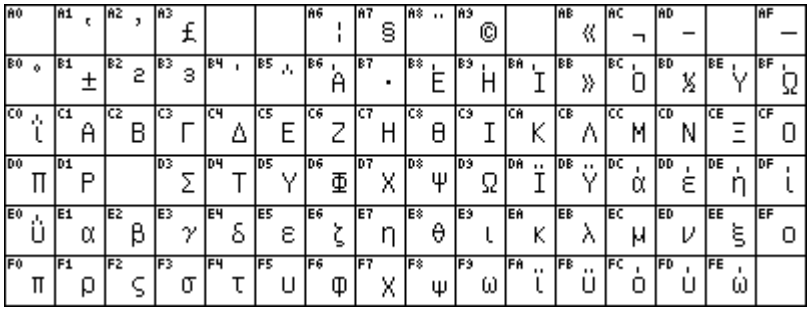
2.6 - la norme ISO 8859-6 ou Arabe

Cette norme code l'arabe. Même si vous installez une police compatible avec cette norme sur votre PC, vous ne pourrez pas lire de documents écrits en arabe car il vous manquera l'outil qui intervertit la direction d'écriture.



2.7 - la norme ISO 8859-7 ou Grec

Cette norme code le grec moderne (celui de Demis Roussos, donc, et pas celui d'Homère). On la nomme également ELOT-928.
À ma connaissance, contrairement à ce qui a lieu pour le cyrillique, elle n'a pas de norme concurrente.



2.8 - la norme ISO 8859-8 ou Hébraïque

Cette norme code l'hébreu moderne et le yiddish. Même remarque que pour la norme ISO 8859-6 pour ce qui est de la direction d'écriture.

2.9 - la norme ISO 8859-9 ou Latin 5

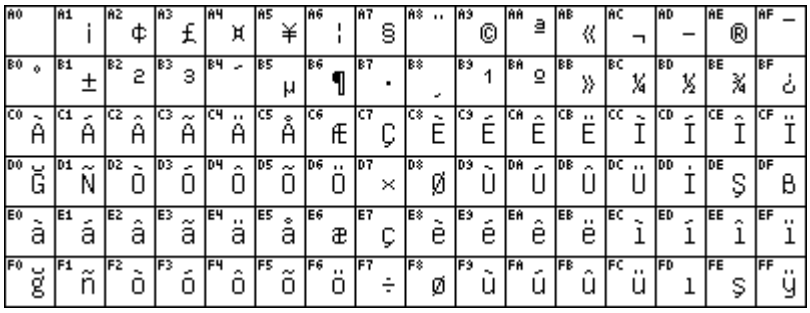
J'espère que vous suivez toujours... ici N=9 mais on parle bien de Latin 5, ce qui est logique puisque les normes ISO 8859-5 à 8 codent des alphabets non-latins.
Cette norme est la même que la norme ISO 8859-1 mais les lettres spécifiquement islandaises

ð þ ý

des positions xF0, xFE et xFD ont été remplacées par les lettres spécifiquement turques

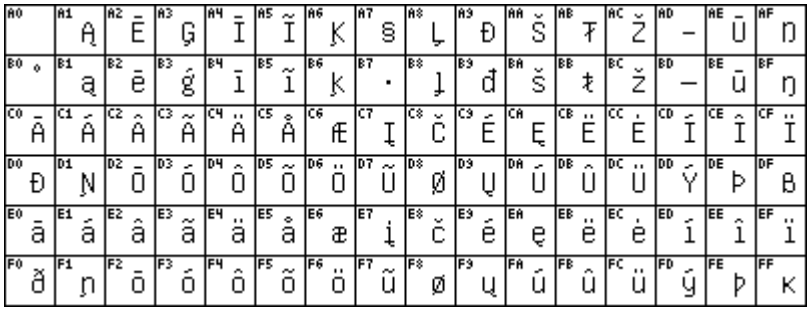
ğ ş ı

Cette norme a remplacé la norme ISO 8859-3 en ce qui concerne le turc. Elle a en effet l'avantage de pouvoir coder dans le même document le turc et les langues d'Europe Occidentale (qui sont tout de même plus fréquentes que le maltais ou l'esperanto !).



2.10 - la norme ISO 8859-10 ou Latin 6

Cette norme, comme toutes les normes ISO 8859-N où N > 9, est récente puisqu'elle ne date que de 1992. Il s'agit en fait d'un réarrangement de la norme ISO 8859-4 dont certaines lettres peu utilisées ont été enlevées pour faire place aux fameuses lettres islandaises et mériter ainsi le nom de "Europe Septentrionale". Peine perdue puisque le nom reste attaché à la norme ISO 8859-4. Bad luck.



2.11 - la norme ISO 8859-11 ou Thaï

Cette norme code le thaï. Vu qu'il s'agit d'une norme récente, je n'ai aucune idée quant à sa diffusion réelle, et je n'ai pas de copains thaïs pour me renseigner. Je suis preneur de toute info à ce sujet.

	A1	A2	A3	A4	A5	A6	A7	A8	A9	AA	AB	AC	AD	AE	AF
	ก	ข	ฃ	ค	ฅ	ฆ	ง	จ	ฉ	ช	ซ	ฌ	ญ	ฎ	ฏ
B0	ฐ	ฑ	ฒ	ณ	ด	ต	ถ	ท	ธ	น	บ	ป	ผ	ฝ	พ	ฟ
C0	ภ	ม	ย	ร	ฤ	ล	ฦ	ว	ศ	ษ	ส	ห	ฬ	อ	ฮ	ฯ
D0	ะ	ั	า	ำ	ิ	ี	ึ	ื	ุ	ู	ฺ					฿
E0	เ	แ	โ	ใ	ไ	ๅ	ๆ	็	่	้	๊	๋	์	ํ	๎	๏
F0	๐	๑	๒	๓	๔	๕	๖	๗	๘	๙	๚	๛				

2.12 - la norme ISO 8859-12 ou Indien

Cette norme n'existe pas encore. La place est réservée pour une hypothétique norme "Indienne". Vu le nombre d'alphabets différents utilisés en Inde, 8 bits me paraissent trop peu nombreux pour cette tâche herculéenne. On verra bien.

2.13 - la norme ISO 8859-13 ou Latin 7

Cette norme est censée remplacer la norme ISO 8859-10 pour ce qui est des langues des Pays Baltes (en effet, il manque une lettre lettonne à cette dernière norme).

A0	A1	A2	A3	A4	A5	A6	A7	A8	A9	AA	AB	AC	AD	AE	AF
	”	¢	£	¤	„	¦	§	Ø	©	Ŗ	«	¬	-	®	Æ
B0	°	±	²	³	“	µ	¶	·	ø	¹	ŗ	»	¼	½	¾	æ
C0	Ą	Į	Ā	Ć	Ä	Å	Ę	Ē	Č	É	Ź	Ė	Ģ	Ķ	Ī	Ļ
D0	Š	Ń	Ņ	Ó	Ō	Õ	Ö	×	Ų	Ł	Ś	Ū	Ü	Ż	Ž	ß
E0	ą	į	ā	ć	ä	å	ę	ē	č	é	ź	ė	ģ	ķ	ī	ļ
F0	š	ń	ņ	ó	ō	õ	ö	÷	ų	ł	ś	ū	ü	ż	ž	’

2.14 - la norme ISO 8859-14 ou Latin 8

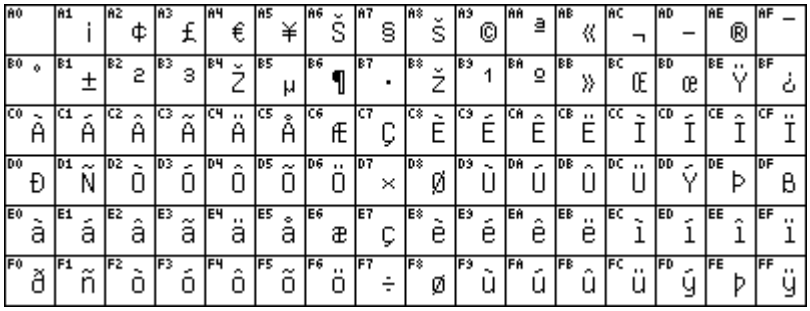
Cette norme est une adaptation de la norme ISO 8859-1 avec les w accentués qui manquaient pour couvrir le gallois, et avec les consonnes surmontées d'un point pour indiquer l'amuissement en irlandais (ce qui n'était pas franchement indispensable, vu qu'on peut tout aussi bien l'indiquer avec un h après la consonne en question). Cette norme code donc toutes les langues celtiques.

A0	A1	A2	A3	A4	A5	A6	A7	A8	A9	AA	AB	AC	AD	AE	AF
	Ḃ	ḃ	£	Ċ	ċ	Ḋ	§	Ẁ	©	Ẃ	ḋ	Ỳ	-	®	Ÿ
B0	Ḟ	ḟ	Ġ	ġ	Ṁ	ṁ	¶	Ṗ	ẁ	ṗ	ẃ	Ṡ	ỳ	Ẅ	ẅ	ṡ
C0	À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î	Ï
D0	Ŵ	Ñ	Ò	Ó	Ô	Õ	Ö	Ṫ	Ø	Ù	Ú	Û	Ü	Ý	Ŷ	ß
E0	à	á	â	ã	ä	å	æ	ç	è	é	ê	ë	ì	í	î	ï
F0	ŵ	ñ	ò	ó	ô	õ	ö	ṫ	ø	ù	ú	û	ü	ý	ŷ	ÿ

2.15 - la norme ISO 8859-15 ou Latin 9

Cette norme est une adaptation de la norme ISO 8859-1 avec la lettre ligaturée "e dans l'o" (œ) qui manquait inexpliquablement du jeu de cette dernière (c'est pour cette raison que les *œ* sont souvent ignorés par les fureteurs, mieux vaut utiliser *oe* dans vos pages web). Une énorme injustice vis-à-vis de notre belle langue française est enfin réparée !
Vu qu'il s'agit d'une norme récente (1998), elle comporte également le symbole de l'euro (€).
Malgré ces petits avantages par rapport à la norme ISO 8859-1, elle n'a pas franchement eu de succès…
Voici par exemple ce que pensait Reuters de la norme ISO 8859-15 (je cite) :

> *« We have just the place for ISO 8859-15 here in London. It is called the Science Museum and is full of charming historical relics. »*



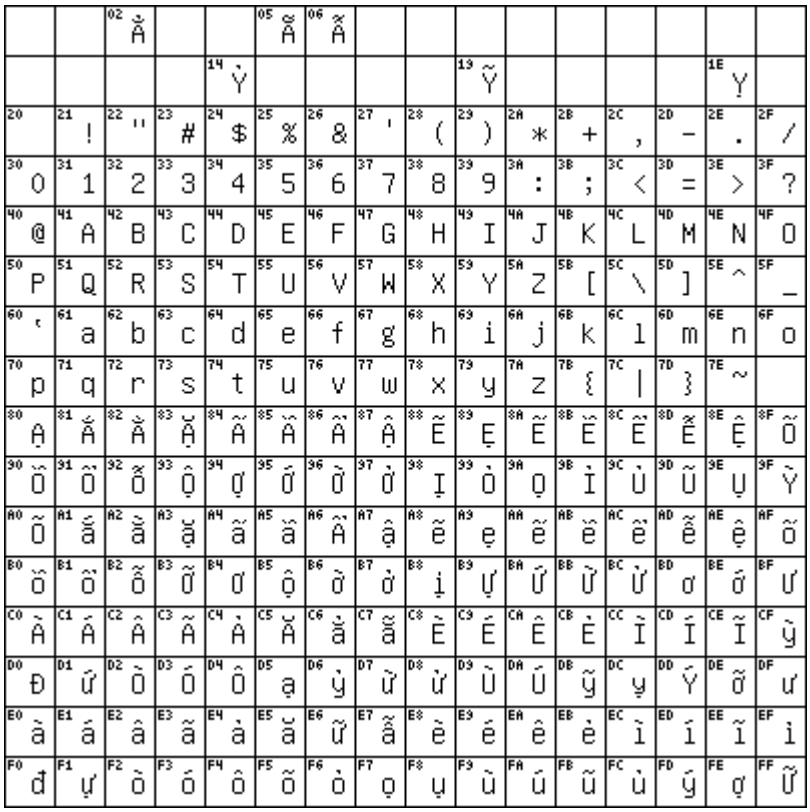
3. les autres normes sur 8 bits

La famille des normes ISO 8859-N n'est pas la seule à coder les systèmes d'écriture sur 8 bits. Puisque ces normes partaient du principe de codage du bloc xA0 à xFF, des langues comme le vietnamien n'y trouvaient pas leur compte (l'alphabet latin adapté au vietnamien utilise un très grand nombre de signes diacritiques).

Les normes sur 8 bits sont extrêmement populaires dans l'industrie de l'informatique car l'octet est une valeur sûre. Malgré les inconvénients dantesques pour tous les gens qui travaillent dans d'autres langues que l'anglais (ça fait quand même du monde ; il y a un milliard de Chinois, ne l'oublions pas), l'excuse selon laquelle l'adoption d'une norme universelle sur 16 bist (Unicode) accroîtrait la taille des fichiers bloque pour l'instant tout progrès significatif.

3.1 - la norme VISCII

Cette norme est en gros la réponse vietnamienne à la norme ascii. Cependant, elle n'est reconnue par aucun fureteur, et il faut utiliser des polices particulières pour coder le vietnamien en HTML à partir des classiques ISO 8859-N.



Il existe également des systèmes plus compliqués pour le vietnamien sur plus de 8 bits.

3.2 - la norme ISO 6429

Ceux qui ont lu attentivement le paragraphe 2 auront remarqué un "trou" entre l'us-ascii et les diverses normes ISO 8859-N.
Ce trou (x80 à x9F) a été rempli par la norme ISO 6249 qui l'a utilisé pour coder des fonctions comme Carriage Return, Form Feed, Backspace, etc. J'imagine que tout cela avait pour but, à l'époque (1991),

de contrôler le comportement des imprimantes matricielles, des telex, et de toute cette sorte d'engins très bruyants.
Il paraît que cette norme est encore utilisée.

3.3 - les jeux étendus Windows

Le fameux "trou" x80 à x9F a également été exploité par Microsoft pour étendre le jeu de caractères utilisé par Windows. Ce jeu étendu apparaît, par exemple, au niveau des polices True Type.
C'est ainsi qu'on a pu écrire *œ* en français dans un document Word ou Excel alors que cette lettre ligaturée avait été "oubliée" par la norme ISO 8859-1.
Selon les pays de localisation, les caractères x80 à x9F ne sont pas les mêmes.

3.3.1 - le jeu de caractères CP1252 ou WinLatin 1

Il s'agit de la norme ISO 8859-1 avec les caractères étendus x80 à x9F.

80 €		82 ‚	83 ƒ	84 „	85 …	86 †	87 ‡	88 ˆ	89 ‰	8A Š	8B ‹	8C Œ		8E Ž	
	91 ‘	92 ’	93 “	94 ”	95 •	96 –	97 —	98 ˜	99 ™	9A š	9B ›	9C œ		9E ž	9F Ÿ
A0	A1 ¡	A2 ¢	A3 £	A4 ¤	A5 ¥	A6 ¦	A7 §	A8 ¨	A9 ©	AA ª	AB «	AC ¬	AD	AE ®	AF ¯
B0 °	B1 ±	B2 ²	B3 ³	B4 ´	B5 µ	B6 ¶	B7 ·	B8 ¸	B9 ¹	BA º	BB »	BC ¼	BD ½	BE ¾	BF ¿
C0 À	C1 Á	C2 Â	C3 Ã	C4 Ä	C5 Å	C6 Æ	C7 Ç	C8 È	C9 É	CA Ê	CB Ë	CC Ì	CD Í	CE Î	CF Ï
D0 Ð	D1 Ñ	D2 Ò	D3 Ó	D4 Ô	D5 Õ	D6 Ö	D7 ×	D8 Ø	D9 Ù	DA Ú	DB Û	DC Ü	DD Ý	DE Þ	DF ß
E0 à	E1 á	E2 â	E3 ã	E4 ä	E5 å	E6 æ	E7 ç	E8 è	E9 é	EA ê	EB ë	EC ì	ED í	EE î	EF ï
F0 ð	F1 ñ	F2 ò	F3 ó	F4 ô	F5 õ	F6 ö	F7 ÷	F8 ø	F9 ù	FA ú	FB û	FC ü	FD ý	FE þ	FF ÿ

C'est le jeu que nous utilisons avec nos PC français sous Windows NT. Cependant il vaut mieux s'en tenir, pour une page web par exemple, au jeu ISO 8859-1. En effet, si Netscape sous Windows reconnaît ces caractères étendus, Netscape sous Unix affiche de faux caractères.

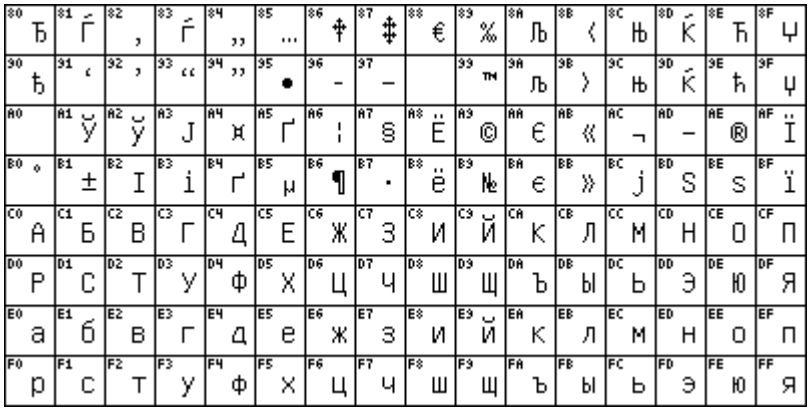
3.3.2 - le jeu de caractères CP1250 ou WinLatin 2

Il s'agit de la norme ISO 8859-2 avec les caractères étendus x80 à x9F.

80 €		82 ‚		84 „	85 …	86 †	87 ‡		89 ‰	8A Š	8B ‹	8C Ś	8D Ť	8E Ž	8F Ź
	91 ‘	92 ’	93 “	94 ”	95 •	96 –	97 —		99 ™	9A š	9B ›	9C ś	9D ť	9E ž	9F ź
A0	A1 ˇ	A2 ˘	A3 Ł	A4 ¤	A5 Ą	A6 ¦	A7 §	A8 ¨	A9 ©	AA Ş	AB «		AD	AE ®	AF Ż
B0 °	B1 ±	B2 ˛	B3 ł	B4 ´	B5 µ	B6 ¶	B7 ·	B8 ¸	B9 ą	BA ş	BB »	BC Ľ	BD ˝	BE ľ	BF ż
C0 Ŕ	C1 Á	C2 Â	C3 Ă	C4 Ä	C5 Ĺ	C6 Ć	C7 Ç	C8 Č	C9 É	CA Ę	CB Ë	CC Ě	CD Í	CE Î	CF Ď
D0 Đ	D1 Ń	D2 Ň	D3 Ó	D4 Ô	D5 Ő	D6 Ö	D7 ×	D8 Ř	D9 Ů	DA Ú	DB Ű	DC Ü	DD Ý	DE Ţ	DF ß
E0 ŕ	E1 á	E2 â	E3 ă	E4 ä	E5 ĺ	E6 ć	E7 ç	E8 č	E9 é	EA ę	EB ë	EC ě	ED í	EE î	EF ď
F0 đ	F1 ń	F2 ň	F3 ó	F4 ô	F5 ő	F6 ö	F7 ÷	F8 ř	F9 ů	FA ú	FB ű	FC ü	FD ý	FE ţ	FF ˙

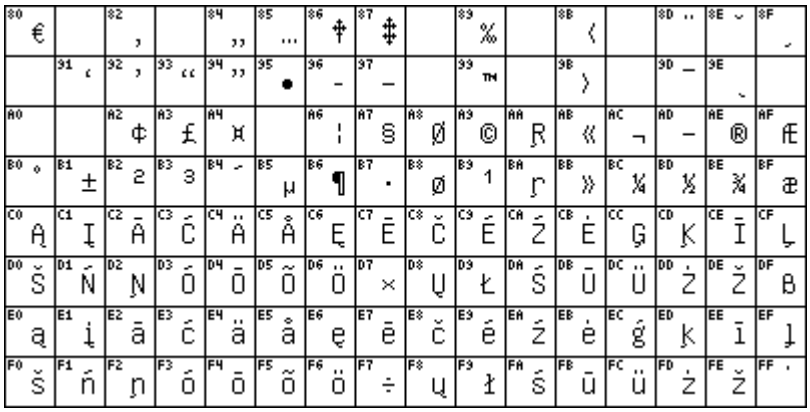
3.3.3 - le jeu de caractères CP1251 ou WinCyrillic

Nous eussions aimé qu'il s'agît de la norme ISO 8859-5 avec simplement des caractères étendus, mais voilà, toutes les lettres sont décalées (l'alphabet commence en xC0 au lieu de commencer en xB0). Elle est donc totalement incompatible à la fois avec l'ISO 8859-5 et la norme soviétique KOI-8.



3.3.4 - le jeu de caractères CP1257 ou WinBaltic

Il s'agit de la norme ISO 8859-13 avec les caractères étendus x80 à x9F.



3.3.5 - et cætera

Vous l'avez compris, Microsoft couvre les mêmes alphabets que les diverses normes ISO 8859-N en exploitant les caractères supplémentaire x80-x9F. Nous avons ainsi CP1253 ou WinGreek, CP1254 ou WinTurkish, CP1255 ou WinHebrew qui sont compatibles avec les jeux ISO 8859-N correspondants, et CP1256 ou WinArabic qui ne l'est pas (à l'instar de WinCyrillic).
Il existe également le jeu CP1258 ou WinVietnamese qui n'est pas compatible avec le viscii, mais qui l'est avec l'ascii.

3.4 - la norme ISIRI 3342

Cette norme code l'écriture arabe modifiée employée pour écrire le persan. Il s'agit d'une norme nationale iranienne, et non d'une norme ISO.

4. les algorithmes UTF

UTF signifie Unicode Transformation Format. Les algorithmes UTF décrivent des ruses pour coder Unicode (qui est sur 2 octets) en utilisant, autant que possible, un seul octet.
Le système (je simplifie) est le suivant : 6 bits sont utilisés pour coder normalement les lettres de l'alphabet latin et les deux bits restants servent à indiquer, pour des caractères exotiques, le nombre de bits parmi ce qui suit utilisés pour leur codage.

4.1 - UTF-1

C'est le plus simple des formats UTF. Son inconvénient est que, en cas de rupture du flux de données, on ne sait pas retrouver où on en était.

4.2 - UTF-2 rebaptisé plus tard UTF-8

Ce format utilise un nombre de bits standardisé par rapport à l'UTF-1 (toujours un multiple de 8), ce qui permet de reprendre un codage interrompu.

L'intérêt de l'UTF-8 est sa capacité à préserver les standards existants comme l'ASCII à 7 bits.
Sa principale caractéristique est d'utiliser une méthode de codage variable, entre 1 et 6 octets. Les premiers octets constituent l'image de l'ensemble US-ASCII (7 bit), i.e. x00-x7F. Le premier octet du caractère (il peut y en avoir jusqu'à 6 !) définit la **séquence** à laquelle il appartient. Si cet octet commence par 0, il définit alors le codage US-ASCII et sa longueur est 1. Si les n (n > 1) premiers bits sont égaux à 1, le n+1 étant égal à 0, alors la longueur du caractère sera n. Les bits restants définissent la séquence de codage.

Normalement, tous les outils X11 et GNU développés après le 1er janvier 1999 doivent intégrer le format UTF-8.
Les désavantages principaux de ce format sont :
- le codage ISO 8859-1 n'est pas reconnu de façon transparente, et il n'est pas pensable de passer à la moulinette l'énorme quantité de documents déjà existants.
- tous les caractères n'ont pas la même longueur en octets.

4.3 - UTF-7,5

Ce format utilise sept bits au lieu de six pour coder l'alphabet latin et permet la transparence par rapport à l'ISO 8859-1. L'inconvénient qui en résulte est la plus grande taille en octets nécessaire au codage des alphabets non-latins.
Ce format n'est pas reconnu comme étant une norme par l'ISO.

4.4 - UTF-7

Ce format reprend la philosophie de l'UTF-8 mais n'utilise pas le 8ème bit pour être compatible avec certains mailers. Il semble être quasiment aussi efficace que l'UTF-8.

4.5 - SCSU

Ce format (Standard Compression Scheme for Unicode) reprend la philosophie de l'UTF-8 mais permet la transparence par rapport à l'ISO 8859-1. Il nécessite des algorithmes très lourds.

5. les normes sur plus de 8 bits

Ces normes sont nées en Extrême-Orient, où les utilisateurs des *milliers* d'idéogrammes nécessaires à l'écriture des langues du groupe CJK ont bien ri quand ils ont vu que 2^8 ça ne faisait que 256.
Exit donc les ISO 8859-N, et place à des normes spécifiques à chaque pays et à son système d'écriture.

5.1 - la norme JIS X 0208

Comme d'habitude, les Japonais furent les plus rapides. La norme JIS X 0208, de 1976, est la toute première norme sur deux octets. Il s'agit d'une norme nationale japonaise, et non d'une norme ISO. Elle code les alphabets latin, cyrillique et grec, les syllabaires japonais (**kana**) et un grand nombre d'idéogrammes chinois utilisés en japonais (**kanji**).
Cette norme aurait pu, en fait, servir de base à la norme plus complète Unicode. Cependant, pour des raisons de compatibilité avec des imprimantes, seuls 8836 caractères peuvent être utilisés.

5.2 - la norme GB 2312 ou Guo-Biao

Cette norme code les idéogrammes chinois simplifiés en usage en RPC et à Singapour. Il s'agit d'une norme nationale de la RPC.
La première mouture de cette norme codait 6763 idéogrammes ; la mouture actuelle code plus de 26000 idéogrammes.

5.3 - le protocole HZ (HanZi)

Il s'agit d'une adaptation de la norme précédente aux mailers. Elle n'utilise donc que deux fois sept bits.

5.4 - la norme CNS 11643

Cette norme code les idéogrammes chinois traditionnels en usage à Taïwan et à Hong Kong. Il s'agit d'une norme nationale taïwanaise. Cette norme code 48027 idéogrammes.

5.5 - la norme KS C 5601

Cette norme code le syllabaire coréen (**hangul**) et les idéogrammes chinois (parfois) utilisés en coréen (**hanja**).

5.6 - le format EUC

Extended Unix Code utilisé sur certaines stations de travail. Ce format regroupe en fait le JIS X 0208 japonais, le GB 2312 chinois simplifié, le CNS 11643 chinois traditionnel, et le KS C 5601 coréen.

5.7 - le codage Big5

Il ne s'agit en fait pas d'une véritable norme, mais du standard industriel imposé de facto par l'industrie taïwanaise pour le codage des idéogrammes chinois traditionnels. Il est complètement différent de la norme taïwanaise CNS 11643 et sa méthode de codage ressemble à celle de la norme GB 2312 utilisée

en RPC pour les idéogrammes simplifiés, ce qui permet de passer de l'une à l'autre avec des scripts. Ce standard a le désavantage d'utiliser deux fois 8 bits au lieu de deux fois 7 bits (problèmes de mailers). Ce standard permet de coder 13051 idéogrammes traditionnels.

5.8 - la norme Guohui

Elle est très semblable à la norme précédente et a l'avantage d'être une norme officielle taïwanaise.

5.9 - UTF-16

Ce format code tous les caractères sur 16 bits. Point.
Un calcul rapide permet de s'apercevoir qu'il permet ainsi de coder 65536 caractères, c'est-à-dire presque tous les systèmes d'écriture du monde, y compris certains dont vous n'avez jamais entendu parler comme les hiéroglyphes méroïtiques ou les runes magyares.
Ce format suit fidèlement les directives de la norme Unicode ou ISO 10646 pour ce qui est de l'ordre des caractères, aboutissant ainsi à une norme *véritablement universelle*.

Ces caractères sont groupés en blocs indivisibles de 128, pour faciliter le classement par systèmes d'écriture.

Exemple : bloc des caractères de x3480 à x34FF

5.10 - UCS-4

Ce format est le même que le précédent, mais sur 31 bits. Il peut être compris par les systèmes informatiques actuels mais n'est guère utilisé en raison de sa lourdeur (plein de 0 inutiles pour les alphabets européens).

5.11 - le jeu de caractères étendu WGL4

Il s'agit d'un jeu de caractères étendu de chez Microsoft (du même type de ceux présentés au paragraphe 3.3) mais conforme à la norme ISO 10646.
Il combine en un seul jeu tous les caractères des jeux 1250 à 1254. Certaines polices True Type pour Windows NT sont normalement WGL4 : Times New Roman, Lucida Unicode et Verdana notamment.

6. les normes descriptives ou de balisage

Les normes descriptives ne codent pas directement les caractères mais les "décrivent" au sein d'un document. HTML en est un exemple bien connu. Comme tous les systèmes ne codent pas de la même façon les guillemets, il vaut mieux remplacer, dans le code HTML d'une page web, les occurrences de

"

par

"

qui sera universellement reconnu par tous les fureteurs. Il en va de même pour d'autres signes tipographiques comme &, ©, etc.

6.1 - la norme ISO 8879

Cette norme décrit les jeux de caractères Latin 1, Latin 2, Grec et Cyrillique, mais pour les systèmes **SGML**, un langage de balisage destiné à décrire la mise en forme de documents partageables par des systèmes informatiques différents.

6.2 - la norme ERCS

Extended Reference Concrete Syntax for SGML. Cette norme décrit les jeux de caractères d'Extrême-Orient (CJK) pour les systèmes SGML.

6.3 - Canonical XML

Il existe une sous-spécification de **XML** appelée Canonical XML qui doit normalement être compatible avec l'UTF-8.

7. les polices de caractère True Type

Jusqu'à Windows 95, les polices True Type étaient codées sur 8 bits et permettaient donc d'afficher 256 caractères dans une police donnée. Reprenant la philosophie des jeux de caractères Windows, les polices True Type affichent l'us-ascii sur les 128 premiers caractères, des caractères étendus Windows de x80 à x9F, et des caractères nationaux sur les caractères suivants.

Il existe ainsi des polices correspondant aux divers jeux Windows CP1250 à 1258 et, ce qui est intéressant, des polices correspondant à des systèmes d'écriture qui n'ont rien d'ISO ou pour lesquels il n'existe pas de codage en-dehors d'Unicode : le vietnamien, les langues amérindiennes, l'alphabet phonétique international, etc.

Pour créer une page web en vietnamien, par exemple, il suffit de la coder en utilisant une police True Type vietnamienne et de la mettre à disposition sur son site pour que les internautes puissent la télécharger.

Inconvénient : cela marchera très bien sous Windows, mais les internautes sous Unix ou MacOS ne comprendront rien…

À partir de Windows NT, les polices True Type sont codées sur 16 bits et permettent donc d'afficher une partie des caractères universels de la norme ISO 10646. On peut télécharger plusieurs polices de cette sorte sur le site de Microsoft (654 caractères), et également sur le site de la société Bitstream (plus de 8500 caractères).

8. le codage d'une page web (HTML)

Lorsque votre fureteur charge une page HTML, il va l'afficher avec le codage par défaut de votre système d'exploitation (pour nous, ISO 8859-1). D'où le genre de problèmes décrits au paragraphe 2.2.

La solution à ce problème est d'ajouter une ligne forçant le codage des caractères dans la partie comprise entre les tags `<HEAD>` et `</HEAD>` du code HTML de la page.

Exemples :
`<meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1">` pour forcer l'affichage en ISO 8859-1.
`<meta http-equiv="Content-Type" content="text/html; charset=windows-1250">` pour forcer l'affichage en WinLatin 2.
`<meta http-equiv="Content-Type" content="text/html; charset=KOI8-R">` pour forcer l'affichage en cyrillique (norme russe KOI8-R).
`<meta http-equiv="Content-Type" content="text/html; charset=iso-2022-kr">` pour forcer le codage d'après la norme coréenne KS C 5601.
`<meta http-equiv="content-type" content="text/html; charset=x-UNICODE-2-0-UTF-7">` pour forcer l'affichage en Unicode avec l'algorithme de transformation UTF-7.

Il est également possible de choisir, à la main, l'affichage correspondant à la page chargée.

Sous Internet Explorer, il faut cliquer sur le menu *Affichage*, puis sur l'item *Codage*.
Ma version (5.00) propose une quantité impressionnante de codages, à savoir : ISO 8859-1, 2, 5, 6, 7, 8, 9, 13 ; CP1250 à 1258 ; Chinois simplifié, Chinois traditionnel, Japonais, Coréen, Thaï (sans plus de détails) ; KOI8-R et U ; les anciens DOS Europe Centrale, arabe, cyrillique, hébreu ; ASMO 708 ; UTF-8.
Remarque : d'après mes sources, l'ISO 8859-6 et l'ASMO 708 sont la même chose, mais Internet Explorer propose quand même les deux !

Sous Netscape, il faut cliquer sur le menu *Afficher* et sur l'item *Encodage*.
Ma version (4.5) propose les codages suivants : ISO 8859-1, 2, 5, 7, 9 ; CP1250, 1251, 1253 ; Japonais ou Coréen Auto-Detect (qu'es aquò?), Shift_JIS et EUC-JP (deux implémentations légèrement différentes de JIS X 0208), Big5, EUC-TW (càd CNS 11643), GB 2312 ; KOI8-R ; UTF-7, UTF-8.

Le problème principal est que le codage ainsi défini s'applique à toute la page, d'où l'impossibilité d'afficher sur la même page un paragraphe en français et un en tchèque, par exemple, à moins de coder la page en Unicode, ce qui est beaucoup plus fastidieux, car au lieu de taper directement le caractère avec un éditeur de texte, il faut le décrire de la façon suivante

&#*D*;

où *D* représente le code (unique) du caractère dans la norme ISO 10646, en notation décimale.
Exemples :
`&#257;` pour ā
`&#328;` pour ŋ
`&#945;` pour α
`&#12354;` pour le **hiragana** japonais prononcé "a"

Annexe 1 - Codage des systèmes d'écriture pour les langues europeennes et les principales autres langues du monde

Langue	ISO 639-1	ISO 639-2/T	ISO 639-2/B	ISO 15924	codage (*)
albanais	sq	sqi	alb	la	ISO 8859-1, ISO 8859-2
allemand	de	deu	ger	la	ISO 8859-1
anglais	en	eng	eng	la	ISO 8859-1
arménien	hy	hye	arm	hy	UTF-7, UTF-8
azéri (†)	az	aze	aze	la	ISO 8859-9 (sauf ə), UTF-7, UTF-8
basque	eu	eus	baq	la	ISO 8859-1
biélorusse	be	bel	bel	cy	ISO 8859-5
breton	br	bre	bre	la	ISO 8859-1
bulgare	bg	bul	bul	cy	ISO 8859-5
catalan	ca	cat	cat	la	ISO 8859-1
corse	co	cos	cos	la	ISO 8859-1
croate	hr	hrv	scr	la	ISO 8859-2
danois	da	dan	dan	la	ISO 8859-1
erse	gd	gdh	gae	la	ISO 8859-1
espagnol	es	esp	spa	la	ISO 8859-1
estonien	et	est	est	la	ISO 8859-4, ISO 8859-10
féroïen	fo	fao	fao	la	ISO 8859-1
finnois	fi	fin	fin	la	ISO 8859-1
français	fr	fra	fre	la	ISO 8859-1 (sauf œ), ISO 8859-15
frison	fy	fry	fry	la	ISO 8859-1
galicien	gl	glg	glg	la	ISO 8859-1
gallois	cy	cym	wel	la	ISO 8859-1 (sauf w accentués), ISO 8859-14
géorgien	ka	kat	geo	ka	UTF-7, UTF-8
grec	el	ell	gre	el	ISO 8859-7
hongrois	hu	hun	hun	la	ISO 8859-2
irlandais	ga	gai	iri	la, lg	ISO 8859-1
islandais	is	isl	ice	la	ISO 8859-1
italien	it	ita	ita	la	ISO 8859-1
letton	lv	lav	lav	la	ISO 8859-4
lituanien	lt	lit	lit	la	ISO 8859-4, ISO 8859-10
luxembourgeois	lb	ltz	ltz	la	ISO 8859-1
macédonien	mk	mkd	mac	cy	ISO 8859-5
maltais	mt	mlt	mlt	la	ISO 8859-3
néerlandais	nl	nld	dut	la	ISO 8859-1
occitan	oc	oci	oci	la	ISO 8859-1
polonais	pl	pol	pol	la	ISO 8859-2
portugais	pt	por	por	la	ISO 8859-1
rhéto-roman	rm	roh	roh	la	ISO 8859-1
roumain	ro	ron	rum	la	ISO 8859-2
russe	ru	rus	rus	cy	ISO 8859-5
serbe	sr	srp	scc	cy	ISO 8859-5
slovaque	sk	slk	slo	la	ISO 8859-2
slovène	sl	slv	slv	la	ISO 8859-2
suédois	sv	swe	swe	la	ISO 8859-1
tchèque	cs	ces	cze	la	ISO 8859-2
turc	tr	tur	tur	la	ISO 8859-9
ukrainien	uk	ukr	ukr	cy	ISO 8859-5

afrikaans	af	afr	afr	la	ISO 8859-1
arabe	ar	ara	ara	ar	ISO 8859-6
bengali	bn	ben	ben	bn	UTF-7, UTF-8
esperanto	eo	epo	epo	la	ISO 8859-3
chinois (1)	zh	zho	chi	ha	GB 2312, HZ
chinois (2)	zh	zho	chi	ha	Big5, CNS 11643, Guohui
coréen (3)	ko	kor	kor	hg	KS C 5601
coréen (4)	ko	kor	kor	kh	KS C 5601
haoussa	ha	hau	hau	la	UTF-7, UTF-8
hébreu	he	heb	heb	he	ISO 8859-8
hindi	hi	hin	hin	dv	UTF-7, UTF-8
indonésien	id	ind	ind	la	ISO 8859-1
japonais	ja	jpn	jpn	ja	JIS X 0208
malais	ms	msa	may	la	ISO 8859-1
marathe	mr	mar	mar	dv	UTF-7, UTF-8
ouolof	wo	wol	wol	la	ISO 8859-1 (sauf ŋ), UTF-7, UTF-8
ourdou	ur	urd	urd	ar	UTF-7, UTF-8
persan	fa	fas	per	ar	ISIRI 3342
souahéli	sw	swa	swa	la	ISO 8859-1
tagalog	tl	tgl	tgl	la	ISO 8859-1
tamoul	ta	tam	tam	ta	UTF-7, UTF-8
télougou	te	tel	tel	te	UTF-7, UTF-8
thaï	th	tha	tha	th	ISO 8859-11
vietnamien	vi	vie	vie	la	Aucun satisfaisant (5)
xhosa	xh	xho	xho	la	ISO 8859-1
yorouba	(6)	yor	yor	la	UTF-7, UTF-8
zoulou	zu	zul	zul	la	ISO 8859-1

Remarques :
(*) il s'agit du codage par défaut ou officiel. Il est évident qu'on peut coder l'anglais d'après n'importe quelle autre norme ISO 8859-N, par exemple !
(†) la voyelle azérie ə peut être remplacée par ä
(1) RPC et Singapour
(2) Taïwan et Hong Kong
(3) Corée du Nord
(4) Corée du Sud
(5) Aucune séquence d'Unicode ne décrit le vietnamien ! Un vrai scandale, il n'y aucun moyen de le coder de façon universelle.
(6) le yorouba n'apparaît pas dans la norme ISO 639-1.

Annexe 2 - Glossaire

hangul : syllabaire phonétique pouvant noter tous les sons de la langue coréenne. En Corée du Nord, la langue est notée avec le hangul seul ; en Corée du Sud, avec un mélange de hangul et de hanja.

hanja : idéogrammes chinois utilisés en Corée du Sud pour noter les mots de façon étymologique plutôt que phonétique.

HTML : Hyper Text Mark-up Language. Code permettant d'afficher une page web de la même façon indépendamment du système d'exploitation et du fureteur utilisés (on y croit tous).

ISO : Organisation Internationale de Normalisation.

ISO 15924 : norme provisoire qui spécifie les digraphes pour les systèmes d'écriture du monde.

ISO 639-1 : norme ISO qui spécifie les digraphes pour les principales langues du monde.

ISO 639-2 : norme ISO qui spécifie les trigraphes pour la plupart des langues du monde. Malheureusement, la norme est elle-même subdivisée en **/T** (terminologie) et **/B** (bibliographie). D'autres normes (comme la norme DVB-SI) n'ont pas pris la peine de spécifier s'il fallait utiliser la colonne T ou la colonne B !

kana : syllabaires phonétiques pouvant noter tous les sons de la langue japonaise. Il en existe deux : le *hiragana* pour les mots japonais, et le *katakana* pour les emprunts aux langues étrangères. Chaque série compte presque 70 caractères.

kanji : idéogrammes chinois utilisés au Japon pour noter les mots de façon étymologique plutôt que phonétique. Les réformes des années 50 et 70 ont restreint le nombre de kanji essentiels à 2000 :
881 idéogrammes de base (kyôiku kanji)
941 idéogrammes d'usage courant (tôyô kanji)
En tout 1822, donc, plus 188 autres plus rares utilisés par exemple pour écrire les noms de famille.

KOI : Code pour l'échange de données. Norme soviétique pour le codage de l'alphabet cyrillique.

séquence : les 65536 caractères potentiels d'Unicode sont divisés en 256 séquences (ou *rangées*) correspondant grosso modo chacune à un ensemble de systèmes d'écriture similaires.
La séquence x00 correspond à l'ISO 8859-1, la séquence x04 correspond aux alphabets cyrilliques, les idéogrammes couvrant, eux, plusieurs séquences !

SGML : Standard Generalised Mark-up Language. Le langage de référence pour les documents en ligne. Le HTML est une sous-partie du SGML.

UCS : le standard Unicode est basé sur le système de codage défini dans la norme ISO 10646, qui elle-même définit deux méthodes de codage à 16 et 31 bits, respectivement : 10646.UCS-2 (16 bits) et 10646.UCS-4 (31 bits). UCS signifie Universal Character Set (ensemble universel de caractères).

Unicode : la norme qui permet de coder tous les systèmes d'écriture du monde. Actuellement, elle répertorie officiellement 38885 caractères issus de 25 systèmes d'écriture différents.

WG2 : Working Group 2 - Les gens de l'ISO qui travaillent sur la norme ISO 10646.

WG3 : Working Group 3 - Les gens de l'ISO qui travaillent sur les normes à 7 et à 8 bits. La norme ISO 2375 donne la liste des normes de codage reconnues (une norme de normes… ça devient puissant).

XML : Extensible Mark-up Language. Version simplifiée du SGML destinée au world wide web. On peut trouver une FAQ à l'URL suivante : http://www.ucc.ie/xml/

Annexe 3 - Sources

J'ai surtout mis à contribution les deux sites suivants pour le contenu technique de ce document :

www.czyborra.com

et

http://www.gms.lu/~fr_jap/unic_std.htm