Le codage des caractères

Le codage des caractères a fait l'objet de multiples normes. En 1963 un premier code (à 7 bits) de représentation des caractères était défini aux Etats-Unis et connu sous le nom de code ASCII[1]. Les organisations internationales de normalisation (comme l'ISO[2]) qui se préoccupaient de définir un code universel accepté par toutes les machines et assurant la compatibilité des supports et la possibilité des échanges, devaient en publier une version améliorée connue sous le nom de code ISO-8859. Ce code a connu des variantes selon les pays dont, pour la France, le code 8859-1 (appelé Latin-1) puis le code 8859-15 (appelé Latin-9) qui intègre de nouveaux symboles. De plus, certains codes non utilisés par l'ISO-8859 connaissent des variantes selon le système d'exploitation (par exemple Windows-1252).

Enfin en 1992, la norme la norme ISO/CEI 10646 (appelée Unicode) définit plus d'un million de caractères ainsi que des représentations appelées UTF[3] (UTF-8, UTF-16 et UTF-32) qui utilisent des codes de 1 à 4 octets. Le code ASCII standard correspond aux codes sur un seul octet de l'UTF8 (dont le premier bit est toujours 0), il est donné par la table suivante :

UTF8 (codes sur 1 octet)				b6	0	0	0	0	1	1	1	1
				b5	0	0	1	1	0	0	1	1
				b4	0	1	0	1	0	1	0	1
b3	b2	b1	b0									
0	0	0	0		NUL	DLE	SP	0	@	P	`	p
0	0	0	1		SOH	DC1	!	1	A	Q	a	q
0	0	1	0		STX	DC2	"	2	B	R	b	r
0	0	1	1		ETX	DC3	#	3	C	S	c	s
0	1	0	0		EOT	DC4	$	4	D	T	d	t
0	1	0	1		ENQ	NAK	%	5	E	U	e	u
0	1	1	0		ACK	SYN	&	6	F	V	f	v
0	1	1	1		BEL	ETB	'	7	G	W	g	w
1	0	0	0		BS	CAN	(	8	H	X	h	x
1	0	0	1		TAB	EM	)	9	I	Y	i	y
1	0	1	0		LF	SUB	*	:	J	Z	j	z
1	0	1	1		VT	ESC	+	;	K	[	k	{
1	1	0	0		FF	IS4	,	<	L	\	l	|
1	1	0	1		CR	IS3	-	=	M	]	m	}
1	1	1	0		SO	IS2	.	>	N	^	n	~
1	1	1	1		SI	IS1	/	?	O	_	o	DEL

SP désigne l'espace.

[1] American Standard Code for Information Interchange
[2] International Organization for Standardization
[3] Unicode Transformation Format

Les caractères accentués (ainsi que les signes de multiplication et de division) sont codés sur 2 octets dont le premier est 11000011 et le second (10xxxxxx) est donné par la table suivante :

2ème octet				b7	1	1	1	1
				b6	0	0	0	0
				b4	0	0	1	1
				b4	0	1	0	1
b3	b2	b1	b0					
0	0	0	0		À	Đ	à	ð
0	0	0	1		Á	Ñ	á	ñ
0	0	1	0		Â	Ò	â	ò
0	0	1	1		Ã	Ó	ã	ó
0	1	0	0		Ä	Ô	ä	ô
0	1	0	1		Å	Õ	å	õ
0	1	1	0		Æ	Ö	æ	ö
0	1	1	1		Ç	×	ç	÷
1	0	0	0		È	Ø	è	ø
1	0	0	1		É	Ù	é	ù
1	0	1	0		Ê	Ú	ê	ú
1	0	1	1		Ë	Û	ë	û
1	1	0	0		Ì	Ü	ì	ü
1	1	0	1		Í	Ý	í	ý
1	1	1	0		Î	Þ	î	þ
1	1	1	1		Ï	ß	ï	ÿ