Chap 4. ASCII

Livre p 261 – Chap 21 Représentation des textes

1. American Standard Code for Information Interchange : ASCII

a) Historique

Différentes normes ont été inventées pour communiquer à distance avec uniquement deux caractères :
Le morse en 1844, le code Baudot (ou Murray code) en 1874, le code ASCII en 1963 puis normalisé ANSI (American National Standards Institute) en 1986.
Le code ASCII est codé sur 7 bits (car on utilisait un bit pour le contrôle d'erreurs), il permet donc de coder 128 caractères différents :

b) Table ASCII standard

de 0 à 31 : caractères de contrôle
de 48 à 57 : chiffres
de 65 à 90 : lettres majuscules
de 97 à 122 : lettres minuscules
Astuce : On peut accéder directement à un caractère en appuyant sur Alt puis son code ASCII en décimal dans les logiciels…
Exemple : Le texte « Hello world ! »
Sera codé en hexadécimal dans la machine : « 48 65 6C 6C 6F 20 77 6F 72 6C 64 20 21 »

Dec	Hex	Char	Dec	Hex	Char	Dec	Hex	Char	Dec	Hex	Char
0	00	Null	32	20	Space	64	40	@	96	60	`
1	01	Start of heading	33	21	!	65	41	A	97	61	a
2	02	Start of text	34	22	"	66	42	B	98	62	b
3	03	End of text	35	23	#	67	43	C	99	63	c
4	04	End of transmit	36	24	$	68	44	D	100	64	d
5	05	Enquiry	37	25	%	69	45	E	101	65	e
6	06	Acknowledge	38	26	&	70	46	F	102	66	f
7	07	Audible bell	39	27	'	71	47	G	103	67	g
8	08	Backspace	40	28	(	72	48	H	104	68	h
9	09	Horizontal tab	41	29	)	73	49	I	105	69	i
10	0A	Line feed	42	2A	*	74	4A	J	106	6A	j
11	0B	Vertical tab	43	2B	+	75	4B	K	107	6B	k
12	0C	Form feed	44	2C	,	76	4C	L	108	6C	l
13	0D	Carriage return	45	2D	-	77	4D	M	109	6D	m
14	0E	Shift out	46	2E	.	78	4E	N	110	6E	n
15	0F	Shift in	47	2F	/	79	4F	O	111	6F	o
16	10	Data link escape	48	30	0	80	50	P	112	70	p
17	11	Device control 1	49	31	1	81	51	Q	113	71	q
18	12	Device control 2	50	32	2	82	52	R	114	72	r
19	13	Device control 3	51	33	3	83	53	S	115	73	s
20	14	Device control 4	52	34	4	84	54	T	116	74	t
21	15	Neg. acknowledge	53	35	5	85	55	U	117	75	u
22	16	Synchronous idle	54	36	6	86	56	V	118	76	v
23	17	End trans. block	55	37	7	87	57	W	119	77	w
24	18	Cancel	56	38	8	88	58	X	120	78	x
25	19	End of medium	57	39	9	89	59	Y	121	79	y
26	1A	Substitution	58	3A	:	90	5A	Z	122	7A	z
27	1B	Escape	59	3B	;	91	5B	[	123	7B	{
28	1C	File separator	60	3C	<	92	5C	\	124	7C	|
29	1D	Group separator	61	3D	=	93	5D	]	125	7D	}
30	1E	Record separator	62	3E	>	94	5E	^	126	7E	~
31	1F	Unit separator	63	3F	?	95	5F	_	127	7F	□

c) ASCII étendu

Le code ASCII standard sur 7 bits ne comporte aucun caractère accentué et peu de caractères spéciaux, aussi il a été étendu sur 8 bits pour coder 128 caractères supplémentaires.
Problème, le code ASCII étendu possède de nombreuses variantes suivant le pays ou le système sur lequel il est utilisé, les plus courants étaient OEM au début, puis ANSI. (Table OEM et ANSI au dos)
Exemple : Le texte « Élève » sera codé « C9 6C E8 76 65 » en ANSI, mais « 90 6C 8A 76 65 » en OEM.

Mais il existe beaucoup d'autres versions, elles sont répertoriées dans la norme ISO 8859
(ISO 8859-1 : latin-1 pour l'Europe occidentale par exemple, est la plus utilisée)

2. Unicode

a. Unicode

L'Unicode est système de codage compatible avec tous les pays et tous les systèmes, il est mis à jour régulièrement (554 nouveaux caractères ont été ajoutés en 2019) et répertorie actuellement plus de 100 000 caractères différents (dont les emojis !) …
Exemples : U+0065 représente « e », U+00E9 représente « é », U+03B5 représente « ε ».

b. Universal character set Transformation Format : UTF-8

La norme la plus répandue pour l'encodage des caractères en Unicode est l'UTF-8.
(Utilisé sur 93,8 % du web d'après les chiffres d'août 2019 sur https://w3techs.com/)
Cette norme est compatible avec l'ASCII standard mais n'utilise pas la même taille d'encodage suivant les caractères à coder : de 1 à 4 octets pour coder un caractère.

taille	valeurs autorisées	significatifs
1 octet	**0**xxxxxxx	7 bits
2 octets	**110**xxxxx **10**xxxxxx	11 bits
3 octets	**1110**xxxx **10**xxxxxx **10**xxxxxx	16 bits
4 octets	**11110**xxx **10**xxxxxx **10**xxxxxx **10**xxxxxx	21 bits

Exemples : La lettre « o » est codée sur 1 octet par « 6F », c'est-à-dire « **0**1101111 » ;
la lettre « ô » est codée sur 2 octets par « C3 B4 », c'est-à-dire « **110**00011 **10**110100 »
l'idéogramme « 宗 » est codée par « E5 AE 97 », c'est-à-dire « **1110**0101 **10**101110 **10**010111 »

c. UTF-16

L'UTF-16 qui est une variante utilisée par Windows qui utilise des mots de 16 bits et non plus un octet.

Table ASCII étendu OEM et ANSI :

DEC	HEX	CHAR OEM	CHAR ANSI	DEC	HEX	CHAR OEM	CHAR ANSI	DEC	HEX	CHAR OEM	CHAR ANSI	DEC	HEX	CHAR OEM	CHAR ANSI
128	80	Ç	€	160	A0	á		192	C0	└	À	224	E0	α	à
129	81	ü		161	A1	í	¡	193	C1	┴	Á	225	E1	ß	á
130	82	é	‚	162	A2	ó	¢	194	C2	┬	Â	226	E2	Γ	â
131	83	â	*ƒ*	163	A3	ú	£	195	C3	├	Ã	227	E3	∏	ã
132	84	ä	„	164	A4	ñ	¤	196	C4	─	Ä	228	E4	∑	ä
133	85	à	…	165	A5	Ñ	¥	197	C5	┼	Å	229	E5	σ	å
134	86	å	†	166	A6	ª	¦	198	C6	╞	Æ	230	E6	µ	æ
135	87	ç	‡	167	A7	º	§	199	C7	╟	Ç	231	E7	τ	ç
136	88	ê	ˆ	168	A8	¿	¨	200	C8	╚	È	232	E8	Φ	è
137	89	ë	‰	169	A9	⌐	©	201	C9	╔	É	233	E9	θ	é
138	8A	è	Š	170	AA	¬	ª	202	CA	╩	Ê	234	EA	Ω	ê
139	8B	ï	‹	171	AB	½	«	203	CB	╦	Ë	235	EB	δ	ë
140	8C	î	Œ	172	AC	¼	¬	204	CC	╠	Ì	236	EC	∞	ì
141	8D	ì		173	AD	¡	-	205	CD	═	Í	237	ED	Ø	í
142	8E	Ä	Ž	174	AE	«	®	206	CE	╬	Î	238	EE	∈	î
143	8F	Å		175	AF	»	¯	207	CF	╧	Ï	239	EF	∩	ï
144	90	É		176	B0	░	°	208	D0	╨	Ð	240	F0	≡	ð
145	91	æ	‘	177	B1	▒	±	209	D1	╤	Ñ	241	F1	±	ñ
146	92	Æ	’	178	B2	▓	²	210	D2	╥	Ò	242	F2	≥	ò
147	93	ô	“	179	B3	│	³	211	D3	╙	Ó	243	F3	≤	ó
148	94	ö	”	180	B4	┤	´	212	D4	╘	Ô	244	F4	⌠	ô
149	95	ò	•	181	B5	╡	µ	213	D5	╒	Õ	245	F5	⌡	õ
150	96	û	–	182	B6	╢	¶	214	D6	╓	Ö	246	F6	÷	ö
151	97	ù	—	183	B7	╖	·	215	D7	╫	×	247	F7	≈	÷
152	98	ÿ	˜	184	B8	╕	¸	216	D8	╪	Ø	248	F8	°	ø
153	99	Ö	™	185	B9	╣	¹	217	D9	┘	Ù	249	F9	¨	ù
154	9A	Ü	Š	186	BA	║	º	218	DA	┌	Ú	250	FA	·	ú
155	9B	ø	›	187	BB	╗	»	219	DB	▉	Û	251	FB	√	û
156	9C	£	œ	188	BC	╝	¼	220	DC	▄	Ü	252	FC	³	ü
157	9D	Ø		189	BD	╜	½	221	DD	▌	Ý	253	FD	²	ý
158	9E	×	ž	190	BE	╛	¾	222	DE	▐	Þ	254	FE	■	þ
159	9F	*ƒ*	Ÿ	191	BF	┐	¿	223	DF	▀	ß	255	FF		ÿ