

1 Intro

Comme pour les nombres, les caractères doivent être traduits par chaque machine en une suite de 0 et de 1. L'encodage des caractères a subi plusieurs évolutions majeures.

À l'origine, c'était le « grand bazar » ! En changeant de type d'ordinateur ou de système d'exploitation, les textes ne pouvaient pas être lus correctement car les systèmes d'encodages étaient tous différents.

2 ASCII

En 1960, la norme **ASCII** (American Standard Code for Information Interchange) est créée. Elle définit un encodage des caractères sur 7 bits, permettant ainsi de définir $2^7 = 128$ caractères différents.

Cela était bien suffisant pour les caractères de langue anglo-saxonne et quelques signes de ponctuation (plus quelques encodages particuliers comme le retour à la ligne par exemple).

Decimal	Hexadecimal	Binary	Octal	Char	Decimal	Hexadecimal	Binary	Octal	Char	Decimal	Hexadecimal	Binary	Octal	Char
0	0	0	0	[NULL]	48	30	110000	60	0	96	60	1100000	140	`
1	1	1	1	[START OF HEADING]	49	31	110001	61	1	97	61	1100001	141	a
2	2	10	2	[START OF TEXT]	50	32	110010	62	2	98	62	1100010	142	b
3	3	11	3	[END OF TEXT]	51	33	110011	63	3	99	63	1100011	143	c
4	4	100	4	[END OF TRANSMISSION]	52	34	110100	64	4	100	64	1100100	144	d
5	5	101	5	[ENQUIRY]	53	35	110101	65	5	101	65	1100101	145	e
6	6	110	6	[ACKNOWLEDGE]	54	36	110110	66	6	102	66	1100110	146	f
7	7	111	7	[BELL]	55	37	110111	67	7	103	67	1100111	147	g
8	8	1000	10	[BACKSPACE]	56	38	111000	70	8	104	68	1101000	150	h
9	9	1001	11	[HORIZONTAL TAB]	57	39	111001	71	9	105	69	1101001	151	i
10	A	1010	12	[LINE FEED]	58	3A	111010	72	:	106	6A	1101010	152	j
11	B	1011	13	[VERTICAL TAB]	59	3B	111011	73	;	107	6B	1101011	153	k
12	C	1100	14	[FORM FEED]	60	3C	111100	74	<	108	6C	1101100	154	l
13	D	1101	15	[CARRIAGE RETURN]	61	3D	111101	75	=	109	6D	1101101	155	m
14	E	1110	16	[SHIFT OUT]	62	3E	111110	76	>	110	6E	1101110	156	n
15	F	1111	17	[SHIFT IN]	63	3F	111111	77	?	111	6F	1101111	157	o
16	10	10000	20	[DATA LINK ESCAPE]	64	40	1000000	100	@	112	70	1110000	160	p
17	11	10001	21	[DEVICE CONTROL 1]	65	41	1000001	101	A	113	71	1110001	161	q
18	12	10010	22	[DEVICE CONTROL 2]	66	42	1000010	102	B	114	72	1110010	162	r
19	13	10011	23	[DEVICE CONTROL 3]	67	43	1000011	103	C	115	73	1110011	163	s
20	14	10100	24	[DEVICE CONTROL 4]	68	44	1000100	104	D	116	74	1110100	164	t
21	15	10101	25	[NEGATIVE ACKNOWLEDGE]	69	45	1000101	105	E	117	75	1110101	165	u
22	16	10110	26	[SYNCHRONOUS IDLE]	70	46	1000110	106	F	118	76	1110110	166	v
23	17	10111	27	[ENG OF TRANS. BLOCK]	71	47	1000111	107	G	119	77	1110111	167	w
24	18	11000	30	[CANCEL]	72	48	1001000	110	H	120	78	1111000	170	x
25	19	11001	31	[END OF MEDIUM]	73	49	1001001	111	I	121	79	1111001	171	y
26	1A	11010	32	[SUBSTITUTE]	74	4A	1001010	112	J	122	7A	1111010	172	z
27	1B	11011	33	[ESCAPE]	75	4B	1001011	113	K	123	7B	1111011	173	{
28	1C	11100	34	[FILE SEPARATOR]	76	4C	1001100	114	L	124	7C	1111100	174	
29	1D	11101	35	[GROUP SEPARATOR]	77	4D	1001101	115	M	125	7D	1111101	175	}
30	1E	11110	36	[RECORD SEPARATOR]	78	4E	1001110	116	N	126	7E	1111110	176	~
31	1F	11111	37	[UNIT SEPARATOR]	79	4F	1001111	117	O	127	7F	1111111	177	[DEL]
32	20	100000	40	[SPACE]	80	50	1010000	120	P					
33	21	100001	41	!	81	51	1010001	121	Q					
34	22	100010	42	"	82	52	1010010	122	R					
35	23	100011	43	#	83	53	1010011	123	S					
36	24	100100	44	\$	84	54	1010100	124	T					
37	25	100101	45	%	85	55	1010101	125	U					
38	26	100110	46	&	86	56	1010110	126	V					
39	27	100111	47	'	87	57	1010111	127	W					
40	28	101000	50	(88	58	1011000	130	X					
41	29	101001	51)	89	59	1011001	131	Y					
42	2A	101010	52	*	90	5A	1011010	132	Z					
43	2B	101011	53	+	91	5B	1011011	133	[
44	2C	101100	54	,	92	5C	1011100	134	\					
45	2D	101101	55	.	93	5D	1011101	135]					
46	2E	101110	56	.	94	5E	1011110	136	^					
47	2F	101111	57	/	95	5F	1011111	137	_					

3 ASCII étendu : ISO-8859-1 (Latin-1)

Par la suite, pour augmenter le nombre de caractères à représenter, le 8ème bit de l'octet a été ajouté pour pouvoir représenter $2^8 = 256$ caractères (comme les caractères accentués par exemple). Les normes qui se sont développées sur ce modèle ont fait en sorte de conserver une compatibilité avec l'ASCII d'origine (par exemple le A majuscule reste codé 41 en hexadécimal dans ces nouveaux encodages) : cela définit des tables ascii étendues.

La norme **ISO-8859-1 (Latin-1)** est utilisée pour encoder tous les caractères de langue européenne (de l'Ouest) comme le Français.

ISO-8859-1																
	x0	x1	x2	x3	x4	x5	x6	x7	x8	x9	xA	xB	xC	xD	xE	xF
0x	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
1x	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
2x	SP	!	"	#	\$	%	&	'	()	*	+	,	-	.	/
3x	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4x	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5x	P	Q	R	S	T	U	V	W	X	Y	Z	[\]	^	_
6x	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7x	p	q	r	s	t	u	v	w	x	y	z	{		}	~	DEL
8x	PAD	HOP	BPH	NBH	IND	NEL	SSA	ESA	HTS	HTJ	VTS	PLD	PLU	RI	SS2	SS3
9x	DCS	PU1	PU2	STS	CCH	MW	SPA	EPA	SOS	SGCI	SCI	CSI	ST	OSC	PM	APC
Ax	NBSP	ı	¢	£	¤	¥	¦	§	¨	©	ª	«	¬	­	®	¯
Bx	°	±	²	³	´	µ	¶	·	¸	¹	º	»	¼	½	¾	¿
Cx	À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î	Ï
Dx	Ð	Ñ	Ò	Ó	Ô	Õ	Ö	×	Ø	Ù	Ú	Û	Ü	Ý	Þ	ß
Ex	à	á	â	ã	ä	å	æ	ç	è	é	ê	ë	ì	í	î	ï
Fx	ø	ñ	ò	ó	ô	õ	ö	÷	ø	ù	ú	û	ü	ý	þ	ÿ

4 Unicode et UTF-8

Pour unifier complètement l'ensemble des caractères existant dans toutes les langues du monde, le standard Unicode (lié à la norme ISO/CEI 10646) a été développé à partir de 1991 par le Consortium Unicode. Il vise au codage de texte écrit en donnant à tout caractère un nom et un identifiant numérique appelé **point de code**.

Le répertoire Unicode peut contenir plus d'un million de caractères, ce qui est bien trop grand pour être codé par un seul octet. La norme Unicode définit donc des méthodes standardisées pour coder cet index sous forme de séquence d'octets.

UTF-8 est l'une d'entre elles et la plus majoritairement utilisée. En UTF-8, chaque caractère est codé sous forme de séquences **de un à quatre octets**.

Les caractères ASCII restent codés sur un seul octet pour garder la comptabilité avec l'ASCII d'origine. Par exemple le A majuscule reste codé par l'octet 65 (en base 10).

Les caractères dont le point de code est supérieur à 127 (caractère non ASCII) se codent sur 2 à 4 octets. Le caractère € (euro) se code par exemple sur 3 octets : 226, 130, et 172.

5 Encodage en Python

Les fonction **ord(X)** et **chr(N)** renvoient respectivement les points de code du caractère X ou le caractères du point de code N.

Exemple :

`chr(65)` \mapsto renvoie 'A'

`chr(231)` \mapsto renvoie ç

`ord('A')` \mapsto renvoie 65

`ord('ç')` \mapsto renvoie 231

La console Python travaille en utf8 par défaut.

Il est possible de convertir une chaîne de caractères (type str) en chaîne d'octets (type bytes).

En fonction du type d'encodage, la taille de la chaîne d'octets obtenue sera variable.

Évidemment, il n'est pas possible d'encoder en mode 'ascii' une chaîne de caractères contenant des accents ou tout autre caractère non défini dans la table ASCII.

Les fonctions d'encodage ou de décodage sont simplement 'chaîne'.encode('nom-encodage') et b'chaîne'.decode('nom-encodage').

`'€'.encode('utf8')`

\hookrightarrow `b'\xe2\x82\xac'` (soit en base 10 : 226, 130, 172)

`'déjà'.encode('utf8')`

\hookrightarrow `b'd\xc3\xa9j\xc3\xa0'`

`'déjà'.encode('utf16')`

\hookrightarrow `b'\xff\xfe\x00\xe9\x00j\x00\xe0\x00'`

`'déjà'.encode('ascii')`

\hookrightarrow `ERROR`

`b'd\xc3\xa9j\xc3\xa0'.decode('utf8')`

\hookrightarrow `'déjà'`

`b'd\xc3\xa9j\xc3\xa0'.decode('utf16')`

\hookrightarrow 𐄂𐄂

6 Déclarer un encodage en programmation

En HTML5, on commencera toujours par "prévenir" le navigateur de l'encodage utilisé. Il est recommandé d'encoder les pages html en utf8.

`<meta charset='utf8 '>`

En Python, la 1ère ligne du programme permet de déclarer l'encodage sous la forme :

`# coding: utf-8`

Les bons éditeurs de texte proposent toujours de choisir l'encodage lors de l'enregistrement d'un document.