

Représentation d'un texte en machine

Nous savons qu'un ordinateur est uniquement capable de traiter des données binaires, comment sont donc codés les textes dans un ordinateur ? Ou plus précisément, comment sont codés les caractères dans un ordinateur ?

Rappel : ASCII

Avant 1960 de nombreux systèmes de codage de caractères existaient, ils étaient souvent incompatibles entre eux. En 1960, l'organisation internationale de normalisation (ISO) décide de mettre un peu d'ordre dans ce bazar en créant la norme ASCII (American Standard Code for Information Interchange). À chaque caractère est associé un nombre binaire sur 8 bits (1 octet). En fait, seuls 7 bits sont utilisés pour coder un caractère, le 8e bit n'est pas utilisé pour le codage des caractères. Avec 7 bits il est possible de coder jusqu'à 128 caractères ce qui est largement suffisant pour un texte écrit en langue anglaise (pas d'accents et autres lettres particulières).

ASCII TABLE

Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char
0	0	[NULL]	32	20	[SPACE]	64	40	@	96	60	`
1	1	[START OF HEADING]	33	21	!	65	41	A	97	61	a
2	2	[START OF TEXT]	34	22	"	66	42	B	98	62	b
3	3	[END OF TEXT]	35	23	#	67	43	C	99	63	c
4	4	[END OF TRANSMISSION]	36	24	\$	68	44	D	100	64	d
5	5	[ENQUIRY]	37	25	%	69	45	E	101	65	e
6	6	[ACKNOWLEDGE]	38	26	&	70	46	F	102	66	f
7	7	[BELL]	39	27	'	71	47	G	103	67	g
8	8	[BACKSPACE]	40	28	(72	48	H	104	68	h
9	9	[HORIZONTAL TAB]	41	29)	73	49	I	105	69	i
10	A	[LINE FEED]	42	2A	*	74	4A	J	106	6A	j
11	B	[VERTICAL TAB]	43	2B	+	75	4B	K	107	6B	k
12	C	[FORM FEED]	44	2C	,	76	4C	L	108	6C	l
13	D	[CARRIAGE RETURN]	45	2D	-	77	4D	M	109	6D	m
14	E	[SHIFT OUT]	46	2E	.	78	4E	N	110	6E	n
15	F	[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	[NEGATIVE 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	[ENG OF 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	[SUBSTITUTE]	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	[DEL]

Comme vous pouvez le constater dans le tableau ci-dessus, au "A" majuscule correspond le code binaire (100001) ((65) ou (41))

Exercice :

Quel est le code binaire du "a" minuscule en ASCII?

Comme vous pouvez le constater, certains codes ne correspondent pas à des caractères (de 0 à (32)), nous n'aborderons pas ce sujet ici.

ISO-8859-1 et ANSI

La norme ASCII convient bien à la langue anglaise, mais pose des problèmes dans d'autres langues, par exemple le français. En effet l'ASCII ne prévoit pas d'encoder les lettres accentuées. C'est pour répondre à ce problème qu'est née la norme ISO-8859-1. Cette norme reprend les mêmes principes que l'ASCII, mais les nombres binaires associés à chaque caractère sont **codés sur 8 bits**, ce qui permet d'encoder jusqu'à 256 caractères. Cette norme va être principalement utilisée dans les pays européens puisqu'elle permet d'encoder les caractères utilisés dans les principales langues européennes (la norme ISO-8859-1 est aussi appelée "**latin1**" car elle permet d'encoder les caractères de l'alphabet dit "latin")

Problème, il existe beaucoup d'autres langues dans le monde qui n'utilisent pas l'alphabet dit "latin", par exemple le chinois ou le japonais ! D'autres normes ont donc dû voir le jour, par exemple la norme "GB2312" pour le chinois simplifié ou encore la norme "JIS_X_0208" pour le japonais.

Cette multiplication des normes a très rapidement posé problème. Imaginons un français qui parle le japonais. Son traitement de texte est configuré pour reconnaître les caractères de l'alphabet "latin" (norme ISO-8859-1). Un ami japonais lui envoie un fichier texte écrit en japonais. Le français devra modifier la configuration de son traitement afin que ce dernier puisse afficher correctement l'alphabet japonais. S'il n'effectue pas ce changement de configuration, il verra s'afficher des caractères ésotériques.

Microsoft a développé une norme presque équivalente (27 caractères de plus) : ANSI

Unicode

Pour éviter ce genre de problème, en 1991 une nouvelle norme a vu le jour : Unicode

Unicode a pour ambition de rassembler tous les caractères existant afin qu'une personne utilisant Unicode puisse, sans changer la configuration de son traitement de texte, à la fois lire des textes en français ou en japonais

Unicode est uniquement une table qui regroupe tous les caractères existant au monde, il ne s'occupe pas de la façon dont les caractères sont codés dans la machine. Le standard Unicode est constitué d'un répertoire de 137 929 caractères, couvrant une centaine d'écritures

Unicode accepte plusieurs systèmes de codage : UTF-8 (8-bit Unicode Transformation Format), **UTF-16**, **UTF-32**. Le plus utilisé, notamment sur le Web, est UTF-8.

Pour encoder les caractères Unicode, UTF-8 utilise un nombre variable d'octets : les caractères "classiques" (les plus couramment utilisés) sont codés sur un octet, alors que des caractères "moins classiques" sont codés sur un nombre d'octets plus important (jusqu'à 4 octets). Un des avantages d'UTF-8 c'est qu'il est totalement compatible avec la norme ASCII : Les caractères Unicode codés avec UTF-8 ont exactement le même code que les mêmes caractères en ASCII.

Exercice :

Quel est le code binaire du "b" minuscule Unicode codé avec UTF-8 ?

Questions issues de la Banque Nationale de Sujets

Q1 - Le code ASCII permet de représenter en binaire les caractères alphanumériques. Quel est son principal inconvénient ?

Réponses :

- A- Il utilise beaucoup de bits.
- B- Il ne différencie pas les majuscules des minuscules.
- C- Il ne représente pas les caractères accentués.
- D- Il n'est pas compatible avec la plupart des systèmes informatiques.

Q2 - Quel est un avantage du codage UTF8 par rapport au codage ASCII ?

Réponses :

- A- il permet de coder un caractère sur un octet au lieu de deux
- B- il permet de coder les majuscules
- C- il permet de coder tous les caractères
- D- il permet de coder différentes polices de caractères

Q3 - On considère les codes ASCII en écriture hexadécimale (en base 16). Le code ASCII de la lettre A est 0x41, celui de la lettre B est 0x42, celui de la lettre C est 0x43, etc. Quel est le code ASCII, en hexadécimal, de la lettre X (c'est la 24e lettre de l'alphabet usuel).

Réponses :

- A- 0x58
- B- 0x64
- C- 0x7A
- D- 0x88

Q4 - Laquelle de ces affirmations concernant le codage UTF-8 des caractères est vraie ?

Réponses :

- A- le codage UTF-8 est sur 7 bits
- B- le codage UTF-8 est sur 8 bits
- C- le codage UTF-8 est sur 1 à 4 octets
- D- le codage UTF-8 est sur 8 octets

Q5 - Parmi les noms suivants, lequel **n'est pas** celui d'une méthode d'encodage des caractères ?

Réponses :

- A- UTF-16
- B- ASCII
- C- Arial
- D- Unicode

Q6 - Parmi les caractères ci-dessous, lequel ne fait pas partie du code ASCII ?

Réponses :

- A- a
- B- B
- C- @
- D- é

Q7 - Sur une page web qui s'affiche sur notre navigateur on peut lire : « En conséquence, l'Assemblée Nationale reconnaît et déclare, en présence [...] » Quelle peut être la cause des affichages étranges de cette page ?

Réponses :

- A- l'encodage des caractères n'est pas celui attendu par le navigateur
- B- le texte original est en japonais
- C- la taille des caractères n'est pas celui attendu par le navigateur
- D- la connexion à Internet présente des coupures

Q8 - Olivier visite son site préféré pour relire le programme de NSI. Il devrait lire le texte ci-dessous :

L'enseignement de spécialité de numérique et sciences informatiques du cycle terminal de la voie générale vise l'appropriation des fondements de l'informatique pour préparer les élèves à une poursuite d'études dans l'enseignement supérieur, en les formant à la pratique d'une démarche scientifique et en développant leur appétence pour des activités de recherche. Le paramétrage de son navigateur lui donne l'affichage ci-dessous :

L'enseignement de spécialité de numérique et sciences informatiques du cycle terminal de la voie générale vise l'appropriation des fondements de l'informatique pour préparer les élèves à une poursuite d'études dans l'enseignement supérieur, en les formant à la pratique d'une démarche scientifique et en développant leur appétence pour des activités de recherche.

Quel type d'encodage Olivier doit-il choisir dans les paramètres de son navigateur pour afficher correctement le texte ?

Réponses :

- A- ANSI
- B- LATIN-1
- C- UTF-8
- D- ASCII