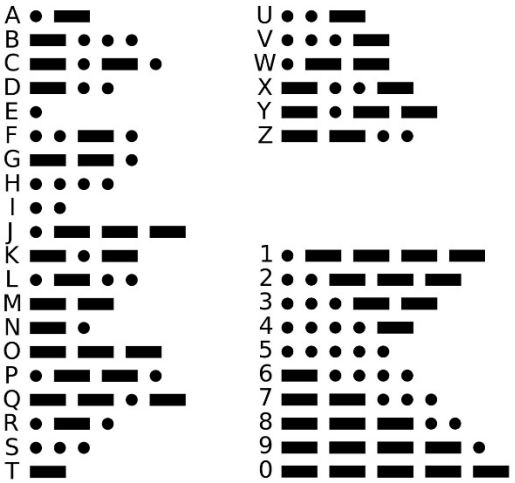

REPRESENTATION DES DONNEES :
Types et valeurs de base (3)

- *Représentation d'un texte en machine*

"Thomas Edison a enseigné le code morse à sa femme afin qu'ils puissent communiquer en secret en tapant dans les mains."

Numérique et Sciences Informatiques
1ère

Support de cours :
Jean-Christophe BONNEFOY

Objectifs :

- Identifier l'intérêt des différents systèmes d'encodage
- Convertir un fichier texte dans différents formats d'encodage

1. Un peu d'histoire

Si le caractère existe depuis environ deux millénaires, sa représentation abstraite sous forme numérique est plus récente. Elle a notamment été développée pour le télégraphe. Cette abstraction permettant daméliorer lefficience des communicatio ns. Cependant, au milieu du XX^{ème} siècle, chaque matériel (notamment les imprimantes) avait leur propre codage. Tout ordinateur était livré avec ses sous-programmes et ses tables permettant de transposer les codes d'un matériel à l'autre. Lémergence dun codage unifié sest cependant heurtée à des différences dapproche conventionnelles et culturelles du concept de caractère.

- 1690 : première expérience de transmission d'une information codée au Jardin du Luxembourg (Paris) à l'aide d'un télégraphe.
- 1832 : invention de l'alphabet Morse international par Samuel Morse.
- 1912 : Création de l'American Institute of Electrical Engineers (aujourd'hui IEEE) pour définir des standards industriels américains.
- 1926 : création de l'AFNOR, organisation française pour la normalisation des standards français.
- 1947 : création de l'ISO (Organisation Internationale de Normalisation). Cette organisation a pour but de produire des normes internationales appelées normes ISO.
- 1961 : apparition de la première version du code ASCII (American Standard Code for Information Interchange)
- 1991 : apparition du standard Unicode.

2. Représentation d'un caractère : une histoire de norme

Un fichier contient une représentation de données. Par exemple un fichier peut contenir une représentation dun texte. Bien souvent on dit plus simplement que le fichier contient le texte. Le contenu du fichier nest lui -même quune suite de 0 et de 1, des bits. On choisit donc de coder chacune des lettres, plus généralement chacun des caractères, par une représentation binaire.

Ce choix dutiliser un codage donné est arbitraire.

2.1 La norme ASCII

Le codage ASCII est une norme de codage de caractères en informatique ancienne et connue pour son influence incontournable sur les codages de caractères qui lui ont succédé. ASCII contient les caractères nécessaires pour écrire en anglais.

LASCII définit seulement 128 caractères numérotés de 0 à 127 et codés en binaire de 000 0000 à 111 1111. Sept bits suffisent donc pour représenter un caractère codé en ASCII. Pour alléger les notations, le caractère codé est représenté le plus souvent par l'équivalent en hexadécimal ou en décimal du nombre binaire associé.

Exemple : Lettre « E » : $(100\ 0101)_2 = (45)_{16} = (69)_{10}$

				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	Exemple : E = 100 0101$_{(2)}$ = 69$_{(10)}$. Appuyez sur "ALT", saisissez 69 et relâchez "ALT". Convaincu !								
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		HT	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	FS	,	<	L	\	l	|
1	1	0	1		CR	GS	-	=	M	]	m	}
1	1	1	0		SO	RS	.	>	N	^	n	~
1	1	1	1		SI	US	/	?	O	_	o	DEL

Tableau du code ASCII

- le caractère « a » s'écrit : $110\ 0001_2$ (61_h en hexadécimal)
- le code $010\ 1011_2$ correspond au caractère : « + » ($2B_h$ en hexadécimal)
- l'espace (SP) se code : $010\ 0000_2$ (20_h en hexadécimal)

Remarque : Dans la pratique, on code un caractère sur 8 bits (1 octet) avec le bit de poids fort toujours égale à 0.

2.2 La norme ISO-8859-1

Le code ASCII a été conçu pour représenter des textes écrits en anglais, il ny a donc pas daccents, de tréma … Par exemple, en français les caractères é, è, ç, à, ù, ô, æ, œ sont fréquemment utilisés alors qu'ils ne figurent pas dans la table ASCII. Il va donc falloir étendre la table ASCII pour pouvoir coder les nouveaux caractères avec un 8eme bit. Cela donne la norme ISO-8859-1 (souvent appelé Latin-1). On trouve dans cette norme quasiment tous les caractères utilisés dans la langue française. Il manque cependant le œ !

Charset ISO-8859-1 (Latin 1)

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0			32	0 48	@ 64	P 80	‘ 96	p 112			32	° 176	À 192	Ð 208	à 224	ð 240
1			! 33	1 49	A 65	Q 81	a 97	q 113			¡ 161	± 177	Á 193	Ñ 209	á 225	ñ 241
2			" 34	2 50	B 66	R 82	b 98	r 114			¢ 162	² 178	Â 194	Ò 210	â 226	ò 242
3			# 35	3 51	C 67	S 83	c 99	s 115			£ 163	³ 179	Ã 195	Ó 211	ã 227	ó 243
4			$ 36	4 52	D 68	T 84	d 100	t 116			¤ 164	´ 180	Ä 196	Ô 212	ä 228	ô 244
5			% 37	5 53	E 69	U 85	e 101	u 117			¥ 165	µ 181	Å 197	Õ 213	å 229	õ 245
6			& 38	6 54	F 70	V 86	f 102	v 118			¦ 166	¶ 182	Æ 198	Ö 214	æ 230	ö 246
7			’ 39	7 55	G 71	W 87	g 103	w 119			§ 167	· 183	Ç 199	× 215	ç 231	÷ 247
8			(40	8 56	H 72	X 88	h 104	x 120			¨ 168	¸ 184	È 200	Ø 216	è 232	ø 248
9			) 41	9 57	I 73	Y 89	i 105	y 121			© 169	¹ 185	É 201	Ù 217	é 233	ù 249
A			* 42	: 58	J 74	Z 90	j 106	z 122			ª 170	º 186	Ê 202	Ú 218	ê 234	ú 250
B			+ 43	; 59	K 75	[91	k 107	{ 123			« 171	» 187	Ë 203	Û 219	ë 235	û 251
C			, 44	< 60	L 76	\ 92	l 108	| 124			¬ 172	¼ 188	Ì 204	Ü 220	ì 236	ü 252
D			- 45	= 61	M 77	] 93	m 109	} 125			173	½ 189	Í 205	Ý 221	í 237	ý 253
E			. 46	> 62	N 78	^ 94	n 110	~ 126			® 174	¾ 190	Î 206	Þ 222	î 238	þ 254
F			/ 47	? 63	O 79	_ 95	o 111				¯ 175	¿ 191	Ï 207	ß 223	ï 239	ÿ 255

www.langbox.com - LangBox International

On remarquera que les 128 premiers caractères correspondent à l'ASCII.

- le caractère « à » s'écrit : $E0_h$, soit $1110\ 0000_2$
- le code EB_h correspond au caractère : « ë »

En France, depuis l'apparition de l'euro, c'est le codage ISO-8859-15 (souvent appelé Latin-9) qui est utilisé. C'est une sorte de mise à jour de la norme ISO-8859-1. Il permet d'utiliser tous les caractères que l'on veut dans notre langue française.

2.3 La norme Unicode

Le problème de la norme précédente se trouve dans l'échange de texte à l'échelle mondiale. 256 caractères ne suffisant pas pour représenter les lettres de tous les alphabets utilisés (pensons au russe, à lhébreu, etc.), un nouveau standard a été introduit : **Unicode**.

Il vise à donner à tout caractère de n'importe quel système d'écriture (toute langue confondue) :

- un identifiant numérique que l'on appelle **Point de code**
- un descriptif

Le point de code est noté U+xxxx, où xxxx est en hexadécimal, et comporte 4 à 6 chiffres. Il est compris entre U+0000 et U+10FFFF. Toutes les places ne sont pas occupées, on estime le taux de remplissage à environ 17%. Il reste donc encore de la place !

Point de code	Caractère	Descriptif
U+0061	a	LATIN SMALL LETTER A
U+00E9	é	LATIN SMALL LETTER E WITH ACUTE
U+03A3	Σ	GREEK CAPITAL LETTER SIGMA
U+265E	♞	BLACK CHESS KNIGHT
U+2167	Ⅷ	ROMAN NUMERAL EIGHT
U+0924	त	DEVANAGARI LETTER TA

Exemples de caractères dans la norme Unicode

Le codage de cette table est multiple. Le codage le plus couramment utilisé se nomme **UTF-8**. Son principe est le suivant :

Premier Point de code	Dernier Point de code	Nombre d'octets nécessaires pour coder le caractère	Représentation binaire	Nombre de bits maximum pour le caractère
U+0000	U+007F	1	**0**xxx xxxx	7
U+0080	U+07FF	2	**110**x xxxx **10**xx xxxx	11
U+0800	U+FFFF	3	**1110** xxxx **10**xx xxxx **10**xx xxxx	16
U+10000	U+10FFFF	4	**1111 0**xxx **10**xx xxxx **10**xx xxxx **10**xx xxxx	21

Où xxxx représente le codage binaire du point de code

Exemple :

Caractère	Point de code	Codage binaire du point de code	Représentation binaire du caractère en UTF-8
9	U+0039	011 1001	**0**011 1001
é	U+00E9	1110 1001	**110**0 0011 **10**10 1001
烈	U+F99F	1111 1001 1001 1111	**1110** 1111 **10**10 0110 **10**01 1111
𝍮	U+1031E	1 0000 0011 0001 1110	**1111 0**000 **10**01 0000 **10**00 1100 **10**01 1110