Encodage des caractères

1 Représentation des textes

La représentation des textes dans un ordinateur s'appuie sur un encodage. Une table associe un nombre à chaque caractère. Les systèmes d'exploitation possèdent ces tables et font la conversion en binaire.

L'informatique créée aux Etats-Unis, a pu se contenter 128 caractères en anglais pendant longtemps, mais ces caractères sont devenus insuffisants notamment avec les langues latines et leurs nombreux caratères accentués. Les tables se sont alors agrandies et se sont harmonisées. Mais d'autres langues, comme l'arabe, le corréen et le japonais qui utilisent des caractères très particuliers, propres à leurs langues, ont conduit à étendre encore les tables de caractères.

On peut retenir 3 tables de codage importantes :

- Le codage ASCII créé dans les années 1960.
- La norme ISO 8859-1 et 8859-15
- Le codage unicode et sa représentation en UTF-8

2 Le codage ASCII

La table de caratères ASCII (American Standard Code for Information Interchange) a été élaborée au début des années 1960.

Cette norme définit un jeu de 128 caractères représenté par un octet, mais seuls 7 bits sont utilisés d'où le nombre de caractères utilisés puisque sur 7 bits on dispose de $2^7 = 128$ valeurs différentes.

A chaque codage binaire d'un caractère est associé un nombre décimal égal à cette valeur binaire convertie.

Exemple

Voici quelques caractères de la table ASCII:

- Le caratère A, en majuscule, correspond au nombre décimal 65_{10} dans la table ASCII que l'on note 41_{16} en hexadécimal et de codage binaire 01000001_2.
- Le caractère a, en minuscule, correspond au nombre décimal 97_{10} dans la table ASCII que l'on note 61_{16} en hexadécimal et de codage binaire 01100001_2.

Remarque

La table ASCII contient des caractères spéciaux entre 00 et 32 comme les caractères blancs (espace, tabulation, etc), retours à la ligne, suppressions et des caractères de contrôle. Ce sont des caractères non imprimables

A ces 32 caractères spéciaux, il y en a un dernier de valeur décimale 127_{10} qui correspond à la touche del du clavier.

Table de caractères ASCII

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
000	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
001	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
002	SP	!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
003	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
004	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
005	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
006	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
007	p	q	r	s	t	u	v	w	x	y	z	{	l	}	~	DEL

Cette table est à double entrée. La colonne de gauche représente les trois premiers chiffres hexadécimaux et la première ligne le dernier chiffre hexadécimal du codage du caractère.

On lit la table de la façon suivante:

- on repère le caractère dans la table situé à l'intersection d'une ligne et d'une colonne;
- sur la ligne de ce caractère, on note les trois chiffres hexadécimaux de la première colonne;
- sur la colonne de ce caratère, on note le chiffre hexadécimal de la première ligne.
- on concatène les trois chiffres de la première colonne avec le chiffre de la première ligne.

Exemple

- Le A est situé sur la ligne contenant les chiffres 004 (première colonne) et sur la colonne contenant le chiffre 1 (première ligne). Donc le caractère A est représenté par la valeur hexadécimale 0041.
- Le symbole % est situé sur la ligne repérée 002 et sur la colonne repérée 5 donc il est représenté par le code hexadécimal 0025 et codé 00100101 en binaire.

3 Norme ISO 8859

Les caractères imprimables de la table ASCII sont devenus rapidement insuffisants pour transmettre des textes dans d'autres langues que l'anglais.

La table ASCII ne contient aucun caractère accentué ! Pour remédier à ce problème l'ISO, **Organisme International de Normalisation** a proposé la table ISO 8859 qui utilise 256 caractères sur un octet complet

Pour définir le plus de caractères possible, la norme ISO définit plusieurs tables 8859-n où n est le numéro de la table. Il y a 16 tables pour les différentes langues.

En France, on a utilisé la table 8859-1, appelée aussi *latin-1* et considérée comme multi-langue et la table ISO-8859-15 qui introduit le symbole monétaire € et qui gère mieux le français.

Remarque

Toutes les tables ISO 8859-n contiennent les 128 caractères ASCII et 128 autres caractères propre à une langue (latine, arabe, grec, cyrillique, hebreu, ...).

Il existe 10 tables différentes pour les langues latines !

Tables de caractères ISO 8859-1 et 8859-15

ISO/CEI 8859-1																
	x0	x1	x2	x3	x4	x5	x6	x7	x8	x9	xA	xB	xC	xD	xE	xF
0x	*positions inutilisées*															
1x																
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	{	|	}	~	
8x	*positions inutilisées*															
9x																
Ax	NBSP	¡	¢	£	¤	¥	¦	§	¨	©	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	ð	ñ	ò	ó	ô	õ	ö	÷	ø	ù	ú	û	ü	ý	þ	ÿ

ISO/CEI 8859-15																
	x0	x1	x2	x3	x4	x5	x6	x7	x8	x9	xA	xB	xC	xD	xE	xF
0x	*non utilisé*															
1x																
2x		!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
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	{	|	}	~	
8x	*non utilisé*															
9x																
Ax		¡	¢	£	€	¥	Š	§	š	©	ª	«	¬		®	¯
Bx	°	±	²	³	Ž	µ	¶	·	ž	¹	º	»	Œ	œ	Ÿ	¿
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	ð	ñ	ò	ó	ô	õ	ö	÷	ø	ù	ú	û	ü	ý	þ	ÿ

Les tables précédentes montrent que les 128 premiers caractères sont ceux de la table ASCII. Finalement, on a complété la table ASCII avec 128 nouveaux caratères (selon les langues et pays).

On peut remarquer quelques différences entre les tables ISO 8859-1 et ISO 8859-15 comme les fractions 1/4 et 1/2 peut utilisés en France et remplacés par les caractère OE, oe collés.

4 La norme Unicode

La norme ISO 8859 permet d'encoder un grand nombre de caractères mais cela ne suffisait toujours pas. La multiplicité des tables rend difficile la rédaction d'un texte en plusieurs langues qui utilisent différentes tables ! Par exemple écrire un texte mélangeant le japonais et le français.

L'ISO a donc défini un jeu universel de caractères sous la norme ISO 10646 appelée **UNICODE**.

Cette norme associe à chaque caractère (lettre, idéogramme, emoji,…) un **point de code**. Ce point de code est un **numéro** en **hexadécimal** préfixé par **U+**.

Exemple

- Le caractère "A" a pour point de code **U+0041** et nom unique **LATIN CAPITAL LETTER A** (on remarque que c'est sa valeur hexadécimale dans la table ASCII).
- Le caractère "?" a pour point de code **U+003F** et nom unique **QUESTION MARK** (on remarque aussi que c'est sa valeur hexadécimale dans la table ASCII).

Unicode est une norme qui définit les techniques pour encoder en binaire les **points de code** de façon plus ou moins économique. Ces encodages sont appelés **Universal Transform Format** et notés **UTF-n** où n désigne le nombre minimal de bits pour représenter un point de code (n=8, 16 ou 32).

UTF-8 est le format d'encodage qui utilise au minimum 8 bits pour coder les caractères. Il est compatible avec la table de caractère ASCII. Les 128 premiers points de code sont ceux de la table ASCII.

Selon les caractères et la valeur des points de code, l'encodage en UTF-8 utilise 1, 2, 3 ou 4 octets.

Le tableau ci-dessous donne l'encodage des caratères en UTF-8.

Plage	**Suite d'octets (en binaire)**	**Nombre de bits utilisés**
U+0000 à U+007F	0xxxxxxx	7 bits
U+0080 à U+07FF	110xxxxx 10xxxxxx	11 bits
U+0800 à U+FFFF	1110xxxx 10xxxxxx 10xxxxxx	16 bits
U+10000 à U+10FFFF	11110xxx 10xxxxx 10xxxxxx 10xxxxxx	21 bits

Exemple d'encodage en UTF-8

1. Commençons par un caractère ASCII dont le point de code est **U+005A**. On remarque que ce point de code appartient bien à la plage $[U+0000; U+007F]$ donc est codé sur 1 octet commençant par 0. En fait, il est encodé en UTF-8 comme en ASCII par la valeur binaire $5A_{16} = 01011010_2$

2. Prenons comme second exemple, un caractère accentué que l'on retrouve dans la table ISO 8859-15 dont le point de code est **U+00C9**. Le point de code appartient à la plage $[U+0080; U+07FF]$ donc il est encodé en UTF-8 sur 2 octets (contrairement à l'encodage ISO qui n'en utilisait qu'un seul). Comment est-il encodé en binaire ?

   L'encodage binaire sera de la forme $110xxxxx\ \ 10xxxxxx$ ou chaque x représente un bit issu du point de code.

   On convertit en binaire $00C9_{16} = 0000\ 0000\ \ 1100\ 1001_2$ et on remplace chaque bit x par un bit de C9.

Sur 2 octets	1	1	0	x	x	x	x	x		1	0	x	x	x	x	x	x
U+00C9				0	0	0	1	1				0	0	1	0	0	1
UTF-8	1	1	0	0	0	0	1	1		1	0	0	0	1	0	0	1

   En écriture hexadécimale, le caractère de point de code **U+00C9** est encodé en UTF-8 par **C3 89**.

3. Pour finir, le caractère de point de code **U+FFFD** appartient à la plage $[U+0800; U+FFFF]$. Ce caractère est donc encode en UTF-8 sur trois octets. Comment est-il encodé en binaire ?

L'encodage binaire sera de la forme $1110xxxx\ \ 10xxxxxx\ \ 10xxxxxx$ ou chaque x représente un bit issu du point de code.

On convertit en binaire $FFFD_{16} = 1111\ 1111\ \ 1111\ 1101_2$ et on remplace chaque bit x par un bit de C9.

Sur 3 octets	1	1	1	0	x	x	x	x	1	0	x	x	x	x	x	x	1	0	x	x	x	x	x	x
U+FFFD					1	1	1	1			1	1	1	1	1	1			1	1	1	1	0	1
UTF-8	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1	0	1	1	1	1	0	1

En écriture hexadécimale, le caractère de point de code **U+FFFD** est encodé en UTF-8 par **EF BF BD** soit .

Remarque

En python, on peut déclarer, sous réserve de coder en UTF-8, directement un caractère par son point de code unicode.

Par exemple pour le symbole : c='$\backslash uFFFD$'

Il existe une méthode **encode()** qui donne le codage binaire noté en hexadécimal d'un caratère dont on connait le point de code:

Avec c='$\backslash uFFFD$' et en exécutant c.encode(), cela renvoie $b'\backslash xe2\backslash x82\backslash xac'$ qui est le code hexadécimal de l'encodage en UTF-8 du caractère.

Sitographie

- Sur wikipedia : Norme ISO 8859
- Sur wikipedia : UTF-8

[1]:
```
c='\u00c3'
c.encode()
```

[1]:
```
b'\xc3\x83'
```

[2]:
```
c
```

[2]:
```
'Ã'
```

[3]:
```
p=b'\xc3\xb8'
```

[4]:
```
p.decode()
```

[4]:
```
'ø'
```

[5]:
```
c='\uFFFD'
c.encode()
```

[5]:
```
b'\xef\xbf\xbd'
```

[6]:
```
c
```

[6]:
```
''
```

[7]:
```
c='\u00e9'
```

[8]:
```
c.encode()
```

[8]:
```
b'\xc3\xa9'
```