Histoire des sciences

L'encodage des caractères alphanumériques

1963 : Naissance du code ASCII

Afin de représenter les caractères alphanumériques avec un ordinateur, le **code ASCII** (*American Standard Code for Information Interchange*) est développé pour permettre l'uniformisation du codage et faciliter les échanges d'informations. À ses débuts, le code ASCII permettait uniquement de représenter les caractères principaux de la langue anglaise, ainsi que des caractères « non imprimables » comme le retour à la ligne, codés sur 7 bits.

```
 !"#$%&'()*+,-./
0123456789:;<=>?
@ABCDEFGHIJKLMNO
PQRSTUVWXYZ[\]^
`abcdefghijklmno
pqrstuvwxyz{|}~
```

1986 : Apparition de la norme ISO 8859-1

Extension de la norme américaine ASCII, la norme **ISO 8859-1**, encore appelée Latin-1, est formée d'un ensemble de 191 caractères de l'alphabet latin, tous codés sur un unique octet. Elle permet de représenter la très grande majorité des caractères de nombreuses langues d'Europe occidentale. Des extensions existent pour d'autres alphabets, comme le cyrillique ou le polonais.

1988 : Le standard Unicode

Grâce au standard Unicode, les caractères de tous les systèmes d'écriture du monde peuvent posséder un nom et un identifiant numérique. Les caractères sont représentés selon la norme U+xxxx où xxxx est un nombre composé de 4 à 6 caractères hexadécimaux. L'**interopérabilité** entre différents logiciels devient possible. Ce standard code actuellement 137 929 caractères.

1996 : la norme UTF-8 est adoptée

La **norme d'encodage UTF-8** est une représentation d'Unicode qui possède un encodage de taille variable selon le caractère à coder, ce qui permet de limiter le coût en mémoire et d'assurer la compatibilité avec la norme ASCII, plus ancienne. Les caractères de la table ASCII restent ainsi codés sur 7 bits, tandis que les autres caractères sont représentés sur 2, 3 ou 4 octets (voir ci-dessus). L'encodage UTF-8 est utilisé par plus de 95 % des sites web en octobre 2020.

A	é	語	
41	C3 | A9	E8 | AA | 9E	F0 | 90 | 8E | 84

1. Combien de caractères différents peut-on coder sur 7 bits ?

Sur 7 bits, on peut coder 2^7 caractères différents, soit 128 caractères différents.

2. Pouvez-vous imaginer une des limites du code ASCII ?

De nombreux caractères européens comme les lettres accentuées n'étaient pas représentables.

3. Quel est l'intérêt d'avoir ajouté un 8e bit de codage dans la norme Latin-1, en termes de quantité de caractères représentables ?

L'ajout d'un 8e bit permet de doubler le nombre de caractères représentables, soit 256 en tout.

4. Dans le traitement de texte de votre choix, saisir Alt+65. Quel caractère de la table UTF-8 obtenez-vous ? Justifier que le caractère A soit codé par la valeur 41 en UTF-8.

On obtient un A majuscule. Son code UTF-8 est 41 car $41_{16} = 65_{10}$.

5. À quels caractères de la table UTF-8 correspondent les saisies Alt+0192, Alt+0201 et Alt+0200 ?

Alt+0192 correspond au caractère À, Alt+0201 correspond au caractère É et Alt+0200 à È.

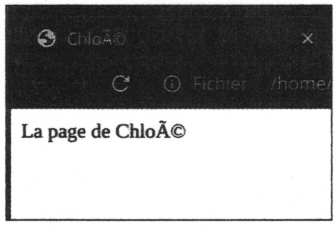
© BORDAS/SEJER, 2021 – Photocopie non autorisée.

Activité 4 / Représentation d'un texte en machine

```
1   <!doctype html>
2   <html lang="fr">
3   <head>
4     <meta charset="iso-8859-1">
5     <title>Chloé</title>
6   </head>
7   <body>
8     La page de Chloé
9   </body>
10  </html>
```

Chloé a trouvé sur le Web la page HTML présentée ci-contre, qu'elle a modifiée elle-même.

Lorsqu'elle ouvre le fichier modifié dans son navigateur, elle est surprise de voir le résultat ci-dessous :



1. Que peut-on remarquer ?

Le caractère accentué « é » du prénom Chloé n'est pas affiché correctement.

Curieuse, Chloé ouvre le fichier HTML avec un éditeur hexadécimal, qui permet de voir le contenu du fichier sous forme d'octets à gauche et les caractères décodés à droite. Un espace est codé en hexadécimal par un octet de valeur 20.

```
00000000: 3c 21 64 6f 63 74 79 70 65 20 68 74 6d 6c 3e 0a  <!doctype html>.
00000010: 3c 68 74 6d 6c 20 6c 61 6e 67 3d 22 66 72 22 3e  <html lang="fr">
00000020: 0a 3c 68 65 61 64 3e 0a 20 20 3c 6d 65 74 61 20  .<head>.  <meta
00000030: 63 68 61 72 73 65 74 3d 22 69 73 6f 2d 38 38 35  charset="iso-885
00000040: 39 2d 31 22 3e 0a 20 20 3c 74 69 74 6c 65 3e 43  9-1">.  <title>C
00000050: 68 6c 6f c3 a9 3c 2f 74 69 74 6c 65 3e 0a 3c 2f  hloÃ©</title>.</
00000060: 68 65 61 64 3e 0a 3c 62 6f 64 79 3e 0a 20 20 4c  head>.<body>.  L
00000070: 61 20 70 61 67 65 20 64 65 20 43 68 6c 6f c3 a9  a page de ChloÃ©
00000080: 0a 3c 2f 62 6f 64 79 3e 0a 3c 2f 68 74 6d 6c 3e  .</body>.</html>
```

On pourrait modifier directement les octets du fichier à l'aide d'un éditeur hexadécimal, ce qui serait une autre solution au problème de Chloé.

2. Sachant qu'il y a correspondance entre le texte et son codage, compléter la valeur des octets associés au texte « La page de Chloé ». Comment est codé le caractère « é » dans le fichier ?

4c 61 20 70 61 67 65 20 64 65 20 43 68 6c 6f c3 a9. Le caractère « é » est codé c3 a9.

Chloé a trouvé sur le Web la table de caractères ISO 8859-1 suivante :

3. En utilisant la table, justifier l'affichage « Ã© » visible sur la page web.

c3 correspond bien au caractère Ã et a9 correspond bien au caractère ©.

4. Le fichier a-t-il été encodé selon la norme ISO 8859-1 ? Quel codage devrait représenter la lettre « é » ?

Non, car en ISO 8859-1, le « é » aurait dû être codé sur un seul octet : e9.

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

5. L'éditeur de texte de Chloé lui indique que son fichier est encodé en UTF-8. Proposer une solution pour résoudre son problème.

Le fichier HTML indique qu'il est codé en ISO 8859-1, alors que le fichier est en réalité codé en UTF-8. Le navigateur n'utilise donc pas la bonne table de caractères pour afficher le texte. Il faudrait remplacer ISO 8859-1 par UTF-8 dans l'en-tête HTML. La solution recommandée sur le Web est de remplacer le caractère é par &eacute, directement dans le code HTML de la page web, en utilisant un codage UTF-8.

© BORDAS/SEJER, 2021 – Photocopie non autorisée.