1 / Da ASCII code !

Au commencement, chaque caractère était identifié par un code unique qui est un entier naturel et la correspondance entre le caractère et son code était appelée un ***Charset****. Le code n'étant pas utilisable tel quel par un ordinateur qui ne comprend que le binaire, il fallut donc représenter les codes par des octets, et cela fut appelé* ***Encoding****.*

Dans de nombreux grimoires anciens on découvre le code ASCII qui était utilisé pour représenter du texte en informatique. ASCII signifiait American Standard Code for Information Interchange. Il paraît que ce code est toujours en usage...

1.1 Activité – Taille d'un texte

Quelle est la taille (en octets) de la phrase : « Enfin ! Je viens de comprendre ce qui s'est produit. » (attention, il faut compter les espaces, et signes de ponctuation…) ?

1.2 Activité – Utilisation de la table ASCII

1) À l'aide de la table ASCII, coder en binaire la phrase suivante : « L'an qui vient ! ».

2) Voici maintenant une exclamation codée en binaire :

01000010 01110010 01100001 01110110 01101111 00100001

Retrouver cette exclamation !

3) Peut-on coder en binaire la phrase « Un âne est-il passé par là?» à l'aide de la table ASCII ?

2 / Quand la table ASCII ne suffit plus

Il va donc falloir étendre la table ASCII pour pouvoir coder les nouveaux caractères. Les mémoires devenant plus fiables et, de nouvelles méthodes plus sûres que le contrôle de parité ayant été inventées, le 8ième bit a pu être utilisé pour coder plus de caractères.0

3 / De la difficulté de convenir d'une norme

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

Windows-1252 (CP1252)

	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	€		‚	ƒ	„	…	†	‡	ˆ	‰	Š	‹	Œ		Ž	
9x		‘	’	“	”	•	–	—	˜	™	š	›	œ		ž	Ÿ
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

3.1 Activité – Utilisation d'un logiciel

On commence parce se connecter au site suivant :

https://wiki.inria.fr/sciencinfolycee/Convertisseur_texte/binaire/hexa_en_ligne

Voici le code binaire d'un texte :

01000010 01110010 01100001 01110110 01101111 00101100 00100000 01110100 01110101 00100000
01100001 01110011 00100000 01110000 01110010 01100101 01110011 01110001 01110101 01100101
00100000 01110100 01101111 01110101 01110100 00100000 01110100 01110010 01101111 01110101
01110110 11101001 00101110 00101110 00101110

À l'aide du logiciel fourni sur le site, retrouver le texte contenu dans le code.

4 / Quand le net s'affole...

Nous avons tous un jour reçu un courriel bizarre ou lu une page web telle que celle-ci :

Prenons lâ€™exemple typique de la lumiÃ¨re Ã©mise par un phare maritime : elle est dâ€™abord indivisible, son coÃ»t de production Ã©tant alors indÃ©pendant du nombre d'utilisateurs ; elle possÃ¨de une propriÃ©tÃ© de non-rivalitÃ© (elle ne se dÃ©truit pas dans l'usage et peut donc Ãªtre adoptÃ©e par un nombre illimitÃ© d'utilisateurs) ; elle est Ã©galement non excluable car il est impossible dâ€™exclure de lâ€™usage un utilisateur; mÃªme si ce dernier ne contribue pas Ã son financement.

5 / Et l'Unicode vint...

L'Unicode est une table de correspondance Caractère-Code (Charset), et l'UTF-8 est l'encodage correspondant (Encoding) le plus répandu. Maintenant, par défaut, les navigateurs Internet utilisent le codage UTF-8 et les concepteurs de sites pensent de plus en plus à créer leurs pages web en prenant en compte cette même norme ; c'est pourquoi il y a de moins en moins de problèmes de compatibilité.

6 / Quelques précisions sur l'UTF-8

6.1 Des règles, encore des règles

L'encodage UTF-8 utilise 1, 2, 3 ou 4 octets en respectant certaines règles :

• Un texte en ASCII de base (appelé aussi US-ASCII) est codé de manière identique en UTF-8. On utilise un octet commençant par un bit 0 à gauche (bit de poids fort).

Caractère	Point de code (hexadécimal)	Valeur scalaire		Codage UTF-8
		décimal	binaire	binaire
A	U+0041	65	1000001	01000001

• Les octets ne sont pas remplis entièrement. Les bits de poids fort du premier octet forment une suite de 1 indiquant le nombre d'octets utilisés pour coder le caractère. Les octets suivants commencent tous par le bloc binaire 10.

Définition du nombre d'octets utilisés

Représentation binaire UTF-8	Signification
0xxxxxxx	1 octet codant 1 à 7 bits
110xxxxx 10xxxxxx	2 octets codant 8 à 11 bits
1110xxxx 10xxxxxx 10xxxxxx	3 octets codant 12 à 16 bits
11110xxx 10xxxxxx 10xxxxxx 10xxxxxx	4 octets codant 17 à 21 bits

• Dans la norme ISO 8859-1 le « é » est codé 1110 1001, en UTF-8 on le code sur deux octets en respectant les précisions apportées dans le tableau ci-dessus. Les bits imposés sont en gras, le code du « é » est écrit en commençant par la droite et l'octet de gauche est rempli par des zéros (en italique). Voilà ce que l'on obtient : **110***000*11 **10**101001. On pourra remarquer que le codage ISO s'inscrit bien dans le codage UTF-8.

6.2 Activité – Coder en UTF-8

Le symbole € correspond à la valeur décimale 8364.

1) Convertir cette valeur en binaire.

2) Combien d'octets doit-on utiliser en UTF-8 pour coder ce nombre convenablement (les moitiés d'octet sont interdites) ?

3) Donner le codage UTF-8 correspondant.