Activité : Encodage du texte

Les caractères ASCII

La première table de caractères a été créée aux États-Unis en 1960. C'est la table ASCII reproduite ci-dessous.

	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	{	|	}	~	DEL

1) Quels caractères correspondent aux codes hexadécimaux 006E, 0053 et 0049 ? Donner pour chaque caractère sa valeur binaire et décimale.

2) Quels sont les codes hexadécimaux des caractères '=' et 'x' ?

3) Quel est le principal problème de cette table de caractères ?

Norme UNICODE et encodage UTF-8

La norme Unicode a été créée au début des années 2000 et contient tous les caractères existants dans le monde et dans toutes les langues. L'encodage en binaire d'un caractère se fait en 2 temps :

- Chaque caractère de cette norme est repéré par un point de code : U+v où v est une valeur hexadécimale.
- Selon la valeur du point de code, le caractère est encodé en UTF-8 sur 1, 2, 3 ou 4 octets suivant le tableau ci-dessous :

Plage	Suite d'octets (en binaire)	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

1) Les caractères ASCII ont un point de code qui correspond au code hexadécimal de la table.

- **a** Quel est le point de code du caractère A ?
- **b** Combien d'octets faut-il pour encoder un caractère ASCII en UTF-8 ?

2) Comment sont encodés en UTF-8 les caractères de la table ASCII ?

3) Rechercher sur le web :

- **a** le point de code associé au caractère 'È' (alt+212) ? 'æ' (alt+145) ? '£' ?
- **b** le caractère associé au point de code est U+20AC ?

4) Le caractère de point de code U+20AC est encodé en UTF-8.

- **a** Donner le nombre d'octets utilisés pour l'encodage de ce caractère en UTF-8.
- **b** En déduire, en vous aidant du tableau, l'encodage de ce caractère.