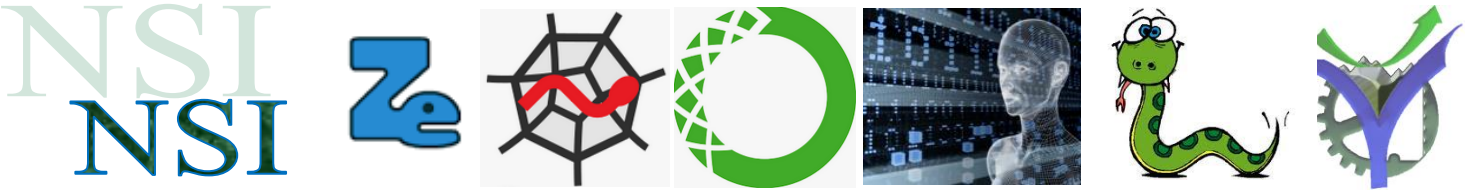


Découvrir et approfondir Unicode

Unicode est un système de **codage** informatique qui a pour but d'unifier les échanges de texte au niveau international. Avec **Unicode**, chaque caractère informatique est décrit par un nom et un **code** l'identifiant de manière unique quelque soit le support informatique ou le logiciel utilisé.

Codage Unicode \uXXXX - Déchiffrer, Décoder, Encoder
https://www.dcode.fr › codage-unicode

A rédiger sur copie sauf l'exercice 2 page 3 qui sera fait sur la feuille réponse

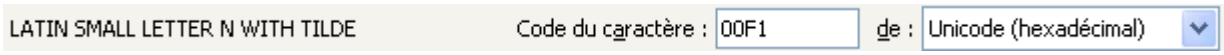
1 Le codage Unicode

1.1 Identification des caractères avec un numéro et un nom unique

Avec 7 bits il est bien évident qu'il est impossible de coder 'tous les caractères du monde'. Il a bien fallu normaliser de nouveau les tables de codage. Une tentative a eu lieu avec la norme ISO 8859 https://fr.wikipedia.org/wiki/ISO/CEI_8859

Le standard de codage Unicode s'est imposé, il propose une représentation unique de chaque caractère **par son nom** et **un code appelé point code**. Le code est noté U+xxxx où chaque x représente une valeur hexadécimale.

Exemple : le caractère ñ a pour code Unicode U+00F1 et pour nom LETTRE MINUSCULE LATINE N TILDE (trouvé dans Word Supplément Latin-1) :



Ce codage permet de mélanger tous caractères dans un même document. Il est compatible avec le code ASCII pour les 128 premiers caractères.

Quelques codes Unicode :

	Codage Unicode	
Caractère	**Nom du caractère**	**Point Code**
☮	**Symbole de paix**	**U+262E**
♫	**Beamed eighth notes**	**U+262B**
€	**Symbole Euro**	**U+20AC**
A	**Latin majuscule A**	**U+0041**
è	**Latin e minuscule accent grave**	**U+00E8**

1.2 Encodage du caractère en binaire

Une fois connu le point code du caractère noté U+xxxx pour déterminer son code binaire effectif il faut tout d'abord choisir une méthode de codage UTF-8, UTF-16, UTF-32. La méthode la plus couramment utilisée aujourd'hui est l'UTF-8.

L'UTF-8 encode les caractères avec un mot de 1 à 4 octets organisés comme suit :

Number of bytes	Bits for code point	First code point	Last code point	Byte 1	Byte 2	Byte 3	Byte 4
1	7	U+0000	U+007F	0xxxxxxx			
2	11	U+0080	U+07FF	110xxxxx	10xxxxxx		
3	16	U+0800	U+FFFF	1110xxxx	10xxxxxx	10xxxxxx	
4	21	U+10000	U+10FFFF	11110xxx	10xxxxxx	10xxxxxx	10xxxxxx

Le premier octet indique le nombre d'octet total du mot. Si ce premier octet commence par un 0 le code est écrit sur un seul octet, compatible asccii. Si le premier octet commence par un 1 alors il faut compter le nombre de 1 se trouvant à la suite avant le premier 0 pour connaitre le nombre total d'octets du code. Ensuite tous les octets commencent par 10 de la sorte il n'y a aucune confusion possible pour le décodage.

Exemple pour un codage sur 3 octets :

1110xxxx | 10xxxxxx | 10xxxxxx

Le premier octet, encadré en jaune, commence par : 1 1 1 0 x x x x il y a trois valeurs 1 avant le premier 0 donc nous avons trois octets au total. Les x indiquent le nombre de bits disponibles pour écrire en binaire la valeur du point code ici 4 + 6 + 6 = 16.

Exemple codage du caractère è en UTF-8 :

Le point code vaut : U+00E8 soit 1110 1000 en binaire.

Ce code a besoin de 8 bits pour être écrit.

En UTF-8 il faut deux octets. Le codage est de la forme : 110xxxxx | 10xxxxxx

Il ne reste plus qu'à remplir les x en commençant par la droite, les poids faibles, et en complétant par des 0 le résultat est donné ci-dessous :

1 1 0 0 0 0 1 1 | 1 0 1 0 1 0 0 0

En hexadécimal cela nous donne : 0x**C3A8**

Exercice 1. En utilisant les données et la méthode explicitée ci-dessus déterminez les codages UTF-8 des caractères ci-dessous <u>en détaillant les calculs</u> :

☮ U+262E € U+20AC A U+0041

Cette page est à rendre avec votre copie

Exercice 2. Décodage de caractères Unicode

Vous recevez le flux de bits suivants composés de caractères Unicode : CF 81 CE A9 40 C3 A8

Pour décoder le message suivre les étapes ci-dessous :

a) Écrire le message en binaire

b) Repérer le découpage du codage UTF-8 en 1,2 ou 3 octets, (il n'y a pas de caractère encodé sur 4 octets dans cet exemple), pour cela après avoir recopié vos valeurs précédentes dans le tableau ci-dessous entourez les différents groupes avec de la couleur :

0

1 1 0

1 1 1 0

1 0

Les trois premiers formats de codage de l'UTF-8

c) Dans chacun des groupes identifiés **isoler l'information point code** du caractère reçu, pour vous aider vous devez trouver la séquence sous la forme ci-dessous :

1 1 0 1 0

1 1 0 1 0

0

1 1 0 1 0

La suite des caractères reçus est donc :