LE STANDARD UNICODE UTF-8

1. INTRODUCTION

Le standard UNICODE a été créé dans le but de normaliser la création et l'échange de documents au niveau mondial.
Utilisé dans le codage des pages HTML et de l' XML, par les logiciels de programmation (comme Java par exemple), il est l'élément incontournable de tout concepteur en informatique.
Unicode a été conçu au départ sur le modèle du jeu de caractères ASCII qui ne pouvait coder que 128 caractères « occidentaux » essentiellement en majuscule et minuscule mais sans accent. Unicode permet de coder tous les caractères utilisés par toutes les langues écrites du monde,plus d'un million de caractères sont réservés à cet effet.
Donc :

- A chaque caractère de l'alphabet international, on doit associer une représentation binaire
- Si des nouveaux caractères sont découverts, la compatibilité ascendante ne doit pas détruire pas les codes précédents.

La version actuelle au moment de la rédaction de ce document est la version 6.1.0 (31/01/12)
Sur le site http://www.unicode.org, vous trouverez la version actuelle

2. L'UTF-8

L'UTF-8 (UTF = Unicode Transformation Format) est un des systèmes d 'encodage couramment utilisé. Il remplit les conditions suivantes :

- Les caractères faisant partie de l'ASCII sur 7 bits restent inchangés. L'objectif est que les anciens fichiers codés avec l'ASCII soit encore reconnus par l'UTF-8 ;
- Les caractères encodés peuvent occuper de 1 à 4 octets dans le fichier (de 8 à 32 bits)

2.1. LE CODAGE

Les caractères sont codés de la façon suivante :

- entre U+0000 et U+007F : elles sont stockées sur un seul octet (on retrouve l'ASCII)
- entre U+0080 et U+07FF : elles sont stockées sur deux octets
- entre U+0800 et U+FFFF : elles sont stockées sur trois octets
- entre U+10000 to U+10FFFF : elles sont stockées sur quatre octets

Pour qu'un logiciel reconnaisse le nombre d'octets utilisés, certains bits ne peuvent pas être modifiés, en l'occurrence les bits de poids forts (en gras ci-dessous). Les X représentent les bits modifiables en fonction du caractère à coder.

	N° UNICODE	Représentation binaire UTF-8	Signification
1	U+0000 à U+ 007F	0XXXXXXX	1 octet codant 1 à 7 bits
2	U+0080 à U+ 07FF	110XXXXX 10XXXXXX	2 octets codant 8 à 11 bits
3	U+0800 à U+ FFFF	1110XXXX 10XXXXXX 10XXXXXX	3 octets codant 12 à 16 bits
4	U+10000 à U+ 1FFFFF	11110XXX 10XXXXXX 10XXXXXX 10XXXXXX	4 octets codant 17 à 21 bits

2.2. EXEMPLES :

le caractère $

- $: U+0024 est compris entre U+0000 et U +007F. Donc correspond au cas 1.
  0024 = 0000 0000 0010 0100 (les 0 à gauche ne comptent pas)
  On ne garde que les 7 bits les plus à droite et on les met à droite du 0, qui lui est figé.

  $ sera donc codé : 0 0100100 = 24 en hexadécimal (il existait en ASCII et n'occupe pas plus d'un octet)

le caractère ½

- ½ = U+00BD est compris entre U+0080 et U + 07FF . Donc correspond au cas 2.
  00BD = 1011 1101 (je ne mets plus les 0 à gauche)
  On place les bits aux endroits disponibles en commençant par la droite. Ensuite on termine par des zéros.

  1ère étape :
  1011 1101 → 110XXXXX 10111101
  2de étape :
  1011 1101 → 110XXX10 10111101

  3ème étape : on place des 0 aux endroits inutilisés
  110XXX10 10111101 → 11000010 10 111101

  ½ sera finalement codé : 1100001010 111101 soit C2 BD soit 2 octets.

le caractère ∫ (intégrale)

- ∫ = U+222B est compris entre U+0800 et U + FFFF . Donc correspond au cas 3.
  222B = 10 0010 0010 1011
  on remplace les X par les valeurs de chaque bit en commençant par la droite.

  1110XXXX 10XXXXXX 10XXXXXX devient donc :

  111000101000100010101011 soit :
  1110 0010 1000 1000 1010 1011 = E288AB

Le cas 4 ne sera pas traité ici mais le principe est le même.