Le codage des caractères

Activité

|  |  |
| --- | --- |
| En recherchant un extrait du « Seigneur des  anneaux », nous sommes tombés sur la page WEB ci-contre.  1. Indiquer pourquoi une partie du texte n’est pas compréhensible.  2. Préciser quel type de lettres pose problème. |  |

1ère solution : le code ASCII

Le code ASCII (**A**merican **S**tandard **C**ode for **I**nformation **I**nterchange), défini aux Etats-Unis en 1963, est basé sur un tableau contenant les caractères les plus utilisés en langue anglaise. Chaque caractère est représenté sur 7 bits (ce qui donne 27= 128 combinaisons possibles).

a. Coder, à l'aide de la table du §3.5, la phrase suivante : "La compagnie de l’anneau".

b. Retrouver le texte correspondant au code ASCII suivant : (46 72 6F 6E 64 6F 6E 20 6C 65 20 48 6F 62 62 69 74)16.

c. Justifier, pourquoi il n'est pas possible de coder correctement le texte "Mon précieux" à l'aide du code ASCII.

2ème solution : le code ISO 8859–1 (latin 1)

La nécessité de représenter des caractères non présents dans la table ASCII tels que ceux de l’alphabet latin comme le « à », le « é », « ç »... impose l’utilisation d’un autre code.

Ces codes sont des extensions du code ASCII. Pour cela le 8ième bit est utilisé ce qui permet de coder 256 caractères (128 caractères supplémentaire par rapport au code ASCII de base). On parle de code ASCII étendu. L’ISO, organisation internationale de normalisation, propose plusieurs variantes de ce code, adaptée aux différentes langues. Nous utilisons la norme **ISO-8859-1** nommée aussi **ISO-Latin1**.

Coder le texte "Mon précieux" en utilisant la table « ASCII étendu » du §3.5

3ème solution : Unicode

La généralisation de l’utilisation d’Internet dans le monde a ainsi nécessité une prise en compte d’un nombre beaucoup plus important de caractères. Ce que permet la norme Unicode qui établit une correspondance unique caractère ↔ code numérique (on dit aussi « charset »). Le répertoire Unicode peut contenir plus d'un million de caractères. Unicode définit des méthodes standardisées pour coder et stocker cet index sous forme de séquence d'octets : UTF-8, UTF-16, UTF-32 et leurs différentes variantes. L’UTF-8 est l’encodage (Encoding) le plus répandu. Les navigateurs Internet utilisent le codage UTF-8 par défaut. Le langage de programmation utilisé en NSI (Python) gère l'Unicode par défaut.

L’index des caractères UNICODE est disponible à l’adresse suivante : <http://www.unicode.org/fr/charts/charindex.html>.

Par ailleurs, une description détaillée de l'UTF-8 peut être trouvée sur Wikipedia : <https://fr.wikipedia.org/wiki/UTF-8>

*Description rapide et simplifiée*:

L'encodage UTF-8 utilise 1, 2, 3 ou 4 octets. Si le code du caractère est inférieur ou égal à 127, on n’utilise qu'un octet avec le bit de poids fort à 0. Sinon, 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.

|  |  |
| --- | --- |
| Représentation binaire UTF-8 | Signification |
| 0xxx xxxx | 1 octet utilisé |
| 110x xxxx 10xx xxxx | 2 octets utilisés |
| 1110 xxxx 10xx xxxx 10xx xxxx | 3 octets utilisés |
| 1111 0xxx 10xx xxxx 10xx xxxx 10xx xxxx | 4 octets utilisés |

Sous Unicode le caractère *inférieur ou égal* a comme point de code U+2A7D.

a. Quelle la valeur binaire de 0x2A7D ? Combien d’octets seront nécessaires pour représenter ce caractère en unicode UTF-8 ?

b. Déterminer la représentation binaire puis hexadécimale caractère inférieur ou égal en Unicode UTF-8.