TP7 - Encodage des caractères en HTML

1 Les différentes solutions

Le jeu de caractère utilisé pour encoder une page html peut être différent d'une page à l'autre. Toujours basé sur le code ASCII pour les caractères non accentués, on peut utiliser trois approches pour coder les caractères "spéciaux" (non-inclus dans la table ASCII 0 à 127) :

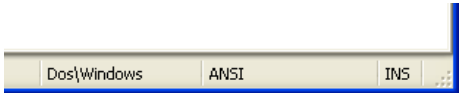
1. On spécifie les caractères autres que ASCII via leur "entity" correspondante (voir TP1). Par exemple, on va coder "€" par les 6 caractères **`€`** et 'é' par les 8 caractères **`é`**
2. On encode la page avec toujours la correspondance "un octet = un caractère" et on utilise pour les octets de 128 à 255 une page de code qui spécifie la correspondance avec un caractère donné.
3. On utilise un encodage sur plusieurs octets par caractères (UTF-8, UTF-16, UTF-32), ce qui permet d'avoir une représentation unique d'un caractère donné.

2 Manipulations

Travail préliminaire Dans la suite, vous allez avoir besoin d'un **éditeur hexadécimal**. Vous utiliserez le logiciel Frhed.

Vérifier si dans le dossier **`C:\temp`** le dossier **`frhed`** existe. S'il n'y est pas, l'y copier depuis le dossier **`P:\M1106`**. Lancer Frhed depuis le disque local. Dans les préférences (CTRL-I), réglez la largeur des lignes à 16 octets et décochez la case d'ajustement automatique juste en dessous.

1. Se connecter sur la page `http://fr.wikipedia.org/wiki/ASCII` et relever le codage hexa des caractères suivants :
   'a' : _____ 'e' : _____ '.' (point) :_____
2. Regler Notepad++ en mode ANSI, avec le jeu de caractère ISO 8859-15 (Menu *Encodage → Langues d'europe occidentale*). Dans un nouveau fichier, saisir la chaîne €.a.à.e.é.è" (6 lettres séparées par 5 points). Ne **pas ajouter** de saut de ligne ou d'espace, et sauvegarder comme **`test_iso8859-15.txt`**



3. Regler Notepad++ en mode ANSI, avec le jeu de caractère Windows-1252. (Re)Taper la même chaîne que précedemment (l'affichage doit faire apparaître la chaîne demandée). Sauvegarder en **`test_1252.txt`**.
4. Regler Notepad++ en mode UTF-8 (sans BOM). (Re)Taper la même chaîne que précedemment (l'affichage doit faire apparaître la chaîne demandée). Sauvegarder en **`test_utf8.txt`**.
5. Ouvrir les trois fichiers avec Frhed et relever le code hexa obtenu pour chacune de ces 6 lettres dans le tableau ci-dessous, en se repérant par rapport au code ASCII du caractère "." (point).

	€	a	à	e	é	è
ISO 8859-15						
Windows-1252						
UTF-8						

Quelle conclusion en tirez vous sur les différences entre UTF-8 et les autres encodages :

_____

6. Ajouter dans chacun des trois fichier ci-dessus un squelette html de façon à le rendre conforme (paires de balises **`html`**, **`head`**, **`body`**, etc.), et les sauvegarder en trois fichiers **`.html`** de même nom que le **`.txt`** correspondant.
7. Les ouvrir dans un navigateur. Lequel s'affiche correctement ?
8. Ajouter dans l'en-tête de chacun de ces fichiers une balise meta portant la spécification correcte de l'encodage de caractère utilisé, et vérifier que l'affichage est correct dans les trois cas. La syntaxe est :
   `<meta charset="XXX">`, avec **`XXX`** qui peut être :
   `UTF-8` ou **`ISO-8859-15`** ou **`windows-1252`**.
   Vérifier que les trois pages s'affichent correctement. Quelle est la méthode d'encodage la plus universelle ? : _____
9. Dans le fichier **`test_iso8859-15.html`**, remplacer l'encodage par **`ISO-8859-1`**. Quelle différence cela fait-il sur l'affichage dans le navigateur :

3 Analyse des différentes solutions

Suite à vos expérimentations, donnez les avantages et inconvénients des différentes solutions d'encodage.

1. Encodage par "entity" HTML :
   - Avantage : _____
   - Inconvénient : _____
2. Encodage via page de code :
   - Avantage : _____
   - Inconvénient : _____
3. Encodage en UTF-8 :
   - Avantage : _____
   - Inconvénient : _____
4. Quel est la différence entre UTF-8, UTF-16 et UTF-32 ?
   - UTF-8 : _____
   - UTF-16 : _____
   - UTF-32 : _____