Indiquer correctement l'encodage des fichiers envoyés par un serveur Web

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 9 juin 2008

https://www.bortzmeyer.org/set-charset-web.html

Idéalement, il n'existerait qu'un seul jeu de caractères universel et un seul encodage de ce jeu et tout le monde l'adopterait et il n'y aurait pas de problème. Mais dans le monde où nous vivons, beaucoup de jeux coexistent sur Internet et encore plus d'encodages différents. Pour que les visiteurs de votre site Web voient les caractères corrects et non pas des signes cabalistiques (comme dans le fameux pastiche de Martine, « Martine [Caractère Unicode non montré [1]][Caractère Unicode non montré]crit en UTF-8 »), il faut un peu de méthode.

Il n'y a en effet pas de truc magique, juste de la méthode. Celle que je suggère, et qui doit être suivie dans cet ordre est exposée ici.

1) Décider d'un jeu de caractères et d'un encodage pour les pages Web publiées. On prend ce qu'on veut, l'**important** est de s'y tenir. Pour le jeu de caractères, Unicode est le choix évident, puisqu'il permet de représenter tous les caractères. Pour des pages en français, son encodage UTF-8 est un bon choix. Mais on peut aussi utiliser ISO-8859-15 ou ISO-8859-1, l'important est de ne pas changer tout le temps.

2) Trouver comment produire les pages à cet encodage. Cela dépend du processus de publication. Si l'encodage choisi est UTF-8, si on édite les pages avec Emacs en UTF-8, il n'y a rien à faire. Si on édite les pages avec Emacs en ISO-8859-1, il fait convertir avec recode <http://www.gnu.org/software/recode/>. Avec un CMS, il faut lire la documentation du CMS. Etc.

3) S'assurer que les pages contiennent l'indication **correcte** de l'encodage. En XML / XHTML, cela se fait dans la déclaration. Exemple : `<?xml version="1.0" encoding="UTF-8" ?>`. En HTML traditionnel, cela se fait via les éléments `<meta>`. Exemple : `<meta http-equiv="content-type" content="text/html; charset=iso-8859-15" />`. Il parait que MSIE ignore l'encodage XML et il peut donc être prudent de mettre le `<meta>` de toute façon, c'est ce que je fais sur mon blog.

4) S'assurer que le serveur HTTP envoie la bonne indication d'encodage. (Cette étape peut être facultative, ça dépend de beaucoup de choses. Dans le doute, je la fais toujours.) Cela se règle, avec Apache, grâce à `AddCharset`. Exemple : `AddCharset UTF-8 .html`. Ensuite, le validateur du W3C <http://validator.w3.org/> doit permettre de vérifier cela. On peut aussi utiliser l'excellente extension Firefox WebDeveloper <http://chrispederick.com/work/web-developer/> (menu "*Information*", option "*View Response Headers*"). Sinon, sur Unix, avec wget :

1. Car trop difficile à faire afficher par LaTeX

```
% wget --server-response --output-document /dev/null http://www.example.org/
  HTTP/1.1 200 OK
  Date: Tue, 13 May 2008 19:22:43 GMT
  Server: Apache/1.3.34 Ben-SSL/1.55
  X-Powered-By: PHP/4.4.8
  Connection: close
  Content-Type: text/html
```

Ici, aucun "*charset*" n'est indiqué dans la ligne `Content-Type`, ce qui est clairement mauvais. Questions :

— Quelle est la valeur par défaut ?
— Est-ce vraiment bien prudent de compter sur cette valeur par défaut ?