Encodage et internationalisation

Charset Iso-8859-1, iso-8859-15, utf-8, lequel choisir ?

Il faut tout d'abord distinguer deux «familles» d'encodage: les «locaux» et les «internationaux».

Les jeux de caractères locaux (dont font partie iso-8859-1 et iso-8859-15 -- parfois désignés comme «latin1» et «latin9») sont destiné à des documents dans un seul système d'écriture (une langue ou un groupe de langue utilisant un même alphabet ou syllabaire).

Au contraire les jeux de caractères internationaux (dont fait partie l'utf-8) sont destinés à encoder des document dans n'importe quel système d'écriture (et donc n'importe quelle langue).

Alors que choisir ?
On peut observer deux cas de figure:

1. Mon site gère différents systèmes d'écriture (par exemple l'espagnol et le russe) ou va le faire à l'avenir.
La seule solution gérable dans ce cas est l'utf-8. En effet, il affiche correctement les différents caractères nécessaires à tous les systèmes permettant de ne pas se préoccuper de la nature du contenu lors de son affichage.

2. Mon site ne gère actuellement qu'un système d'écriture (par exemple anglais et français).
C'est ici que la question se pose vraiment, et la réponse est -- comme d'habitude -- ça dépend.

Voici un aperçu des avantages et inconvénients de chaque solution:

1. L'iso-8859-1…
- est simple à employer (c'est souvent celui "par défaut");
- est l'encodage de base de pas mal de langages de programmation;
- est souvent déjà en place;
- est le plus minimal possible (donc souvent le plus rapide);
mais…
- implique d'utiliser des entités HTML dès que l'on souhaite insérer un caractère ne faisant pas partie des quelques 189 disponibles;
- ne gère que les langues occidentales (latines) et est donc faiblement adaptable.

L'iso-8859-15 est dérivé de l'iso-8859-1. Certains caractères jugés inutiles ont été retirés et d'autres ajoutés à leur place (notamment le symbole de l'euro, «€», et le «e dans l'o», «œ»). Il présente globalement les mêmes caractéristiques que l'iso-8859-1 avec cependant le désavantage d'être apparu plus récemment et d'être donc un peu moins répandu/supporté.

2. L'utf-8…
- gère la plupart des langues utilisables et est donc facilement adaptable;
- permet de se passer de la plupart des entités html (les caractères réservés «<», «>» et «&» devront toujours être échappés en < > et &);
- représente «l'aboutissement» de l'encodage, il est peu probable que vous ayez à changer pour un autre suite à une évolution du site;
mais…
- est plus complexe à employer (parfois mal géré ou provocateur de comportements étranges);
- n'est pas toujours bien géré par les différents langages de programmation;
- implique souvent des modifications d'un site déjà établi;
- code les caractères d'une manière plus complexe (et consomme donc souvent légèrement plus de ressources).

Mais la tendance étant à l'internationalisation et à la portabilité, il devient populaire, et de fait son support est assuré de plus en plus efficacement par les éditeurs texte autant que par les langages de programmation et les bases de données (php6 par exemple travaillera par défaut avec de l'utf-8 en interne, mysql permet la gestion des jeux de caractère depuis sa version 4.1).

En conclusion, on peut dire que la bonne partie des désavantages de l'iso-8859-1 proviennent de sa définition

En conclusion, on peut dire qu'une bonne partie des désavantages de l'iso-8859-1 proviennent de sa définition même, alors que ceux de l'utf-8 disparaissent petit-à-petit. La balance pencherait donc plutôt du côté du dernier.

Cela ne fait pas du premier un encodage obsolète pour autant, il faut seulement prendre conscience des limitations que l'on s'impose et savoir si elles ne risquent pas de devenir gênantes à l'avenir. Rappelez-vous qu'il est bien plus simple de commencer directement dans le bon encodage que de changer à posteriori.

Exemples pour illustrer ces propos:

1. Une entreprise est implantée en Espagne et effectue également des ventes en France. Son site est disponible dans ces deux langues et utilise l'iso-8859-1.
Le problème survient l'année suivante, elle souhaite étendre ses activités au Maroc or pour afficher un contenu en arabe elle devra convertir son site en utf-8.
Dans ce cas, il aurait mieux valut commencer dès le départ le site en utf-8.

2. Dupont maintient pour sa famille un petit site personnel ou il affiche les photos de ses week-end avec un script pour les commenter. Le contenu est exclusivement francophone et le restera.
Dans ce cas, l'utilisation de l'iso-8859-1 lui simplifiera probablement la vie alors que l'utilisation de l'utf-8 ne lui apporterai au mieux aucun bénéfice et au pire des problèmes.