Référencement et lettres accentuées

Retour au sommaire de la lettre

La problématique des lettres accentuées est souvent oubliée par les concepteurs de sites web en termes de référencement. Or la plupart des moteurs ne prennent pas en compte de la même façon un mot avec des lettres accentuées ou non. Quels sont les différents types d'encodage ? Lequel utiliser : Iso-8859-1 ou UTF-8 ? Que veulent dire ces termes ? Comment faire en sorte que les spiders comprennent parfaitement vos mots accentués ? Voici quelques pistes de réponse...

Préambule

Dans une stratégie de référencement naturel, le choix de la sémantique reste très important et peut déterminer la réussite d'une campagne de promotion. Faut-il utiliser les mots clés avec accents ou non dans le cadre de cette stratégie ?

L'utilisation des accents n'est, en effet, pas sans poser de problèmes. A la lecture des forums de discussion spécialisés, les questions concernant des « caractères qui s'affichent mal » reviennent très souvent. Les solutions résident dans la compréhension de termes spécifiques comme ASCII, ISO 8859-15 ou encore UTF-8. Quelles sont les difficultés liées à l'encodage ? Comment faire pour connaitre l'encodage d'un site ? Quels sont les différents types d'encodage existant sur le marché ? Quel encodage vaut-il mieux choisir dans le cadre d'une stratégie de référencement ? Cet article tente d'éclaircir la notion d'encodage et de jeux de caractères.

Les moteurs de recherche prennent-ils en compte les accents ?

Lors d'une recherche dans un moteur de recherche, les résultats sur une requête dépendent de nombreux paramètres comme la localisation géographique (basée notamment sur l'adresse IP…), mais aussi de la personnalisation de l'interface de recherche.

Les accents prennent de plus en plus d'importance aux yeux des moteurs de recherche, afin de rendre exhaustifs les résultats issus de la recherche. L'intention de l'usager peut être différente s'il cherche avec ou sans accent.

Lors d'une recherche avec et sans accent sur Google, les requêtes « séjour » et « sejour » donnent un nombre de résultats identique (plus de 40 millions). Cependant, l'organisation des résultats, leur *ranking*, est différent. Que faut-il choisir entre l'un et l'autre ? Il est difficile de donner une réponse ferme. Varier la forme des mots en utilisant des requêtes avec accent et sans accent peut être, au final, la meilleure solution. En effet, les internautes peuvent, invariablement, taper les mots clés de façon accentuées ou non (il semblerait cependant que, selon certaines études, ils aient une prédilection pour les saisir non accentués). ***Petite astuce en passant*** : si vous désirez afficher une version non accentuée d'un mot dans vos pages tout en ne désirant pas montrer ce qui pourrait ressembler à une faute de frappe, affichez le terme en majuscules. Un mot non accentué en capitales passe toujours mieux que sa forme en minuscules... Et la casse (minuscules/majuscules) n'a pas d'importance pour un moteur de recherche...

En prenant en compte ces considérations, le choix d'un mot clé accentué ou non prend toute son importance. Quelles répercussions cela aura–t–il pour un site web ?

Quelles sont les contraintes liées à l'encodage ?

Voici l'exemple d'un texte présent sur une page web :

Un grand choix de séjours
Découvrez tous les séjours en Corse
Partez dès à présent en Corse, nous vous proposons des séjours à Bastia, des séjours à Calvi, toute l'île de beauté à découvrir.

Afin de pallier à tout problème d'affichage entre les navigateurs, il n'est pas rare de voir une page contenant les accents codés en HTML (cf le tableau ci-dessous) :

Numérique	Symbolique	Description	Affichage
&#224;	&agrave;	a grave	à
&#225;	&aacute;	a aigu	á
&#231;	&ccedil;	c cédille	ç
&#232;	&egrave;	e grave	è
&#233;	&eacute;	e aigu	é
&#234;	&ecirc;	e circonflexe	ê
&#235;	&euml;	e tréma	ë
&#236;	&igrave;	i grave	ì
&#242;	&ograve;	o grave	ò
&#251;	&ucirc;	u circonflexe	û
&#252;	&uuml;	u tréma	ü

Cela donne le codage HTML suivant :

```
Un grand choix de s&eacute;jours
D&eacute;couvrez tous les s&eacute;jours en Corse
Partez d&eacute;s &agrave; pr&eacute;sent en Corse, nous vous proposons des s&eacute;jours &agrave; bastia, des s&eacute;jours &agrave; Calvi, toute l' &icirc;le de beaut&eacute; &agrave; d&eacute;couvrir.
```

Cela peut être long et fastidieux de coder systématiquement tous les accents d'un document (même si la plupart des éditeurs HTML actuels effectuent cette opération de façon automatique). Essayons de comprendre un peu mieux les jeux de caractères et de voir si l'encodage des caractères est obligatoire.

Un peu d'explication sur l'encodage

Pour bien comprendre les système d'encodage, il faut, avant toute chose, détailler le fonctionnement d'un ordinateur.

Un ordinateur est une machine fonctionnant grâce à une succession de calculs binaires (0 et 1). Pour les mesurer, on utilise deux unités de mesure : le bit et l'octet (lui-même composé de 8 bits). Tous les fichiers stockés sur un système informatique sont codés sous la forme d'une suite de bits.

Bien sûr, les textes et les caractères (lettres, chiffres, accents, symboles…) n'échappent pas à cette règle. Ils sont codés en utilisant différents formats d'encodage.

Le ***charset*** (*jeu de caractères* en anglais) indique au navigateur dans quel jeu l'usager travaille. Le navigateur peut alors lui restituer correctement la page.

La grande majorité des **charsets** est codée sur un octet (soit 8 bits). Le premier à voir le jour a été l'**ASCII**, codé sous 8 bits. La première version de l'ASCII dispose d'une caractéristique particulière : seuls 7 bits représentent une information. Le petit dernier est un bit de sécurité permettant de détecter une erreur qui aurait pu intervenir lors de la transmission des informations. On ne dispose donc que de 7 bits pour coder le caractère. Cela donnait un total de 128 (soit 2^7) combinaisons possibles. Cependant, ce jeu de caractères limite le nombre de combinaisons. Utilisé avant tout pour échanger des informations en anglais, il pose de gros problèmes lorsqu'il s'agit de coder dans d'autres langues disposant de caractères accentués.

Pour pallier ce problème, le bit de sécurité fut abandonné pour augmenter le nombre de combinaisons possibles. Cela donna un total de 256 combinaisons et permit à de nombreux jeux de caractères de voir le jour et de pouvoir coder dans différentes langues nécessitant des caractères spéciaux comme le grec, le russe ou tout simplement dans les langues européennes occidentales. Grâce à l'Organisation internationale de normalisation (ISO), **l'ISO-8859-1** (ou ISO-Latin1) voit ainsi le jour.

Par la suite, d'autres jeux de caractères considérés comme des variantes du latin1 se sont développés. **L'ISO-8859-15**, appelé aussi latin9, contient de nouveaux symboles comme le sigle € (non présent dans le latin1). Pour compliquer l'affaire, les constructeurs ont décidé d'avoir leur propre jeu de caractères. Ainsi, par exemple, Windows a décidé d'en développer un (appelé Windows-1252).

Pour éviter tout problème de compatibilité entre les jeux d'encodage latin1 et Windows-1252, les navigateurs (du moins sous Windows) ont pris l'habitude de représenter les pages renseignées comme utilisant l'ISO-8859-1 à l'aide du jeu de caractères Windows-1252.

Plus loin avec l'UTF-8

Mais les 256 possibilités rendent impossibles la rédaction d'un document en plusieurs langues. De nouveaux jeux de caractères - codés non plus sur un mais sur deux octets - ont vu le jour comme l'UCS-2. Cela permet de porter le nombre de combinaisons à 65 536.

Ce type de jeu pose de gros problèmes. En effet, il y a la perte d'une caractéristique essentielle lors de l'utilisation de ces jeux étendus : leur compatibilité avec les 128 premiers caractères de la norme ASCII.

Ces caractères sont précieux car ils sont utilisés pour encoder une page HTML sans devoir multiplier les fonctions d'interprétation. Dès lors, il devient difficile de déchiffrer un fichier texte UCS-2 dans un éditeur qui n'est pas prévu pour lui. La solution est alors d'utiliser un nombre d'octets variables pour représenter les caractères. Ce jeu s'appelle **l'UTF-8** (*Unicode Transformation Format*). Il utilise jusqu'à 6 octets pour représenter un caractère permettant de démultiplier les combinaisons. **L'UTF-8** présente la particularité d'être le jeu de caractères par défaut des fichiers XML (*eXtensible Markup Language*, recommandation du W3C).

Quels sont les différents types d'encodage ?

Il existe deux types d'encodage pour tout type de document, l'encodage réel et l'encodage déclaré.

- ⇨ L' **encodage réel** d'un document

Tout document contenant du texte est codé avec un jeu de caractères spécifique. Pour s'assurer d'un bon encodage des caractères, il faut vérifier les paramètres de configuration de l'éditeur HTML utilisé pour savoir quel est l'encodage utilisé par défaut.

- ⇨ L' **encodage déclaré**

Une fois l'encodage de la page connu, il faut s'assurer que l'information soit correctement transmise aux navigateurs web et éviter à tout prix qu'ils se débrouillent tout seul.

Comment les moteurs de recherche traitent-ils l'encodage ?

Les moteurs de recherche se basent sur plusieurs éléments pour déterminer l'encodage de vos données :

- ⇨ D'abord à partir de l'en-tête HTTP "Content-Type" envoyé par le serveur

L'en-tête HTTP est un cours message que le serveur Web communique au navigateur juste avant de lui transmettre le document en lui-même. Ce message donne des informations précises sur le document qui va être envoyé. Par exemple, il peut indiquer au navigateur que le document n'existe plus (404 not found) ou que son adresse a changé (301 permanent redirection) . Il peut aussi servir à préciser l'encodage du document.

L'encodage doit être déclaré dans l'en-tête HTTP envoyé par le serveur avec la page web. On peut vérifier quelles sont les informations envoyées par le serveur en consultant l' en-tête avec des outils spécifiques comme **Rexswain** (disponible à l'adresse suivante :

http://www.rexswain.com/httpview.html). Si l'en-tête ne communique aucun « charset », c'est qu'aucun encodage n'est spécifié par le serveur.

⇨ Avec la balise meta http-equiv="Content-Type" content="...".; charset=..." pour les documents HTML et les documents XHTML traités comme du HTML

On peut effectivement ajouter à chaque page HTML, une balise META qui répète l'information définie par l'entête http. Cette balise META est utile pour les utilisateurs qui enregistrent les pages web en local sur leur ordinateur.

Exemple :

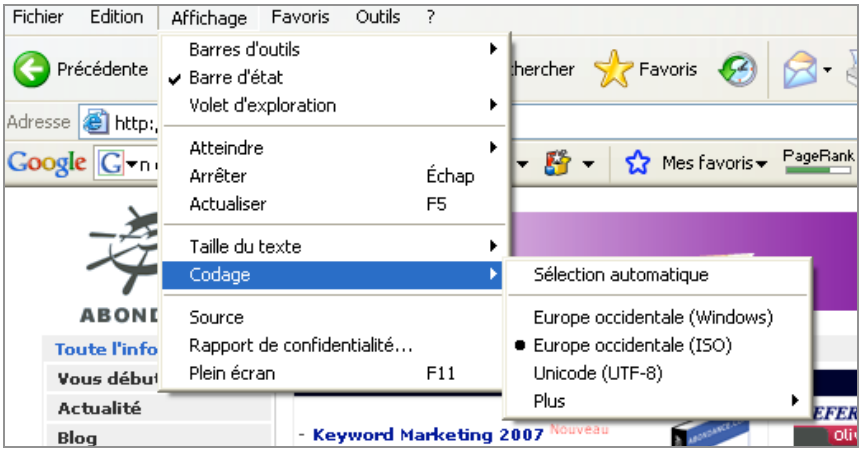
```
<meta http-equiv="Content-Type" content="text/html; charset=utf-8">
```

L'information donnée par la balise META n'est pas prioritaire. En effet, les informations transmises par le serveur prennent le pas systématiquement sur les informations présentes sur la page en elle-même.

Attention ! Il arrive que l'encodage d'une balise META d'un document HTML ne soit pas pris en compte par le navigateur si le serveur dispose d'informations différentes dans l'en-tête http. Il est donc important de procéder à un double encodage au niveau de l'en-tête http et de la page en elle-même. Sans information sur la page, lors de l'enregistrement de la page sur le disque dur d'un ordinateur, le serveur ne transmet plus d'information et risque de rendre le texte illisible.

Comment repérer le codage pour un site web ?

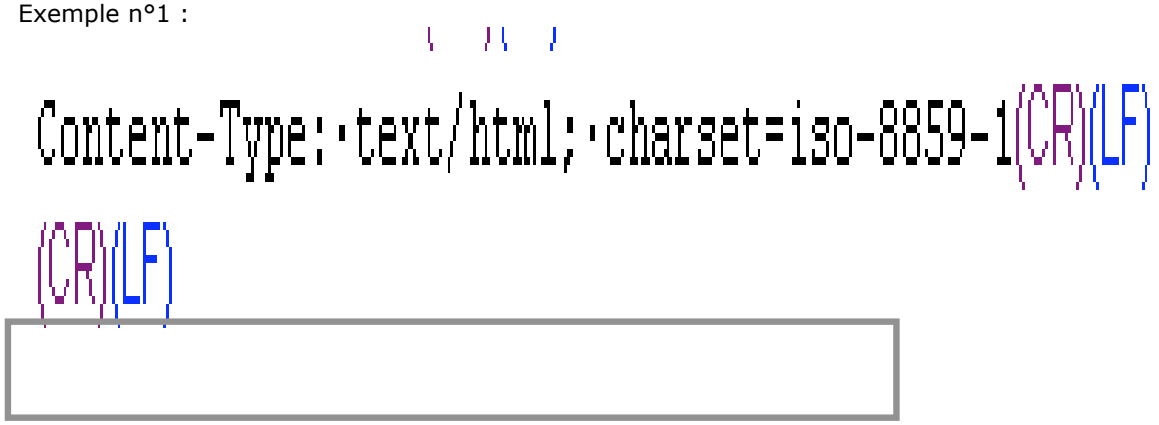
Pour vérifier le type d'encodage, il faut se rendre dans le menu **Affichage** du navigateur et cliquer sur **Codage**.



L'en-tête de la page avec un "lecteur de Header" (par exemple Rexswain) peut également donner des informations sur l'encodage.

Le codage spécifié sur le serveur :

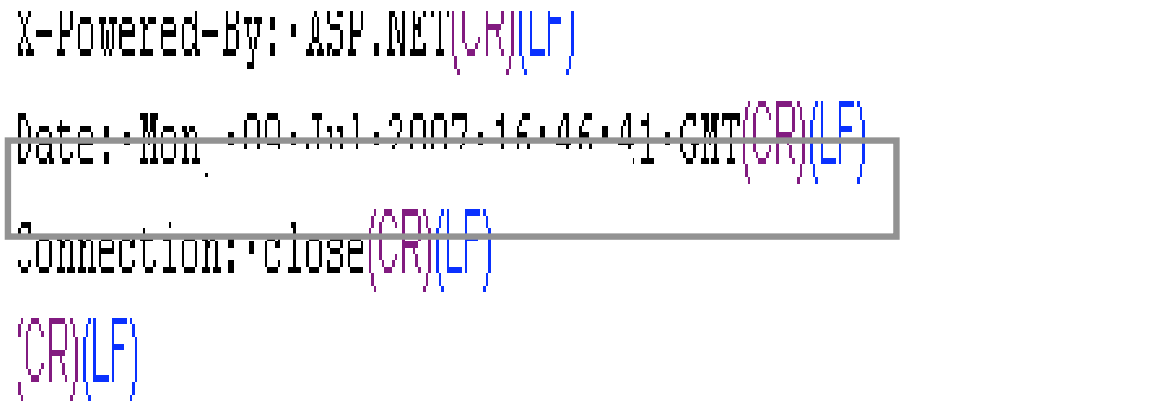
Exemple n°1 :



Dans l'exemple ci–dessus, le codage de la page est "iso-8859-1"

Le codage en dehors du serveur :

Exemple n°2 :



Dans l'exemple, ci–dessus, les informations liées à l'encodage n'existent pas au niveau du serveur. C'est une erreur car leur absence peut avoir des répercussions directes sur l'affichage des pages. Il est donc préférable de les spécifier systématiquement du côté du serveur. Un point à voir d'urgence avec votre webmaster si c'est le cas pour votre site...

Quel type d'encodage faut-il choisir ?

Au final, quel type d'encodage faut-il réellement choisir ? Il faut avant tout savoir à quel(s) public(s) le site est destiné. Dispose–t–il de plusieurs versions linguistiques ? Est–il destiné uniquement à un public européen ?

Si c'est le cas, l'encodage iso-8859-1 est largement suffisant. En revanche, si le site dispose ou envisage d'avoir d'autres versions linguistiques nécessitant un codage particulier comme le japonais, l'encodage **UTF-8** sera beaucoup plus pertinent.

Les problèmes d'encodage sont étroitement liés au bon paramétrage du serveur hébergeant le site. Les *spiders* (robots conçus par les moteurs de recherche pour indexer les pages web dans leurs bases de données) et les navigateurs s'appliquent à lire l'information rencontrée dans l'en-tête HTTP.

Au final, le codage ou non des caractères accentués ne changera rien au référencement si les informations sur le jeu de caractères sont correctement transmises par le serveur. Aucun encodage

n'oblige à utiliser les entités de caractères : tant que l'on reste dans le **même jeu de caractères**, **il n'est pas nécessaire de coder les accents**. Le navigateur ou le robot s'occuperont eux-mêmes de la conversion.

Damien Henckès
Consultant, Nextedia