## L'optimisation de documents Word, PDF et des images

Retour au sommaire de la lettre

Depuis de nombreux mois, les moteurs de recherche n'indexent plus seulement des documents au format HTML, PHP ou autres ASP et CFM. Les formats bureautiques, comme le .doc (Microsoft Word) ou le PDF (Adobe) sont aujourd'hui pris en compte sans problèmes par les moteurs. De plus, il existe de nombreux moteurs de recherche d'images très performants, comme ceux d'AltaVista, d'AllTheWeb ou de Google.

Alors, après tout, pourquoi ne pas tenter d'optimiser des documents de votre site, disponibles sous ces formats, pour qu'ils soient plus facilement trouvés et mieux classés dans les pages de résultats des moteurs ? D'autant plus que les efforts à fournir sont assez faibles (sans spammer pour autant, bien sûr, rappelons-le) pour des résultats souvent excellents...

Dans le cadre de cet article, nous allons essayer de comprendre comment mieux optimiser le référencement de documents images, Word (.doc) et PDF en optimisant les documents eux-mêmes ou leur environnement web (le contenu de la page qui les propose).

### *Images : surtout soigner le nom et le commentaire*

Un fichier image est le plus souvent décrit ainsi dans une page HTML :

```
<IMG SRC="http://www.votresite.com/images/nom-de-l-image.jpg" width="45"
height="52" ALT="commentaires sur l'image">
```

De leur côté, les critères pris en compte par les moteurs pour identifier les images qu'ils proposent dans leurs pages de résultats sont les suivants :

- **Nom de l'image** (ci dessus "nom-de-l-image.jpg"). N'hésitez pas à donner un nom caractéristique à votre image en y incluant des mots clés précis et descriptifs : jacques-chirac.gif, moteur-electricite.jpg, paysage-alpes.jpg, strasbourg.gif, etc.
Les noms d'images n'acceptent pas les caractères accentués, rappelons-le. Mais ce point est peu important pour la plupart des moteurs actuels qui ignorent l'accentuation des mots de toutes façons. Pour séparer les mots, utilisez le tiret (-) ou l'underscore (_), il ne semble pas que les moteurs fassent de disctinction à ce niveau. En revanche, évitez les mots "collés". En d'autres termes, préférez "jacques-chirac.jpg" à "jacqueschirac.jpg". L'utilisation d'un séparateur (tiret ou *underscore*) va "détacher" plusieurs mots dans une même expression et les rendre "réactifs" à une recherche.

- **Format de l'image.** Préférez les formats GIF (.gif) ou JPEG (.jpg). Certains moteurs peuvent indexer d'autres formats (PNG, autres...) mais le "tronc commun" pris en compte par tous les moteurs d'images sont le GIF et le JPEG. Un autre format risquerait d'exclure vos images de l'index.

- **Texte alternatif.** Ce texte, présent dans l'option ALT="..." est très important pour les moteurs de recherche. Il peut être comparé à la balise <TITLE> pour une page web quant à sa fonction et son importance dans le cadre d'un référencement. Nhésitez pas à développer, en une dizaine de mots, ce que représente l'image, en y insérant des mots clés de recherche importants. Exemples :

```
<IMG SRC="http://www.votresite.com/images/jacques-chirac.jpg" width="45"
height="52" ALT="discours de jacques chirac au sommet europeen du
Luxembourg - 10 juillet 2004">

<IMG SRC="http://www.votresite.com/images/cathedrale-strasbourg.gif"
width="100" height="40" ALT="entree ouest de la cathedrale de strasbourg,
alsace, france">
```

Les textes ainsi insérés ne sont pas affichés sur la page (sauf en attendant l'affichage complet de l'image ou, sur certains navigateurs, en passant la souris sur celle-ci). Indiquez-les plutôt en

minuscules non accentuées, notifications comprises par tous les moteurs actuels et notamment Google. De plus, ce texte "alternatif" est prise en compte par bon nombre de moteurs comme critère de pertinence "web". Bien renseigner cette zone aidera donc au bon référencement de vos images comme de vos pages web.

Il existe également deux autres champs, nommé "name" et "title" :

```
<IMG SRC="http://www.votresite.com/images/nom-de-l-image.jpg" width="45"
height="52" NAME="image">

<IMG SRC="http://www.votresite.com/images/nom-de-l-image.jpg" width="45"
height="52" TITLE="Description de l'image">
```

Si l'on est sûr que l'option ALT est prise en compte par les moteurs de recherche, on dispose de moins d'informations sur ces deux champs. Dans l'expectative, vous pouvez indiquer les 3, cela ne pénalisera pas  votre travail :

```
<IMG SRC="http://www.votresite.com/images/nom-de-l-image.jpg" width="45"
height="52" ALT="Description 1 de l'image" NAME="image" TITLE="Description
2 de l'image">
```

Attention cependant à ne pas trop alourdir votre code HTML si votre page contient beaucoup d'images... Privilégiez, dans tous les cas, l'option ALT.

- **Texte du lien**. N'hésitez pas à indiquer, si l'image est affichée en cliquant sur un lien, des mots clés de recherche importants dans le texte du lien pointant sur l'image. Exemple :

```
Cliquez ici pour visualiser <A
href=http://www.votresite.com/images/jacques-chirac.jpg"
target="_blank">une image de Jacques Chirac au sommet européen du
Luxembourg le 10 juillet 2004</A>
```

Ce qui donnera comme résultat :

Cliquez ici pour visualiser une image de Jacques Chirac au sommet européen du Luxembourg le 10 juillet 2004

Vous pouvez également mettre le texte en gras, ce qui donnera au texte constituant le lien un poids encore plus grand par rapport aux critères de pertinence des moteurs :

Cliquez ici pour visualiser **une image de Jacques Chirac au sommet européen du Luxembourg le 10 juillet 2004**

- **Texte "autour de l'image"**. Si vous en avez la possibilité, proposez, proche de l'image dans le code HTML, du texte "visible" (pas de texte en blanc sur fond blanc ou autres joyeusetés ou tentatives de spam svp) explicitant l'image. Exemple :

```
<IMG SRC="http://www.votresite.com/images/cathedrale-strasbourg.gif"
width="100" height="40" ALT="entree ouest de la cathedrale de strasbourg,
alsace, france">Vous pouvez voir, sur l'image ci-contre, une photo de
l'entrée ouest de la cathédrale de Strasbourg (Alsace, France) prise au
grand angle. Son architecture est remarquable... etc. etc.
```

Les moteurs de recherche se servent, pour rechercher leurs images, non seulement du contenu de la balise "Image" (nom de l'image, texte alternatif) mais également de l'environnement de la page. Si éventuellement, le titre et la balise Meta Description de la page contenant l'image peuvent également contenir quelques mots clés descriptifs de celle-ci, cela peut également avoir son importance. Mais ce n'est pas toujours facile...

### *Que faire pour que vos images ne soient pas indexées ?*

Si l'on peut avoir envie de voir les images de son site indexées par Google, on peut également avoir envie.. qu'elles ne le soient pas (pour des raisons de copyright ou autres). Dans ce cas, Google propose une procédure qui vous permettra de ne pas voir vos photographies et autres illustrations indexées par le moteur (voir http://www.google.com/remove.html#images)

Première possibilité : utiliser le fichier robots.txt de votre site (plus d'infos sur ce qu'est le fichier Robots.txt ici : http://docs.abondance.com/robots.html) ainsi :

Pour éviter que l'image spécifique "dogs.jpg" soit indexée, insérez les lignes suivantes dans votre fichier robots.txt :

User-agent: Googlebot-Image
Disallow: /images/dogs.jpg

Pour indiquer à Google qu'aucune image ne doit être "capturée" sur votre site, indiquez les infos suivantes :

User-agent: Googlebot-Image
Disallow: /

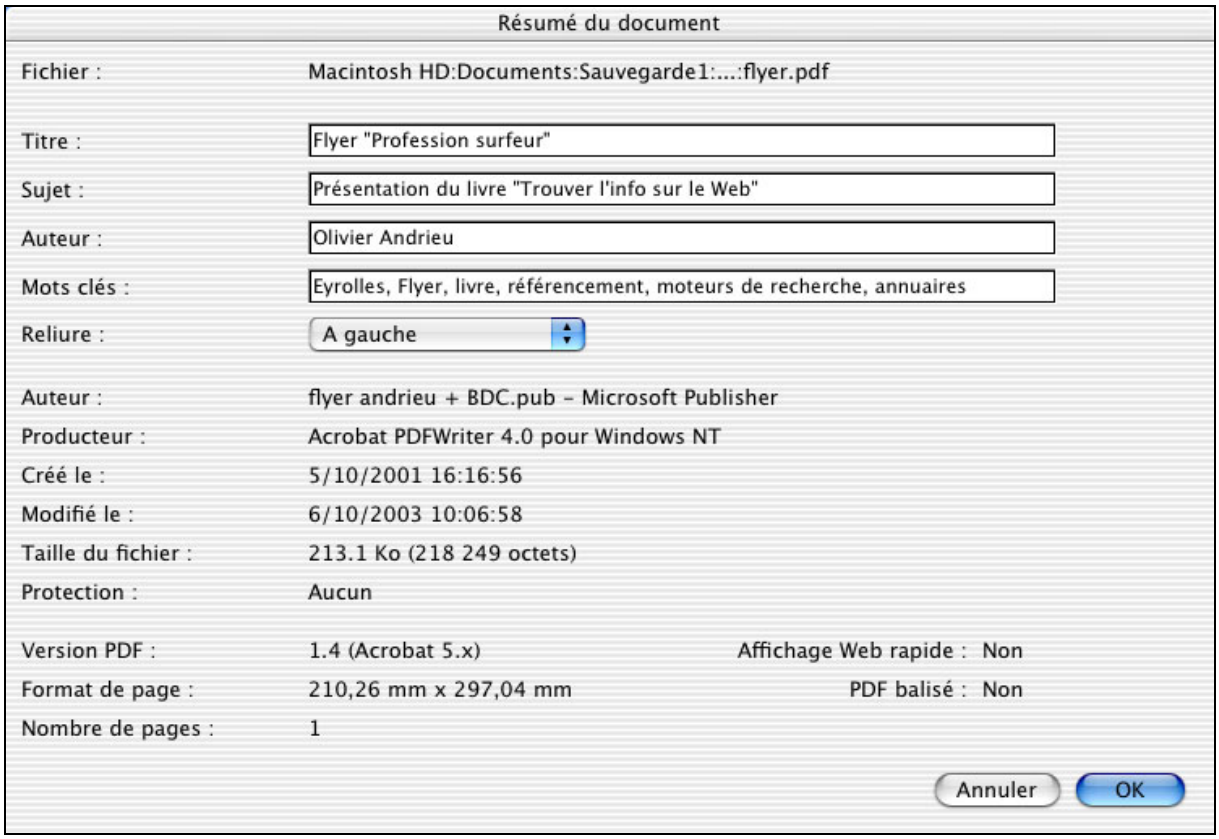
Vous avez également la possibilité d'envoyer une demande écrite à Google (en application du DMCA - "Digital Millenium Copyright Act") selon une procédure indiquée à l'adresse fournie ci-dessus : http://www.google.com/remove.html#images.

### *Documents PDF : soignez le début du texte*

Les moteurs de recherche actuels (Google, AllTheWeb, AltaVista, etc.) indexent des millions de fichiers PDF. Comment faire en sorte que ces fichiers soient optimisés pour un bon positionnement dans les pages de résultats ? Voici quelques indications (la plupart des tests que nous avons effectués l'ont été sur le moteur Google) :

- **Nom du fichier**. Comme pour les images, il est pris en compte. Soignez-le bien et suivez les conseils indiqués ci-dessus pour les images. Ils restent valables pour ce type de fichiers. Exemple : appel-d-offre.pdf, livre-blanc-agriculture.pdf, etc.

- **Meta-données**. Adobe Acrobat permet d'insérer sur chaque document PDF un titre, un sujet (sorte de balise Meta "Description"), le nom de l'auteur et des mots clés (équivalent de la balise Meta "Keywords"), ainsi que d'autres informations. Exemple ci-dessous (sur le logiciel Adobe Acrobat Macintosh):




Ces informations sont très intéressantes car elles permettent de décrire précisemment tout document PDF.

Malheureusement, selon les tests que nous avons faits, ces données ne sont pas prises prises en compte par les moteurs à l'heure actuelle. Deux solutions dans ce cas :

- Elle seront prises en compte à l'avenir (Exalead a notamment commencé à les indexer pour ses recherches sur les formats audio et vidéo et Google, comme nous le verrons par la suite, prend en compte partiellement ces données sur les fichiers Word) et il peut être intéressant de remplir ces champs dès aujourd'hui.

- Soit les moteurs de recherche estiment que ces meta-données subiront la même loi de dégénérescence de leur pertinence que les balises Meta des pages web et ils n'iront pas plus loin que la situation actuelle.

A vous de voir si cela vaut la peine de remplir ces champs, sachant qu'ils ne sont pas lus actuellement par les moteurs mais qu'ils le seront éventuellement d'ici quelques mois... Eventuellement... En tout cas, selon les tests que nous avons effectués, ces champs ne sont quasiment jamais renseignés sur les documents actuellement présents dans les index des moteurs...

- **Texte du fichier.** C'est certainement le champ le plus important, et notamment le premier paragraphe du document qui fournira le titre affiché dans la page de résultats. Google, par exemple, indiquera comme titre de page la première phrase rencontrée affichée en caractères de grande taille. Analysez les résultats retournés par Google sur la requête "référencement filetype:pdf" (http://www.google.fr/search?num=20&hl=fr&ie=ISO-8859-1&newwindow=1&q=r%E9f%E9rencement+filetype%3Apdf&btnG=Recherche+Google&meta=) et vous verrez que les titres des documents affichés sont tous des titres en gros caractères à l'intérieur et toujours au début des documents PDF.

Soignez donc particulièrement le titre de la première page (choix des mots, taille des caractères) de votre document qui fournira au moteur des indices très précieux pour bien classer les fichiers en question.

Tout le texte du fichier PDF est, sinon, indexé : texte des paragraphes, contenus des tableaux, etc. Seules les images ne sont pas prises en compte. Il se peut que le texte en gras ait un "poids" plus fort dans l'algorithme de pertinence de Google, comme pour les pages web, sans certitude cependant, mais avec de fortes présomptions que cela soit le cas. N'hésitez pas, malgré tout, à mettre en exergue les mots importants dans votre document en utilisant le gras et en les insérant dans du texte de liens (Google semble également suivre les liens Web insérés dans les fichiers PDF). En cela, un fichier PDF suit des régles d'optimisation de documents très proches de celles des pages Web.

- **Les en-tête et bas de page**. Ils sont indexés par Google. N'hésitez donc pas à y insérer des phrases très descriptives de ce que propose le document en question.

- La **taille** du document. Pour les pages HTML, Google se limite aux 1024 premiers Kilo-octets (ce qui est déjà beaucoup). Nous n'avons pas d'indications sur le fait qu'il existe ou non une limite de ce type pour les fichiers PDF. Dans le doute, essayez de rester en -dessous du méga-octet, si cela est possible, pour vos fichiers PDF afin d'être sûr qu'ils soient bien indexés (et que vos visiteurs ne perdent pas trop de temps à télécharger vos documents)...

### *Que faire pour que vos fichiers PDF ne soient pas indexées ?*

Même question que pour les images : comment éviter que vos fichiers PDF soient indexés par les moteurs en général et par Google en particulier ? Pour ce faire, insérez les lignes suivantes dans votre fichier Robots.txt :

```
User-agent: Googlebot
Disallow: /*.pdf$
```

Pour indiquer à TOUS les moteurs de ne pas indexer vos fichiers, indiquez plutôt :

```
User-agent: *
Disallow: /*.pdf$
```

Si des fichiers PDF vous appartenant sont déjà présents dans l'index de Google et que vous désirez les enlever, vous devez modifer le fichier robots.txt de votre site comme indiqué ci-dessus puis suivre la procéudre d'urgence indiquée à l'adresse :
http://www.google.com/remove.html#exclude_pages (section "automatic URL removal system")

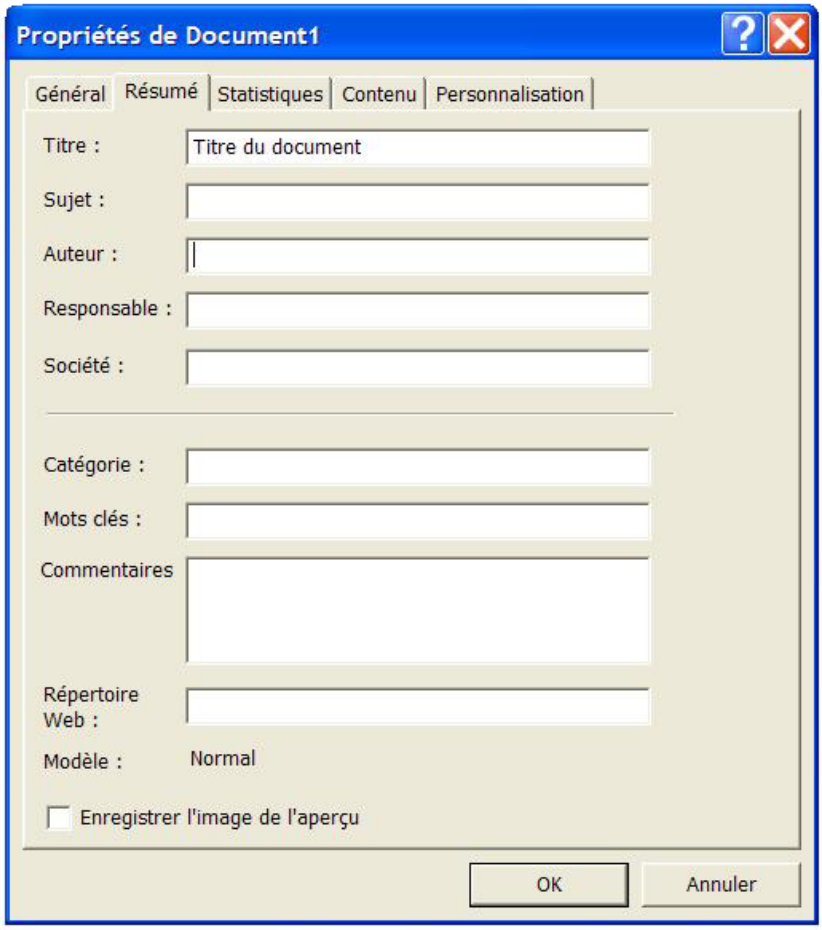
Google viendra ainsi "aspirer" à nouveau votre site pour prendre en compte les modifications que vous lui aurez notifiées dans votre fichier robots.txt.

### *Documents Word : même procédure (ou presque) que pour le PDF*

L'optimisation des fichiers Word (suffixes .doc, .rtf) suit, grosso modo, les mêmes directives que celles énoncées pour les fichiers PDF. A la différence cependant de quelques détails importants. En voici un récapitulatif :

- Le **nom du fichier** est pris en compte selon les mêmes principes que pour les fichiers PDF. Se reporter à ce paragraphe.

- **Meta-données**. Microsoft Word permet, comme Adobe Acrobat, d'assigner des méta-données à un fichier. Voici un exemple (menu "Propriétés") des méta-données d'un fichier Word :




Contrairement aux fichiers PDF, il semblerait selon nos tests que Google prenne en compte dans ses recherche un de ces champs : le **Titre**. Les autres semblent ignorés. Insérez donc, pour chacun de vos documents Word, un titre descriptif précis et contenant des mots clés importants.

- **Texte du fichier** : idem que pour les fichiers PDF. Se reporter à ce paragraphe.

- **En-tête et pieds de page**. Contrairement aux fichiers PDF, ils ne semblent pas pris en compte par Google. Cela ne signifie pas pour autant qu'il faut les bacler (n'oubliez pas que vos fichiers sont également réalisés à destination d'internautes qui vont les lire... Si si...) :-)

- **Taille** du document : idem que pour les fichiers PDF. Se reporter à ce paragraphe.

### *Que faire pour que vos fichiers Word ne soient pas indexées ?*

D'autre part, pour déréférencer un fichier Word de l'index de Google, utilisez la syntaxe suivante dans votre fichier Robots.txt :

```
User-agent: Googlebot
Disallow: /*.doc$
```

ou :

```
User-agent: *
Disallow: /*.doc$
```

Le reste est identique à ce qui est indiqué dans le paragraphe sur les fichiers PDF.

En résumé, que ce soit pour les fichiers PDF ou Word, si vous désirez optimiser ce type de document, prenez surtout bien garde au nom du fichier et au premier paragraphe de texte (le haut de la première page) pour y insérer un titre en grands caractères, contenant les mots importants pour ce document et correspondant exactement au contenu proposé par la suite. Tout devrait bien se passer ensuite pour votre positionnement...