



Mémoire sur le Référencement



Les techniques de référencement naturel d'un site web

Introduction

Depuis plus d'une quinzaine d'années, Internet prend une place considérable et grandissante dans les entreprises et les foyers de tout un chacun.

Actuellement, plusieurs milliards de personnes ont un accès à Internet et, de ce fait, ont la possibilité de mettre en ligne du contenu.

Se pose alors un problème pour le possesseur d'un site Internet : « Comment se faire connaître des internautes? »

En étudiant le comportement de ceux-ci, on constate que le moteur de recherche, par la soumission de requête sur des mots-clés, est l'outil le plus utilisé dans la quête à l'information sur le web.

D'où la nécessité de savoir se faire référencer auprès de celui-ci et de savoir s'y positionner.

Mais alors, qu'est-ce que le référencement?

Trois types de référencement sont à différencier :

- Le référencement mixte: alliant référencement payant et référencement naturel.
- Le référencement payant qui consiste à acheter des mots-clés aux moteurs de recherche afin de positionner plus facilement le site dans la liste des résultats.
- Le référencement naturel : laissant la main « libre » aux moteurs de recherche d'indexer et de positionner le site dans la liste des résultats, par le biais de *robots* (aussi appelés « *spiders* ») appliquant un algorithme visant à scruter la structure et le contenu des pages le constituant.

Le référencement payant n'étant que très récent et l'investissement nécessaire à l'établissement d'une campagne d'achat de mots-clés n'étant bien évidemment pas à la portée de toutes les bourses, c'est tout naturellement qu'ont émergées bon nombre de techniques visant à optimiser le positionnement d'un site et avec elles, le lot habituel d'abus en tout genre.

Maintenant, reste à savoir jusqu'à quel point ces techniques de référencement naturel sont pertinentes.

1- Les techniques basiques de référencement naturel.

1-1) Les Balises

Les premiers moyens mis en œuvre pour référencer un site étaient liés à la simplicité et aux performances limitées des *robots* mais aussi à la capacité de stockage minimale des bases d'indexation.

Ainsi, pour faciliter le travail des robots, les webmasters avaient la possibilité, par l'intermédiaire de balises HTML, de leur fournir des informations relatives aux pages sur lesquelles ils se trouvaient.

Ces balises, appelées balises META, placées entre les balises <HEAD> </HEAD> du document HTML, renseignent sur l'auteur, le contenu mais aussi sur les mots-clés associés au document.

En voici une liste triée par ordre d'importance :

- **<META NAME="publisher" CONTENT="Prénom, NOM">**
Donne un moyen de faire apparaître le nom de l'éditeur ou de la société éditrice dans le code source de la page. Cette balise est actuellement utilisée surtout par les logiciels d'aide à l'édition HTML.
- **<META NAME="Reply-to" CONTENT="ahubert@saliege.com(Alex HUBERT)">**
*Permet de faire apparaître une adresse email de réponse dans le code source de la page
 Par contre, cette adresse sera aussi repérée par les robots mal intentionnés qui parcourent le web à la recherche d'adresses mails.*
- **<META NAME="Copyright" CONTENT="2006 SALIEGE">**
Balise spécifique au Copyright où l'on peut encore mettre le nom de l'entreprise.
- **<META NAME="Category" CONTENT="Association">**
*Permet à un annuaire de savoir dans quelle rubrique il va pouvoir classer le site.
 Il est possible de mettre plusieurs catégories. Elles doivent être écrites en anglais.*
- **<META HTTP-EQUIV="Pragma" CONTENT="no-cache">**
Spécifie au navigateur qu'il doit recharger toute la page avant de l'afficher, et ceci même si elle est déjà disponible par son cache.
- **<META NAME="Identifiant-URL" CONTENT="http://www.saliege.com">**
Cette balise permet simplement de faire apparaître l'URL de la page principale du site dans le code source de la page
- **<META NAME="Revisit-After" CONTENT="7 days">**
*Spécifie avec quelle fréquence le robot peut revenir visiter la page.
 La période spécifiée dans l'attribut CONTENT n'est pas forcément respectée par les robots.*
- **<META HTTP-EQUIV="Expires" CONTENT="Wed, 22 June 2006 15:00:00 GMT">**
*Donne une indication au robot sur la date jusqu'à laquelle il doit garder la page en cache.
 Cette balise sert essentiellement pour des pages dont le contenu est souvent mis à jour.
 Dans le cas contraire, il suffit de remplacer la date par « never ».*
- **<META HTTP-EQUIV="refresh" CONTENT="xx;URL=http://www.google.com">**
*Cette balise redirige automatiquement l'internaute vers la page spécifiée après "xx" secondes.
 Cependant une autre technique sera préférée à cette balise*
- **<META NAME="Description" content="phrase de description">**
*Cette balise, dont la taille ne doit pas excéder 200 caractères, informe les moteurs de recherche sur le contenu de la page.
 Aujourd'hui, nombreux sont les moteurs de recherche qui génèrent dynamiquement un résumé (« snippets ») de la page contenant les mots-clés de la recherche et qui n'utilisent donc plus le contenu de cette balise.*

- **<META NAME="Keywords" content="mot1, exp2, ...">**
*Comme son nom l'indique, cette balise, qui peut contenir jusqu'à 1000 caractères, permet de lister les mots-clés en rapport avec la page.
Il est important de mettre plusieurs synonymes (voire même des antonymes) d'un mot-clé important pour éviter toute ambiguïté.
Les mots-clés peuvent également comporter des fautes d'orthographe volontaires afin de ressortir comme résultat suite à une recherche comportant des erreurs de frappe.
Cette balise, à cause de nombreux abus, est de moins en moins utilisée pour l'indexation des robots.
Il est conseillé de mettre tous les mots-clés en minuscules.
Contrairement à ce que l'on peut penser, répéter les mots-clés ou en mettre trop peut avoir l'effet inverse à l'effet souhaité.
Les spiders comprendront alors cette balise comme une méthode frauduleuse pour augmenter le positionnement et préféreront ne pas indexer la page*
- **<META NAME="Robots" content="xxxx">**
*Cette balise spécifie au robot s'il peut indexer ou non votre page.
En général, elle sert plutôt à exclure des pages du référencement.
Elle sera donc utile dans le cas d'une page en cours de réalisation ou destinée à un usage personnel et qui ne doit pas être accessible.
L'attribut content peut prendre les valeurs suivantes :
- **all** (défaut) : Indique qu'il faut indexer la page et suivre les liens hypertextes.
- **none** : Ne pas indexer la page et de ne pas suivre les liens.
- **index** : La page peut être indexée par les robots.
- **noindex** : Pas d'indexation de la page par les robots.
- **follow** : Donne la permission de suivre les liens hypertextes.
- **nofollow** : Indique qu'il ne faut pas suivre les liens de la page.
- **noimageindex** - Pas d'indexation des images, seulement le texte le sera
- **noimageclick** : Pas d'indexation des liens des images.
- **noarchive** : indique aux robots qu'ils ne doivent pas mettre en cache la page en cours.*



*Si ces balises ont eu leurs heures de gloire lors des débuts du référencement, elles n'ont plus aujourd'hui qu'une influence très limitée.
En effet, l'évolution technologique et les nombreux abus (cf. la partie sur le spamdexing) ont poussé les développeurs en charge des robots à baser leurs indexations sur d'autres critères ceci dans une logique d'offrir à l'utilisateur du moteur de recherche un meilleur service*

Il y a, cependant, d'autres types de balises qui sont plus ou moins prises en compte par les robots.

a. Le titre

GoogleBot, par exemple, donne une place très importante au titre des pages. La balise `<TITLE>` renseigne en général sur le contenu global de la page.

b. Le paragraphe

De la même manière, on sait que la balise `<h1>` (*et ses dérivés h2, h3,...*), qui représente les titres des paragraphes, a également une grande importance.

Malheureusement, comme pour les balises *META*, à cause des abus, les moteurs de recherche n'accordent aujourd'hui plus autant d'importance à ce type de balise qu'ils n'en ont accordé par le passé.

c. La balise ALT des images

Cette balise se présente sous la forme :

```

```

Faisant partie des recommandations du consortium chargé de promouvoir la compatibilité des technologies liées à Internet (le World Wide Web Consortium abrégé W3C), cette balise commence à avoir un certain intérêt quant au référencement des pages.

Elle a pour but de faciliter l'accessibilité au site des personnes non ou mal voyantes.

1-2) Le fichier robots.txt

Une autre technique, utilisant un fichier présent sur la racine du site, est elle aussi à prendre en compte lors du référencement naturel d'un site.

En effet, une fois sur le site, la plupart des *spiders* se met à la recherche du fichier « robots.txt » avant de commencer tout travail d'indexation.

Ce fichier n'a pour seul intérêt que d'exclure des parties du site à l'indexation même si il est également possible, par son intermédiaire, de filtrer les robots autorisés à indexer le site.

- **User-Agent:nom du robot ou *** : permet de préciser le *robot* concerné par les directives qui suivront.

La valeur * désigne tous les *spiders*.

- **Disallow:nom du répertoire ou du fichier** : permet d'indiquer les pages à exclure de l'indexation.

Chaque page ou dossier à exclure doit être sur une ligne à part et doit commencer par /.

La valeur / seule signifie que cela s'appliquera à toutes les pages du site.

NB : La commande **Allow** existe mais est très rarement utilisée.

Prenons l'exemple du fichier robots.txt de Google.fr :

User-agent: * Allow: /searchhistory/ Disallow: /news?output=xhtml& Allow: /news?output=xhtml Disallow: /search Disallow: /groups Disallow: /images Disallow: /catalogs Disallow: /catalogues Disallow: /news Disallow: /nwshp Disallow: /? Disallow: /addurl/image? Disallow: /pagead/ Disallow: /relpage/ Disallow: /relcontent Disallow: /sorry/ Disallow: /imgres Disallow: /keyword/ Disallow: /u/ Disallow: /univ/ Disallow: /cobrand Disallow: /custom Disallow: /advanced_group_search Disallow: /advanced_search Disallow: /googlesite Disallow: /preferences Disallow: /setprefs Disallow: /swr Disallow: /url Disallow: /m? Disallow: /m/search? Disallow: /wml? Disallow: /wml/search? Disallow: /xhtml? Disallow: /xhtml/search? Disallow: /xml? Disallow: /imode? Disallow: /imode/search? Disallow: /jsky? Disallow: /jsky/search? Disallow: /pda? Disallow: /pda/search? Disallow: /sprint_xhtml Disallow: /sprint_wml Disallow: /pqa Disallow: /palm Disallow: /gwt/ Disallow: /purchases Disallow: /hws Disallow: /bsd? Disallow: /linux? Disallow: /mac? Disallow: /microsoft? Disallow: /unclesam? Disallow: /answers/search?q= Disallow: /local? Disallow: /local_url Disallow: /froogle? Disallow: /froogle_	Disallow: /print? Disallow: /books? Disallow: /scholar? Disallow: /complete Disallow: /sponsoredlinks Disallow: /videosearch? Disallow: /videopreview? Disallow: /videoprograminfo? Disallow: /maps? Disallow: /translate? Disallow: /ie? Disallow: /sms/demo? Disallow: /katrina? Disallow: /blogsearch? Disallow: /blogsearch/ Disallow: /blogsearch_feeds Disallow: /advanced_blog_search Disallow: /reader/ Disallow: /uds/ Disallow: /chart? Disallow: /transit? Disallow: /mbd? Disallow: /extern_js/ Disallow: /calendar/feeds/ Disallow: /calendar/ical/ Disallow: /cl2/feeds/ Disallow: /cl2/ical/ Disallow: /coop/directory Disallow: /coop/manage Disallow: /trends? Disallow: /trends/music? Disallow: /notebook/search? Disallow: /music Disallow: /browsersync Disallow: /call Disallow: /archivesearch? Disallow: /archivesearch/url Disallow: /archivesearch/advanced_search Disallow: /base/search? Disallow: /base/reportbadoffer Disallow: /base/s2 Disallow: /urchin_test/ Disallow: /movies? Disallow: /codesearch? Disallow: /codesearch/feeds/search? Disallow: /wapsearch? Disallow: /safebrowsing
--	---

Force est de constater que ce fichier interdit de nombreux dossiers considérés comme sensibles par les membres de l'équipe de développeurs de Google.

Ici, le fichier autorise tous les robots à indexer le site mais restreint l'accès à de nombreux dossiers par le flot de commandes DISALLOW.

Attention : Il ne faut pas laisser de lignes vierges ou blanches (la touche entrée) ; le robot l'interpréterait comme étant la fin du fichier.



*Il est important d'avoir à l'esprit que ce fichier reste accessible à tout le monde (et non pas seulement aux robots).
Il suffit de taper <http://xxx/robots.txt> - en remplaçant xxx par le nom de domaine du site désiré - pour y avoir accès.
De ce fait, il peut être dangereux de spécifier le chemin d'un fichier confidentiel. Préférez les noms de dossiers couplés à une arborescence de dossiers afin que ce fichier ne serve pas de point de départ à une personne mal intentionnée pour attaquer le site.*

1-3) L'optimisation du document

Le référencement naturel d'une page (a fortiori d'un site) passe obligatoirement par un examen de son contenu.

Bien entendu, les robots ne lisent pas intégralement le contenu de la page mais recherchent, entre autre, les occurrences de mots clés (plus précisément la densité d'un mot-clé par rapport aux autres mots présents dans le document html) ou alors, ils établissent une liste de mots-clés en fonction du contenu de la page et des mots qui y sont.

D'où la nécessité d'élaborer une certaine technique de rédaction où les mots-clés pourraient être mis en évidence (notamment par l'usage de la balise).

De plus, l'utilisation de liens internes et externes, de façon pertinente, favorise le positionnement d'un site.

De même, il faut veiller à nommer judicieusement les fichiers afin de faciliter le repérage sur le site au travers de la barre d'adresse. L'utilisation du caractère *underscore* « _ » comme séparateur dans les URL est à proscrire : les robots ne le comprennent pas.

Pour finir, il est nécessaire de créer les conditions pour que les robots des moteurs puissent accéder aux pages en suivant les liens présents sur le site. Il ne faut pas oublier que le robot est censé reproduire le comportement du plus commun des internautes.

Par exemple, une page « Plan du site » peut s'avérer être adaptée dans certains cas de figure.

2- Le cas particuliers

Internet est constitué de documents de types variés et en assurer un référencement performant n'est parfois pas aisé.

Outre l'HTML, les robots sont capables de lire les documents de type texte, Word, Powerpoint, PDF, XML.

Cependant, il est encore difficile pour eux (voire parfois impossible), pour des raisons techniques, de référencer certaines catégories de sites.

2-1) Les sites possédant des jeux de cadres (Frames)

Il est difficile de référencer un site possédant un jeu de cadre car les robots n'aiment guère les frames.

Pourtant, en appliquant quelques manipulations, il est possible d'obtenir un bon référencement pour ce genre de sites.

Généralement lorsque les robots arrivent sur un site possédant des jeux de cadres, ils lisent le contenu des balises `<noframe></noframe>`.

En d'autres termes l'essentiel du site se trouvant entre les balises *frameset* est ignoré.

Pire, ils référencent chacune des pages du jeu de cadres séparément ; ce qui n'est pas le but recherché.

Une solution est de mettre alors les liens vers le contenu des pages entre les balises `<noframe></noframe>`.

Exemple :

```
<noframes>
<body>
<h1>menu Projet Webmaster</h1>
<a href="index.php">Projet Webmaster</a> <br>
Ouverture de la page dans une autre <a href="exemple.html">Publications</a>
</body>
</noframes>
```

Les robots pourront, ainsi, indexer correctement le site puisqu'ils auront accès à des liens, les dirigeant vers du texte.

Le résultat d'une requête concernant le site dans un moteur de recherche permettra d'afficher la page *exemple.html* qui ne comporte pas de jeu de cadres.

Ceci peut être rectifié grâce à un programme Javascript qui permettra d'ouvrir la page contenant les frames et d'afficher le reste du jeu de cadres.

```
<script type="text/JavaScript">
if (parent.frames.length==0) parent.location.href="page_de_frame.html" </script>
```



*Même si ce genre de techniques permet de référencer des sites possédant un jeu de cadres, il est toutefois plus judicieux d'utiliser un code source ne comportant pas de frames.
Cependant, certains robots (notamment Googlebot ont vu leur algorithme corrigé de façon à pouvoir référencer ces sites.*

2-2) Les sites contenant des pages dynamiques

Les pages dynamiques sont caractérisées par des adresses web (URL) complexes, comportant en général un point d'interrogation, éventuellement le caractère « & » ainsi que des noms de variables et des valeurs.

Exemple :

<http://www.monsite.net/documents/articles/article.php?id=12&page=2&rubrique=5>

Le problème est que certains moteurs n'indexent pas les pages ayant des URL de ce type ou alors ils tolèrent au maximum deux paramètres dans l'URL.

Une technique est, en général, de réécrire l'URL, en faisant passer la page dynamique pour un document strictement HTML, de façon à ce que les robots puissent la comprendre.

a- La réécriture d'URL (URL Rewriting)

L'objectif est de transformer une URL du type de celle donnée en exemple en une adresse du type :

<http://www.monsite.net/documents/articles/article-12-2-5.html>

Pour cela, il suffit de spécifier des règles de réécriture dans un fichier « .htaccess » situé sur le serveur, à la racine de chaque dossier contenant les pages dynamiques auxquelles il faudra appliquer ces règles.

Le contenu du fichier *.htaccess* se présente généralement sous la forme suivante :

```
#-----  
# Répertoire : /articles/  
#-----  
  
# Le serveur doit suivre les liens symboliques :  
Options +FollowSymlinks  
  
# Activation du module de réécriture d'URL :  
RewriteEngine on  
  
#-----  
# Règles de réécriture d'URL :  
#-----  
  
# Article sans numéro de page :  
RewriteRule ^article-([0-9+)-([0-9+])\.html$  
  /articles/article.php?id=$1&rubrique=$2 [L]  
  
# Article avec numéro de page :  
RewriteRule ^article-([0-9+)-([0-9+)-([0-9+])\.html  
  /articles/article.php?id=$1&page=$2&rubrique=$3 [L
```

NB :

- Il ne doit pas y avoir de retour chariot sur une ligne de réécriture.
- Chaque fichier *.htaccess* est spécifique à un répertoire.
Il est judicieux dans ce cas d'indiquer en haut de ce fichier l'emplacement du répertoire sur le site.
- Pour insérer des commentaires, il suffit de placer le caractère # en début de ligne.

Explications :

L'instruction *Option +FollowSymlinks* est facultative mais il convient de la conserver.

L'instruction *RewritEngine on* indique que l'on souhaite utiliser le module de réécriture d'URL.

« ^ »

Indique le début de l'URL à récrire.

Ce caractère est facultatif mais il est plus rigoureux de l'utiliser.

« article- »

Ce préfixe qui est une chaîne de caractères quelconque (ici « article ») peut servir à différencier différents schémas d'URL, et il permet à l'internaute de mieux comprendre l'objet de la page.

« () »

Les parenthèses servent à encadrer une variable dont la valeur est récupérée dans la 3ème partie de la ligne.

« [0-9]+ »

Indique que la variable est composée d'un ou plusieurs chiffres.

§ Indique la fin de l'URL à réécrire. Ce caractère est facultatif mais il est plus rigoureux de l'utiliser.

`/articles/`

Cette partie est parfois facultative (cela dépend de la configuration du serveur).

En général il suffit d'indiquer l'emplacement du fichier de manière relative au répertoire dans lequel est situé le fichier `.htaccess`

Sur certains hébergeurs mutualisés, vous devez indiquer le chemin complet vers le fichier, à partir de la racine du site.

« ? » :

Caractère obligatoire précédant la série de variables passées dans l'URL réécrite.

`id=$1`

Indique que la variable nommée `id` prendra la valeur située dans la première paire de parenthèses.

`&`

Caractère utilisé pour séparer 2 variables dans l'URL réécrite.

`rubrique=$2`

Indique que la variable nommée `rubrique` prendra la valeur située dans la deuxième paire de parenthèses.

`[L]`

Drapeau (option) signifiant « Last », indiquant au module de réécriture qu'il doit s'arrêter.

Plus précisément, si l'URL de la page demandée par le visiteur correspond au schéma défini par cette règle, alors le module de réécriture d'URL ne doit pas examiner les autres règles situées dans le reste du fichier `.htaccess`.



La réécriture d'URL suppose au préalable que le serveur sur lequel est installé le site supporte cette procédure.

Il faut aussi identifier un schéma d'URL à partir de leurs formes communes de façon à pouvoir choisir un nouveau format d'URL

Pour finir, il faut, de même, vérifier que tous les liens internes possèdent le même format que celui redéfinit à l'aide du fichier `.htaccess`

Il est bon d'intégrer la gestion de la réécriture d'URL dès la création du site afin d'éviter cette dernière étape.

Cette technique reste encore le moyen le plus efficace pour le référencement naturel des pages dynamiques.

D'autres techniques sont applicables à un certains types de sites dynamiques.

b- Les forums

Les URL des forums comportent un identifiant de session (*session identifier* - abrégé sid) qui permet d'identifier les utilisateurs inscrits du forum.

Ce *sid* n'a rien d'obligatoire dans les URL car cette valeur est déjà stockée sous forme de petits fichiers textes, appelés cookie, conservés par le navigateur internet sur le disque dur de l'internaute.

La session est, quant à elle, présente dans la base de données du site internet.

La valeur dans l'URL n'est utilisée que si aucun cookie n'a été généré.

Il est donc possible de supprimer le sid **dans les URL** (et seulement dans les URL) car les robots étant considérés comme des invités pour le forum, il n'est pas nécessaire de leur fournir le *sid* dans l'URL.

Principe :

La technique consiste en la suppression du sid dans l'URL **pour les invités.**

Pour cela, un certain nombre de manipulations sont à effectuer.

Ces manipulations sont automatisées par des morceaux de code, appelés *mods*.



*Le fonctionnement de cette méthode diffère selon que des **mods** aient été préalablement installés ou non sur le forum.*

c- les blogs

Un blog est un genre particulier de sites internet mais reste tout de même un site internet. Ce qui implique que les règles d'optimisation du contenu peuvent lui être appliqué.

Cependant, le référencement peut être facilité par la présence de flux et notamment par la présence de la balise <link> afin de permettre aux navigateurs (donc aux utilisateurs) ou aux outils spécialisés de détecter le(s) flux associé(s) à votre site.

d- les sites en Flash

La difficulté principale du référencement d'un site en animation Flash réside dans le fait que la balise <body> contenant, a priori, le texte nécessaire à l'indexation de la page par les robots, ne contient qu'un simple fichier au format ShockWave Flash (ayant comme extension .swf) qui leur reste impossible à lire en l'état actuel des choses.

Exemple de code HTML pour un site en Flash :

```
<body>
<object>
  <embed src="votrefichierflash.swf"...></embed>
</object>
</body>
```



La seule solution pour contourner cet obstacle au référencement est la seule utilisation des balises <noembed> ou se trouve le contenu textuel de votre animation ou une énumération des mots clés que vous souhaitez mettre en avant.

NB : Il est fortement conseillé d'éviter l'énumération de mots-clés inutiles (cf. partie sur le spamdexing).

*Rappelons aussi que si le site est composé de plusieurs pages en animation flash, vous n'avez néanmoins qu'une **balise <noembed>** pour toutes les décrire.*

3- Les abus liés au référencement naturel

La course au positionnement a été l'occasion pour certains webmasters d'utiliser des techniques contraire à la l'éthique liée à l'Internet.

3-1) Les pages satellites

Les *pages satellites*, *Miroir*, *Alias* ou *Doorway page* en anglais, sont des pages qui ne servent qu'à fournir aux moteurs de recherche des *mots-clés* qu'ils recherchent. Leur contenu est donc optimisé (par des répétitions, des synonymes, de nombreux liens, etc.) pour un *mot-clé* donné.

Une fois que l'internaute arrive sur un *page satellite*, celle-ci contient généralement un lien qui lui permet de retourner sur la page principale du site.

C'est pour cela que l'on appelle ces pages des *pages satellites*. Elles ne contiennent pas l'information que recherche l'internaute mais servent juste à l'attirer sur le site.

La redirection peut également être automatisée et instantanée (avec la balise `<META refresh>` réglée sur 0 seconde par exemple) afin que cela soit complètement transparent pour l'internaute.

3-2) Le Spamdexing

Le spamdexing est aux moteurs de recherche ce qu'est le spam pour les fournisseurs et les utilisateurs de comptes e-mail.

On définit par le terme **spamdexing** (en français **référencement abusif**) l'ensemble des techniques et méthodes abusives de référencement auprès des moteurs de recherche. Le **spamdexing** consiste ainsi à ajouter des mots-clés sans rapport avec la page et à les dissimuler aux yeux des visiteurs. Ainsi le **spamdexing** est assimilé à du spam dans la mesure où il s'agit d'une tromperie et que cela va à l'encontre de l'intérêt des internautes.

Parmi les techniques souvent considérées comme du *spamdexing* citons les suivantes :

- La mise en place de mots-clés de la même couleur que le fond de page. On parle de **mots-clés invisibles**.
- L'ajout de mots-clés dans les métag tags sans aucun rapport avec la page.
- La répétition de mots-clés (appelé aussi *bourrage de mots-clés*).
- Le détournement de pages web (**pagejacking**).
- L'utilisation abusive d'alias de domaine ou de sous-domaines. La soumission de contenus sous différents noms de domaine n'est pas tolérée par la plupart des moteurs de recherche et des annuaires.
- L'indexation intempestive de pages auprès des moteurs de recherche.

C'est donc naturellement que tout ce qui se rattache au *spamdexing* est prohibé.

3-3) Le Cloaking

Le *cloaking* est une technique qui permet de proposer deux versions d'une même page. L'une pour les internautes et l'autre pour les moteurs de recherche, ceci afin d'améliorer le référencement d'un site.

Cette technique fonctionne sur le même principe que celui des pages satellites, la seule différence est que cette technique est plus aboutie et plus difficile (voire impossible) à détecter par les moteurs actuels.

Principe :

Lorsque vous visitez une page web, votre navigateur commence par envoyer une requête HTTP de ce type au serveur du site:

```
GET:/index.php
HOST:www.actulab.com
USER_AGENT=Mozilla/4.0 (compatible; MSIE 6.0; Windows 98; FREE)
REFERER=http://google.com
REMOTE_ADDR=120.122.121.123
```

GET contient le nom de la page web que vous voulez voir (dans cet exemple index.php).

HOST est le nom du site (dans cet exemple actulab.com).

USER_AGENT est la "signature" de votre navigateur (dans cet exemple Internet Explorer 6, sous Windows 98, connecté par free.fr).

REFERER est le site à partir duquel vous venez (dans cet exemple google.com).

REMOTE_ADDR est votre adresse IP.

Dès réception de cette requête HTTP, le serveur du site génère la page que vous demandez, puis la renvoie vers votre navigateur qui se charge de l'afficher afin que vous puissiez la lire.

De ce fait, il est facile de repérer qui est un robot et qui est un simple internaute.

La technique, utilisant au préalable la variable USER_AGENT, qui fût rapidement démasquée par les moteurs de recherche, s'est perfectionnée depuis, car elle utilise le champ REMOTE_ADDR (contenant l'adresse de protocole internet) et est, pour le moment, indétectable par les robots.



Ces techniques ont conduit à un renforcement des mesures répressives des moteurs de recherche et il convient à l'heure actuelle d'éviter, sous peine de se faire radier de l'index des moteurs qui les détecteraient.

Conclusion

Contrairement à ce que l'on pourrait penser, les techniques de référencement évoluent constamment.

Basées sur des interpolations des webmasters et référenceurs, aucune de ces techniques ne garantie un résultat ayant un réel impact sur le positionnement d'un site.

De plus, les critères de suppression des bases d'indexation étant assez subjectifs, mieux vaut s'assurer de fournir un réel contenu à l'internaute avant de mettre en œuvre toutes ces techniques.

De ce fait, le référencement naturel est une partie non négligeable dans le référencement d'un site internet mais, il conviendra d'y associer d'autres méthodes de référencement et de développement de trafic.

Sources :

www.webrankinfo.com

www.frenchseo.com

<http://www.actulab.com/>

<http://www.sam-mag.com/>

Référencement de votre site web – **Edition ENI**