Olivier **Andrieu**

Réussir son référencement web

2e édition

EYROLLES

Réussir son référencement web

CHEZ LE MÊME ÉDITEUR

I. CANIVET. – **Bien rédiger pour le Web.**
N°12433, 2009, 412 pages.

A. BOUCHER. – **Ergonomie web (2e édition).** *Pour des sites web efficaces.*
N°12479, 2009, 426 pages.

A. BOUCHER. – **Mémento ergonomie web.**
N°12386, 2008, 14 pages.

J. BATTELLE. – **La révolution Google.**
N°11903, 2006, 280 pages.

T. PARISOT. – **Réussir son blog professionnel.**
N°12514, 2009, 300 pages.

S. ROUKINE. – **Améliorer ses taux de conversion web.**
N°12499, 2009, 268 pages.

D. MERCER. – **Réussir son site e-commerce avec osCommerce.**
N°11932, 2007, 446 pages.

M. NEBRA. – **Réussir son site web avec XHTML et CSS (2e édition).**
N°12307, 2008, 316 pages.

E. SLOÏM. – **Mémento Sites web (2e édition).** *Les bonnes pratiques.*
N°12456, 2009, 14 pages.

N. CHU. – **Réussir un projet de site Web (5e édition).**
N°12400, 2008, 246 pages.

S. BORDAGE. – **Conduite de projet Web (4e édition).**
N°12325, 2008, 394 pages.

A. CLARKE. – **Transcender CSS.**
N°12107, 2007, 370 pages.

Olivier **Andrieu**

Réussir son référencement web

2e édition

EYROLLES

ÉDITIONS EYROLLES
61, bd Saint-Germain
75240 Paris Cedex 05
www.editions-eyrolles.com

© Groupe Eyrolles, 2008, 2010, ISBN : 978-2-212-12646-4

Remerciements

Je tiens à remercier ici :

- toutes les personnes qui m'aident depuis plus de quinze ans à suivre le « petit » monde si passionnant des outils de recherche et du référencement. Que de changements dans ce « court » laps de temps et la dernière version de cet ouvrage (janvier 2008)…
- Guillaume Thavaud, de la cellule de veille de la société Brioude Internet, auteur d'articles pour la lettre « Recherche & Référencement » sur le site Abondance.com, pour sa collaboration à de nombreux chapitres de cet ouvrage ;
- Antoine Mussard, de la société VRDCI, et Damien Henckes (consultant indépendant), également pour leur contribution à certains paragraphes et chapitres du livre ;
- Olivier Duffez, lui aussi auteur d'articles pour la lettre « Recherche & Référencement », et qui a contribué aux sections sur l'URL Rewriting et les redirections (chapitre 7) ;
- Christophe, qui se reconnaîtra, pour son aide à la rédaction de certains passages ;
- toutes les personnes que j'aime et qui sont mes « fournisseurs officiels en énergie » ; elles se reconnaîtront également, et particulièrement mon « Ikeagirl » ;
- mes filles, Lorène et Manon, pour leur soutien, ainsi que A. et C.

Olivier Andrieu

Avant-propos

– Je cherche des amis. Qu'est-ce que signifie « apprivoiser » ?
– C'est une chose trop oubliée, dit le renard. Ça signifie « créer des liens... ».
– Créer des liens ?
– Bien sûr, dit le renard. Tu n'es encore pour moi qu'un petit garçon tout semblable à cent mille petits garçons. Et je n'ai pas besoin de toi. Et tu n'as pas besoin de moi non plus. Je ne suis pour toi qu'un renard semblable à cent mille renards. Mais, si tu m'apprivoises, nous aurons besoin l'un de l'autre. Tu seras pour moi unique au monde. Je serai pour toi unique au monde...
– Je commence à comprendre, dit le petit prince. Il y a une fleur... je crois qu'elle m'a apprivoisé...
– C'est possible, dit le renard. On voit sur Terre toutes sortes de choses...
– Oh !, ce n'est pas sur la Terre, dit le petit prince.

Extrait de *Le Petit Prince* d'Antoine de Saint-Exupéry (1943)

Ce livre constituera une aide précieuse, vous apportant de nombreuses informations afin de mieux optimiser, de façon « loyale » et honnête, les pages de votre site pour acquérir une meilleure visibilité dans les résultats des moteurs de recherche. J'espère que cet ouvrage comblera un vide relativement important dans ce domaine. En l'an 2000 est parue la dernière version de mon livre *Créer du trafic sur son site web*, publié aux éditions Eyrolles. Conscient qu'il fallait actualiser les informations contenues dans cet ouvrage, j'ai entrepris l'écriture de *Réussir son référencement web*, dont la première version est parue en janvier 2008. Ce livre a connu, selon les dires de l'éditeur, un grand succès. Tant mieux, cela signifie que la demande est forte sur ce créneau. J'espère sincèrement que cette première mouture a aidé nombre d'entre vous à percer les mystères du référencement et de la visibilité de leur site sur les moteurs de recherche. Au vu des messages que j'ai reçus après sa parution, il semblerait que oui, et je ne peux que remercier ici toutes les personnes qui m'ont fait part de leur expérience depuis ce temps-là.

Le présent livre se veut donc la suite logique de mes précédents ouvrages, bien qu'il ait été fortement mis à jour et enrichi depuis. Le monde du référencement ne connaît que peu de révolutions, mais il est en constante évolution. Aussi, il nous a semblé important de remettre à jour le contenu de ce livre afin de le compléter avec quelques notions essentielles, comme le *duplicate content*, le *TrustRank*, le référencement des images et des vidéos, etc. Tous les chapitres ont été revus pour cette occasion, réactualisés si nécessaire

et surtout complétés avec des informations nouvelles lorsqu'elles nous semblaient importantes.

Contrairement à l'ouvrage *Créer du trafic sur son site web*, qui traitait de la promotion des sites web de façon globale, j'ai voulu recentrer le contenu de ce livre autour du référencement et de ses notions connexes : positionnement, optimisation de site, analyse de l'efficacité d'une stratégie de référencement, etc. Il existe aujourd'hui de nombreux ouvrages sur l'e-marketing, l'affiliation, les communiqués de presse, les liens sponsorisés, la publicité en ligne et les autres manières de faire connaître son site. Aussi, il ne m'a pas semblé pertinent d'expliquer une nouvelle fois ce qui a déjà été écrit par ailleurs. En revanche, quasiment aucun manuel professionnel, concret et précis, n'est sorti depuis quelques années sur le référencement « pur » et l'optimisation de site pour rendre un code HTML réactif par rapport aux critères de pertinence des moteurs de recherche. Il m'a donc paru plus utile d'essayer d'explorer ce domaine de la façon la plus détaillée possible. Et Dieu sait s'il y a des choses à dire ! Vous vous en apercevrez dans les pages qui suivent…

Cet ouvrage ne traitera donc que du référencement de sites web sur les moteurs de recherche. Il est articulé de la façon suivante :

- Les deux premiers chapitres ont pour but de poser les bases du sujet exploré, afin que nous parlions bien tous de la même chose : définitions, enjeux, fonctionnement d'un moteur de recherche, etc.
- Le chapitre 3 présente un début de méthodologie et précise la notion, essentielle aujourd'hui, de « longue traîne », que l'on utilisera abondamment dans le cadre de la recherche de mots-clés et l'évaluation de la qualité d'un référencement.
- Nous aborderons ensuite (chapitres 4 et 5) l'optimisation d'un site en vue de son référencement. Quelles sont les grandes étapes à suivre ? Quels sont les pièges à éviter ? Ces deux chapitres devraient vous fournir un certain nombre d'indications sur les différentes actions à mener afin de rendre votre site le plus réactif possible par rapport aux critères de pertinence *in page* et *off page* des moteurs de recherche.
- Le chapitre 6 sera consacré à la « recherche universelle » et à tous les référencements, multimédias et multisupports, qui en découlent : images, vidéos, actualité, cartographie, PDF, mobile, etc. Un monde en pleine évolution, et parfois même émergent, qu'il faut suivre au jour le jour.
- Le chapitre 7, pour sa part, sera dédié aux contraintes induites par l'optimisation de site web : Flash, frames, JavaScript, notion de duplicate content, etc. Les obstacles sont nombreux dans une stratégie de référencement et nous verrons comment les contourner ou résoudre les problèmes qu'ils posent.
- Une fois votre site optimisé et prêt à être mis en ligne, nous évoquerons au chapitre 8 quelques points qui reviennent de manière récurrente sur le Web : la méthodologie de référencement sur les moteurs, le référencement payant, le format Sitemap, etc.
- À l'inverse du référencement, nous verrons au chapitre 9 comment ne pas être référencé dans les moteurs de recherche. Cela peut être tout aussi important...

- Une fois votre référencement effectué, il vous faudra le suivre (voir chapitre 10) de la meilleure manière possible dans le cadre d'une méthodologie cohérente. Logique…
- Enfin, nous aborderons au chapitre 11 la question de la mise en place pratique de votre stratégie, l'appel à une société spécialisée ou le travail en interne pour un projet de référencement, le choix (parfois complexe) d'un éventuel prestataire, etc.

Ce livre regorge également d'adresses d'outils à utiliser et de sites web à consulter. Véritable mine d'informations, il vous permettra, en tout cas je l'espère, d'offrir une meilleure visibilité à votre contenu au travers des moteurs de recherche. Quand vous aurez multiplié votre trafic par un coefficient que je souhaite le plus fort possible, envoyez-moi un petit message, cela me fera plaisir :-)

Pour terminer cette introduction, il nous semble également important de porter deux faits à votre connaissance :

- Le référencement n'est pas une science exacte. Vous pourrez être d'accord ou non avec certaines affirmations proposées par cet ouvrage. C'est normal. Nous œuvrons dans un domaine empirique où seuls les tests et l'expérience permettent d'avancer. Le dialogue et la concertation font souvent naître de nouvelles idées et permettent de faire avancer les choses. Mais il est difficile d'être toujours totalement sûr de ce que l'on avance dans ce domaine. C'est aussi ce qui en fait sa saveur et son côté passionnant... Le référencement est synonyme de perpétuelle remise en cause de ses acquis ! Une grande école de l'humilité (qui n'est pourtant pas le caractère principal de certains de ses acteurs parfois).
- En juillet 2009, un accord « historique » a été signé entre Yahoo! et Microsoft, impliquant à terme l'abandon de la technologie de recherche de Yahoo! au profit de celle de Microsoft (*http://actu.abondance.com/2009/07/laccord-entre-yahoo-et-microsoft-enfin.html*). Cet accord a été signé au moment où je mettais à jour ce livre (merci aux deux sociétés de n'avoir pas attendu que cette deuxième version soit dans les bacs des libraires pour bouleverser ainsi le marché du *search*). J'ai ainsi pu en tenir compte dans les différents chapitres qui suivent. Cependant, à l'été 2009, il était très complexe de pronostiquer les implications que cet accord aura, dans les mois qui viennent, sur le marché des moteurs de recherche et du référencement. Je ne peux donc que vous inciter à suivre l'actualité (toujours très prolifique) de ce petit monde sur le site Abondance (*http://www.abondance.com/*) afin d'être sûr de ne pas « rater un wagon »...

Mais trêve de bavardages : ce livre se veut le plus pratique possible en vous proposant un maximum d'informations en un minimum de pages. Alors, ne tardez pas et entrez de plain-pied dans le « monde merveilleux – et parfois bien mystérieux – du référencement » !

En vous souhaitant une bonne lecture et… une bonne visibilité sur les moteurs de recherche !

Olivier Andrieu

livre-referencement@abondance.com

Table des matières

CHAPITRE 7

ANNEXE

1

Le référencement aujourd'hui : généralités, définitions

Généralement, le lecteur, avide d'informations, passe assez rapidement le premier chapitre d'un livre dans lequel, pense-t-il, ne seront proposées que des généralités qui lui serviront peu par rapport à ses attentes quotidiennes.

Pourtant, nous ne pouvons que vous inciter à lire assidûment les quelques pages qui suivent. En effet, il est absolument nécessaire, pour bien optimiser son site et réaliser un bon référencement, d'assimiler une certaine somme d'informations au sujet des outils de recherche en général. Vous ne pourrez mettre en place une bonne stratégie de référencement que si vous avez une idée précise de la façon dont fonctionnent les moteurs de recherche et surtout des différents leviers de visibilité qu'ils proposent. Vous pourrez également réagir d'autant plus vite en cas de problème que vous maîtriserez au mieux les méandres parfois complexes de ces outils.

Nous ne pouvons donc que vous engager à lire en détail les pages qui suivent. Elles contiennent des données qu'il vous faudra absolument avoir intégrées avant de continuer votre lecture. Contrairement aux habitudes, nous vous proposons donc de lire ce chapitre plutôt deux fois qu'une, la suite n'en sera que plus limpide…

Référencement versus positionnement

Tout d'abord, dans un livre consacré au référencement, il est nécessaire de bien définir les termes employés, parfois de façon impropre ou erronée, par de nombreux acteurs du domaine (et l'auteur de cet ouvrage en premier, faute avouée…).

Commençons par le commencement avec le terme de « référencement ». Tentons une explication de ce mot au travers d'une analogie avec la grande distribution : lorsque vous allez faire vos courses dans un supermarché, vous vous promenez dans les rayons et y voyez un certain nombre de produits. On dit d'ailleurs, dans le jargon commercial, que ces produits sont « référencés » auprès de la grande surface. En d'autres termes, ils sont « trouvables ». Cependant, ils sont placés parmi des centaines, des milliers d'autres, tous rangés au départ de la même façon dans de nombreux rayons.

Figure 1-1

Dans les grandes surfaces, les produits sont également « référencés »… Source photo : D.R.

Pour mettre en évidence certains d'entre eux, les responsables commerciaux des supermarchés ont alors eu l'idée de les placer au niveau des yeux du consommateur – ou en « tête de gondole », ou encore au niveau des caisses de paiement –, ce qui les rend plus visibles. Certains produits sont alors mis en avant à des endroits stratégiques, beaucoup plus facilement « trouvables » par les clients potentiels. Ils sont ainsi bien « positionnés »… Vous voyez où nous voulons en venir ?

Grande surface et galerie marchande

Pour continuer l'analogie ci-dessus, on peut estimer que la grande surface représente les résultats « naturels » du moteur de recherche, alors qu'une galerie marchande proposera l'équivalent des liens sponsorisés. Un produit peut donc se trouver dans les deux zones d'achat sans qu'il y ait obligatoirement concurrence entre les deux. À retenir...

Pour ce qui est du référencement de votre site web, il en sera de même : lorsque votre site sera « présent » dans les bases de données d'un moteur, on dira qu'il est « référencé ».

C'est une première étape, nécessaire mais pas suffisante, dans le processus de gain de visibilité de votre source d'information. Disons qu'il est « prêt à être vu »… Mais ce référencement devra déjà être optimisé, ce qui représente un vrai travail préliminaire. Nous y reviendrons...

Une phase toute aussi importante sera donc, dans un deuxième temps, de mettre en « tête de gondole » votre site en le positionnant au mieux dans les résultats de recherche pour les mots-clés les plus importants pour votre activité.

Enfin, il faudra une troisième étape, malheureusement souvent négligée, pour vérifier *in fine* que le positionnement a porté ses fruits en évaluant le trafic généré par vos efforts d'optimisation. Croyez-vous que les responsables de supermarchés ne vérifient pas si leurs produits se vendent mieux ou non en fonction de leur emplacement ?

En effet, ce n'est pas parce qu'un produit est placé en tête de gondole qu'il est obligatoirement plus vendu. Tout dépend de l'endroit où se trouve la gondole et du nombre, voire du type, de personnes qui passent devant. En d'autres termes, il ne servira à rien d'être bien positionné sur des mots-clés que personne ne saisit ou sur des moteurs que personne n'utilise… Mais nous reviendrons amplement sur ces notions dans les chapitres qui viennent.

Ne pas mélanger « référencement » et « référencement »…

Le terme de « référencement » est souvent utilisé de manière impropre pour désigner tout le processus d'augmentation de visibilité d'un site web au travers des moteurs de recherche, incluant donc le positionnement, la vérification du trafic généré, etc. Le lecteur voudra bien nous excuser le fait de plonger dans les mêmes travers au sein des pages de ce livre, mais il est parfois difficile de lutter contre certaines habitudes… *Nostra culpa, nostra maxima culpa*…

Liens organiques versus liens sponsorisés

Dans cet ouvrage, nous allons parler de positionnement dans les pages de résultats des moteurs de recherche. Peut-être est-il important de définir clairement sur quelles zones de ces pages de résultats nous allons travailler…

Voici une telle page (voir figure 1-2) pour le moteur de recherche Google France et le mot-clé « référencement » :

- Les zones (1) et (2) sont occupées par des liens sponsorisés, ou liens commerciaux, baptisés AdWords chez Google et qui sont des zones publicitaires payées par des annonceurs selon un système d'enchères avec paiement au clic. Si ces zones ne font pas spécifiquement partie intégrante d'une stratégie de référencement dit « naturel », « organique » ou « traditionnel », le système du lien sponsorisé en est complémentaire. Nous ne l'aborderons cependant pas dans cet ouvrage.
- La zone (3) représente, elle, ce que l'on appelle les liens organiques ou naturels, qui sont fournis par l'algorithme mathématique de pertinence du moteur de recherche. Ils n'ont rien à voir avec la publicité affichée dans les zones ci-dessus. C'est donc dans cet espace que nous allons essayer de vous aider à positionner vos pages.

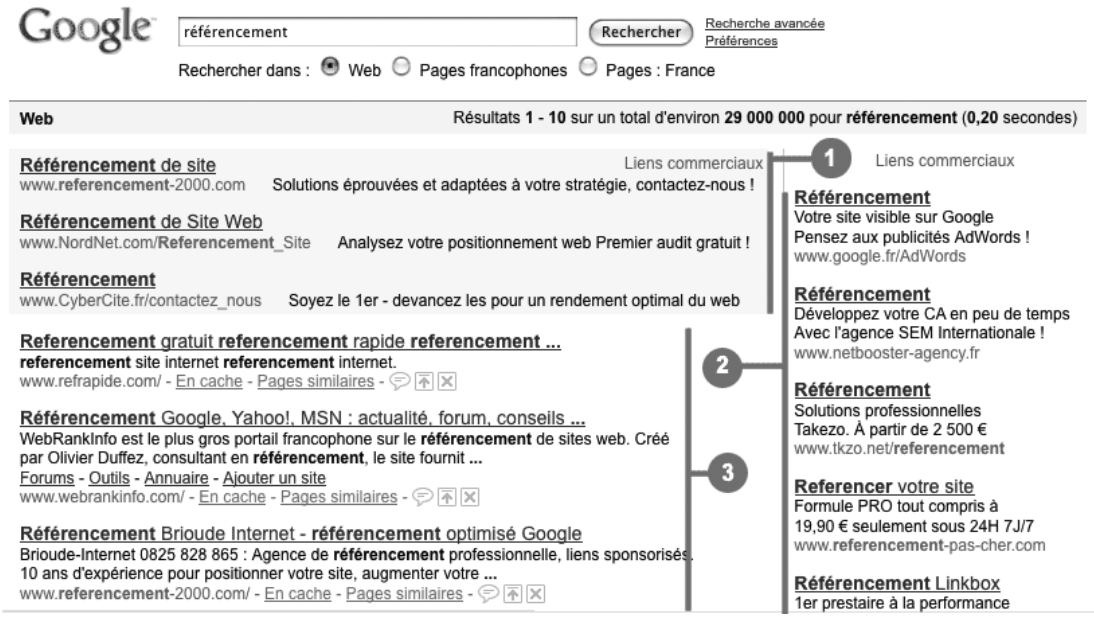


Figure 1-2

Page de résultats sur Google pour le mot-clé « référencement »

Liens « organiques » ou « naturels »

On appelle liens « organiques » (car ils proviennent du cœur même du moteur, ils en représentent la substantifique moelle) ou « naturels » (car aucun processus publicitaire n'intervient dans leur classement), les résultats affichés par le moteur de recherche, la plupart du temps sous la forme d'une liste de 10 pages représentées chacune par un titre, un descriptif et une adresse, en dehors de toute publicité ou promotion pour les services de l'outil de recherche.

Le référencement naturel est indépendant des liens sponsorisés

Il est important de bien comprendre que les deux sources principales d'information dans les pages de résultats des moteurs (liens sponsorisés et liens organiques) sont indépendantes les unes des autres. Être un gros annonceur sur Google ou Yahoo! n'influe donc en rien de façon directe le positionnement de votre site web dans les liens organiques du moteur en question. Heureusement d'ailleurs, car la seule façon d'être pérenne pour un moteur de recherche est de présenter des résultats objectifs et indépendants des budgets publicitaires… Que de « serpents de mer » n'ont pas été imaginés à ce sujet depuis de nombreuses années…

Ceci dit, si par défaut liens naturels et liens sponsorisés sont indépendants, cela ne signifie pas qu'ils n'ont pas d'influence les uns sur les autres. Le fait d'arrêter une importante campagne de liens sponsorisés, par exemple, influera sur le trafic généré sur le site, et donc, la notion de trafic étant aujourd'hui prise en compte par Google dans son algorithme de pertinence, les positionnements en liens naturels peuvent en pâtir… De nombreuses situations de ce type ont vu le jour sur le moteur Google depuis quelques mois…

De la même façon, tous les moteurs de recherche majeurs proposent sur leurs pages de résultats cette dualité liens sponsorisés-liens organiques.

Comme on le voit sur les figures 1-3 et 1-4 (captures effectuées à la mi-2009), sur les moteurs de recherche Yahoo! et Bing, principaux concurrents de Google à l'heure actuelle, les résultats sont également proposés en trois zones distinctes : liens sponsorisés (1 et 2) et liens organiques (3).

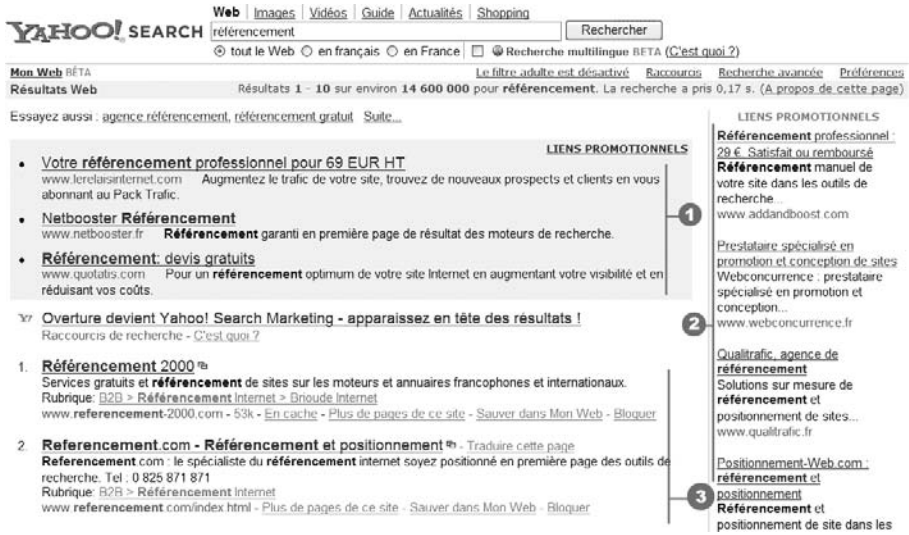


Figure 1-3

Page de résultats sur Yahoo! pour le mot-clé « référencement »

Figure 1-4

Page de résultats sur Microsoft Bing pour le mot-clé « référencement »

La différence se fera, sur un moteur ou l'autre, au niveau de la clarté de différenciation des deux zones : certains outils indiquent clairement ce qui est de la publicité et ce qui ne l'est pas, d'autres semblent moins enclins à marquer une différence nette entre les deux zones, et ce pour des raisons de profits essentiellement : l'internaute, croyant avoir à faire à un lien organique, clique en fait sur une publicité… Reste à estimer le bienfait pour l'internaute de ce type de pratique…

Une stratégie de référencement dit « naturel » ou « traditionnel » aura donc pour vocation de positionner une ou plusieurs page(s) de votre site web dans les meilleurs résultats des liens organiques lorsque les mots-clés importants pour votre activité sont saisis par les internautes.

Attention aux sociétés peu scrupuleuses !

Certaines sociétés vous proposeront parfois comme référencement naturel de l'achat de mots-clés en liens sponsorisés sans le dire expressément. Dans ce cadre, certaines garanties peuvent bien entendu être proposées. Attention aux escrocs qui pullulent dans ce petit monde du référencement (heureusement, il existe également des gens qui travaillent de façon très efficace et professionnelle) !

Les liens sponsorisés apportent-ils une pertinence supplémentaire ?

Aujourd'hui, les liens sponsorisés peuvent amener une pertinence supplémentaire, notamment sur des requêtes à caractère commercial. En effet, ces liens commerciaux sont soumis à une vérification (*a posteriori* ou *a priori*) de la part d'une équipe éditoriale et ils sont censés répondre de la meilleure façon possible à une problématique donnée, mise en lumière par une requête sur un moteur.

La condition *sine qua non* pour que cette pertinence supplémentaire soit réellement efficace sera donc que les procédures de validation des prestataires de liens sponsorisés (Google, Yahoo !, Microsoft…) soient efficaces. Si ce n'est pas le cas, on risque de tomber rapidement dans la gabegie et personne n'aura à y gagner, surtout pas le moteur et la régie. Une raison de plus pour que ces acteurs soignent leurs prestations…

Autre élément important : que les moteurs fassent bien la distinction, dans leurs pages de résultats, entre ce qui est de la publicité et ce qui n'en est pas. Le fournisseur d'accès Internet Free a, par exemple, lancé en mai 2007 une page de résultats associant liens naturels et liens sponsorisés sans qu'il soit facilement possible de les distinguer visuellement (*http://blog.abondance.com/2007/05/une-recherche-pas-trs-free_16.html*). Ce type de stratégie peut tuer le marché de la publicité sponsorisée si elle est appliquée en masse et n'est clairement pas à encourager. Free est d'ailleurs revenu en arrière quelques mois plus tard, conscient de son erreur stratégique…

En bref, on dira que tant que les prestataires de liens sponsorisés et les moteurs de recherche auront comme priorité de servir les internautes avec les meilleurs résultats possibles, tout ira bien et la pertinence s'équilibrera entre liens naturels et liens commerciaux. Toute autre vision du marché risque bien d'être catastrophique pour l'avenir dans un milieu qui reste fragile et sur lequel aucune position n'est acquise et gravée dans le marbre…

Les trois étapes à respecter lors d'un référencement sur un moteur de recherche

Pour mettre en place un référencement réussi, il est nécessaire de passer par plusieurs étapes successives très importantes qui peuvent être représentées par le processus de traitement d'une requête utilisateur par un moteur de recherche.

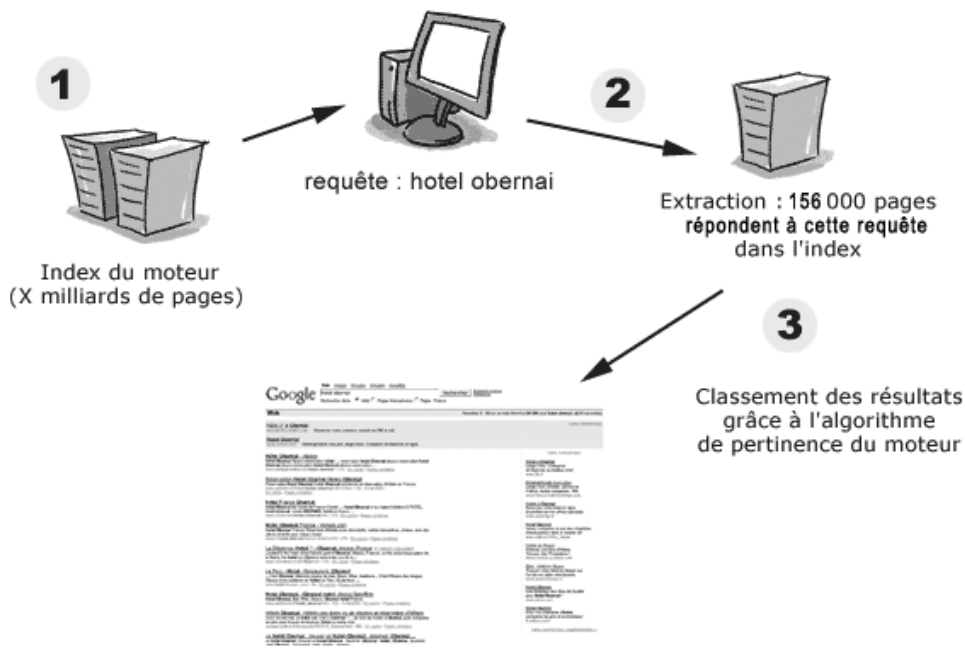


Figure 1-5

Les trois étapes essentielles d'un processus de référencement
Source des dessins : http://www.google.com/intl/en/corporate/tech.html

L'affichage des résultats par un moteur se décompose donc en trois étapes.

1. Extraction depuis son index des pages répondant aux mots de la requête tapée par l'utilisateur (et pas obligatoirement des pages contenant ces mots, comme nous le verrons plus tard).
2. Classement des résultats par pertinence.
3. Affichage.

De la même façon, les étapes à mener dans le cadre d'un *bon* référencement suivront cette même logique.

1. Le moteur se sert d'un index de recherche ; il faudra donc que votre site web soit présent dans cet index. Il s'agit de la phase de référencement. Si votre site propose 100 ou 1 000 pages web, il faudra idéalement qu'elles soient toutes présentes dans l'index du moteur. C'est, bien entendu, une condition *sine qua non* pour qu'elles soient trouvées. Et ce n'est pas sans incidence sur la façon dont votre site doit être pensé lors de sa conception… Nous y reviendrons tout au long de cet ouvrage.

2. L'internaute saisit ensuite un mot-clé (ou une expression contenant plusieurs mots) dans le formulaire proposé par le moteur. Celui-ci extrait de son index général toutes les pages qui contiennent le mot en question (nous verrons, plus loin dans cet ouvrage lorsque nous traiterons du concept de « réputation », que cette affirmation doit être quelque peu révisée). Il faudra donc que vos pages contiennent les mots-clés importants pour votre activité. Cela vous semble évident ? Pourtant, au vu de nombreux sites web que nous ne nommerons pas, cette notion semble bien souvent oubliée… Bref, si vous voulez obtenir une bonne visibilité sur l'expression « hotel obernai », il faudra que les pages que vous désirez voir ressortir sur cette requête contiennent au minimum ces mots.

3. Cependant, la présence de ces mots-clés ne sera pas suffisante. En effet, pour cette expression, Google renvoie plus de 156 000 résultats. Il ne faudra donc pas mettre ces mots n'importe où dans vos pages. Pour faire en sorte que vos documents soient réactifs par rapport aux critères de pertinence des moteurs, et donc qu'ils soient bien positionnés (depuis les 30 premiers résultats jusqu'au « triangle d'or », voir ci-après), il faudra insérer ces termes de recherche dans des « zones chaudes » de vos pages : titre, texte, URL, etc. Nous étudierons tout cela très bientôt.

Pour résumer

Un processus de référencement s'effectue en 3 phases essentielles :

1. Référencement : votre site doit être « trouvable » (« en rayon ») dans l'index du moteur, de la façon la plus complète possible.
2. Identification : une ou plusieurs des pages de votre site doivent se trouver « dans le lot » des pages identifiées car contenant les mots-clés constituant la requête de l'internaute.
3. Positionnement : vos pages doivent être optimisées en fonction des critères de pertinence des moteurs afin d'être classées au mieux dans les pages de résultats pour vos mots-clés choisis au préalable. Pour cela, il faudra (entre autres) placer les termes désirés dans des « zones chaudes » des pages.

Nous espérons que le contenu de ce livre vous aidera à franchir ces trois étapes. C'est en tout cas son ambition !

Positionnement, oui, mais où ?

Depuis que le référencement existe, on a souvent tendance à dire que, sur saisie d'un mot-clé, le but d'un bon positionnement est d'apparaître dans les trois premières pages de résultats des outils de recherche, soit entre la première et la trentième position. Il s'agit effectivement là d'un « contrat » important qu'il ne faudra jamais dépasser. Être classé après la 30e position sur un mot-clé donné équivaut à un trafic quasi nul. En effet, très peu d'internautes dépassent cette fatidique troisième page de résultats lors de leurs recherches. Au-delà, donc, point de salut…

Comment les internautes utilisent-ils les moteurs ?

Voici quelques informations issues d'études sur le comportement des internautes sur les moteurs de recherche.

– Moins d'un tiers des internautes ne consultent que la première page de résultats, et ils sont 45 % à consulter la deuxième et la troisième page ; moins d'un tiers dépassent la quatrième page de résultats.

Source : *Journal du Net – Comment les internautes utilisent les moteurs de recherche* (mai 2009).

http://www.journaldunet.com/ebusiness/le-net/enquete-en-ligne/comment-les-internautes-utilisent-les-moteurs-de-recherche/comment-les-internautes-utilisent-les-moteurs-de-recherche.shtml

– 62 % des utilisateurs de moteurs de recherche cliquent sur un résultat proposé sur la première page de leur moteur favori sans aller plus loin, et ils sont 90 % à ne jamais dépasser la troisième page de résultats.

Source : *iProspect – Search Engine User Behavior Study* (avril 2006).

http://www.iprospect.com/premiumPDFs/WhitePaper_2006_SearchEngineUserBehavior.pdf

– 54 % des internautes ne visualisent que la première page de résultats (19 % vont jusqu'à la deuxième et moins de 10 % à la troisième).

Source : *Impatient Web Searchers Measure Web Sites' Appeal In Seconds* (juin 2003).

http://www.psu.edu/ur/2003/websiteappeal.html

Résumé : *http://actu.abondance.com/2003-27/impatients.html*

Mais être dans les trente premiers résultats n'est pas toujours satisfaisant, loin de là. Vous pouvez être plus exigeant et désirer, par exemple, apparaître uniquement dans les dix premiers liens affichés, soit dans la première page de résultats. Ce qui est évidemment plus difficile selon les mots-clés choisis. Mais il est clair que le trafic généré sera au niveau de votre ambition si vous y arrivez.

Il est encore plus difficile d'être « au-dessus de la ligne de flottaison » (*above the fold* en anglais). Cela signifie que votre lien sera visible dans la fenêtre du navigateur de l'internaute sans que celui-ci ait à utiliser l'ascenseur. Par exemple, en résolution 1 024 × 768 (la plus courante à l'heure actuelle), une page de résultats de Google pour le mot-clé « référencement » apparaît comme sur la figure 1-6.

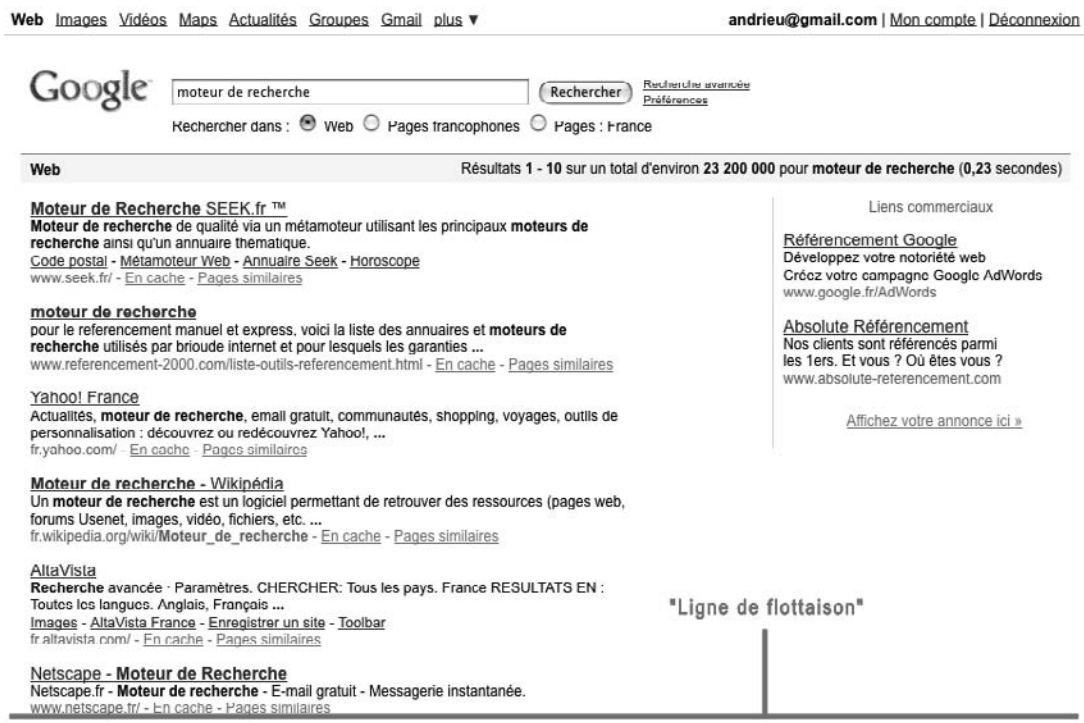


Figure 1-6

Page de résultats de Google en résolution 1 024 × 768

Le navigateur affiche ainsi deux liens sponsorisés (intitulés « Liens commerciaux ») sur la droite de la page mais surtout six liens organiques (issus de l'index de pages web) du moteur sur la partie gauche, visuellement la plus importante.

Le pari sera alors, pour que la situation soit encore meilleure, que vous apparaissiez dans cette zone de six liens. Attention cependant, ce nombre de liens « naturels » peut grandement varier en fonction de nombreux facteurs :

- le type d'information connexe affichée par le moteur selon la requête (plans ou dépêches d'actualité sur Google, exemples d'images répondant à la recherche, etc.) ;
- la présence ou non de liens sponsorisés en position « premium » (jusqu'à trois liens commerciaux peuvent être affichés par Google en début de page sur un fond jaune ou bleu pastel). Comparez les pages de résultats pour un mot-clé comme « référencement » sur Google et Voila (*www.voila.fr*) et vous comprendrez vite que la ligne de flottaison est inaccessible en référencement naturel sur ce dernier moteur ;

- d'une éventuelle proposition de correction orthographique (qui prend plus d'une ligne sur Google) ;
- etc.

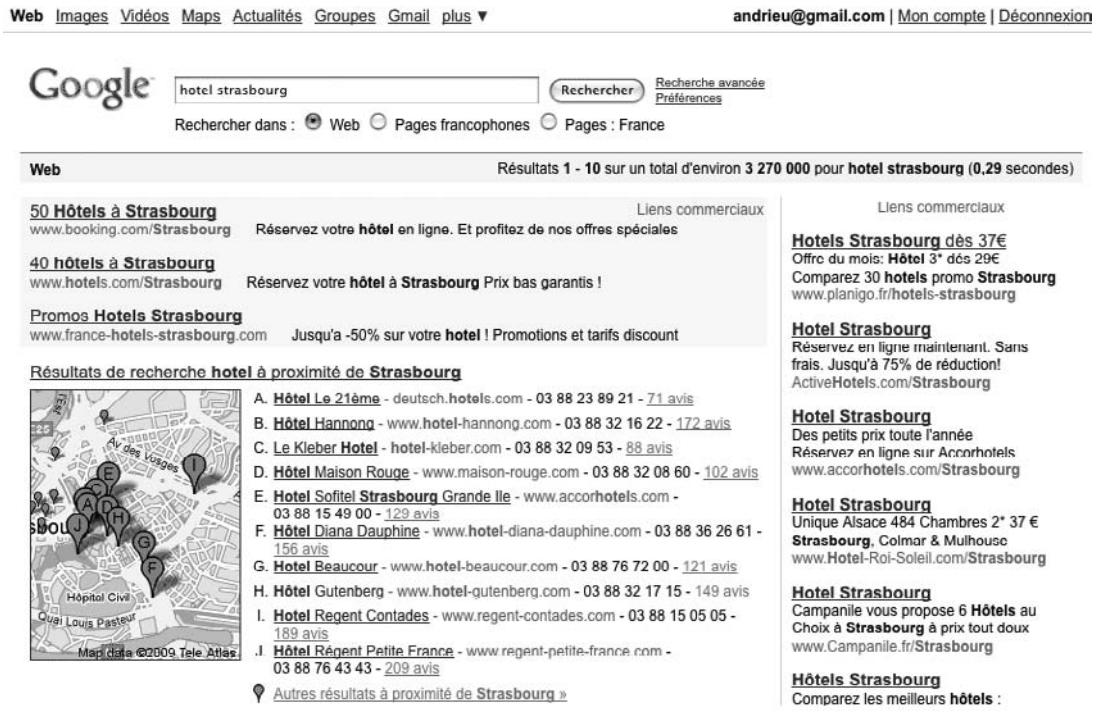


Figure 1-7

Exemple de la requête « hotel strasbourg » sur Google. Dans ce cas, toute la visibilité « au-dessus de la ligne de flottaison » est occupée par les liens commerciaux et les résultats issus de Google Maps. Pour visualiser les liens naturels, il faut « scroller » grâce à l'ascenseur de son navigateur. Le référencement sur Google Maps devient ici primordial (voir chapitre 6).

Résolutions d'écran

Pour connaître les résolutions d'écran le plus souvent utilisées par les internautes, vous pouvez vous servir des données fournies par plusieurs panoramas disponibles sur le Web francophone aux adresses suivantes :

– *http://www.atinternet-institute.com/*
– *http://www.journaldunet.com/chiffres-cles.shtml*

Mais on peut être plus exigeant et tenter de se positionner encore mieux en plaçant un site dans le « triangle d'or » (voir figure 1-8) des pages de résultats. En effet, selon une étude,

menée par les sociétés Enquiro et Dit-It.com en collaboration avec la société EyeTool, spécialisée dans les systèmes d'*eye-tracking* (analyse des mouvements de l'œil), l'œil de l'internaute explore en priorité un « triangle d'or », situé en haut à gauche des pages de résultats de Google. Ainsi, il est possible d'indiquer un taux de visibilité pour chaque rang des liens proposés par le moteur :

- positions 1, 2 et 3 : 100 % ;
- position 4 : 85 % ;
- position 5 : 60 % ;
- positions 6 et 7 : 50 % ;
- positions 8 et 9 : 30 % ;
- position 10 : 20 %.

Figure 1-8

Le triangle d'or de la page de résultats de Google : plus le rouge est vif (l'image originale est en couleurs), plus la zone est lue instinctivement par l'œil des internautes (le trait horizontal épais représente la ligne de flottaison définie dans ce chapitre).

Le triangle d'or de Google

Plus d'informations à propos du « triangle d'or » dans les pages de résultats de Google sont disponibles à cette adresse :

http://www.prweb.com/releases/2005/3/prweb213516.htm

D'ailleurs, on peut même se poser une question : ne vaut-il pas mieux être 11e, donc au-dessus de la ligne de flottaison de la deuxième page de résultats, plutôt que 10e et en dessous de la ligne de flottaison de la première ? Bonne question, effectivement, mais à notre connaissance, aucune étude sérieuse n'a encore été réalisée à ce sujet. Il faut bien avouer qu'elle est quelque peu complexe à mettre en œuvre...

On peut aussi noter que Google a également publié en février 2009 quelques résultats de ses propres études d'*eye-tracking* (*http://googleblog.blogspot.com/2009/02/eye-tracking-studies-more-than-meets.html*).

Figure 1-9

Le triangle d'or de la page de résultats de Google est également présent dans les études internes d'eye-tracking réalisées par le moteur de recherche.

Enfin, signalons également les très intéressantes études, dans le même domaine, effectuées par la société Miratech (*http://www.miratech.fr/newsletter/eye-tracking-google.html*), notamment au sujet de l'interaction entre les liens naturels et commerciaux des pages de résultats de Google.

Figure 1-10

Selon Miratech, le deuxième lien sponsorisé est plus cliqué que le premier, mais le plus fort taux de clic revient quand même au premier lien naturel affiché par Google.

Visibilité dans les résultats des moteurs

Une étude réalisée par un groupe de chercheurs de la Cornell University, aux États-Unis, a tenté en septembre 2005 de comprendre le comportement des internautes lorsqu'ils utilisent un moteur de recherche. Selon ces travaux, plus de 40 % des internautes cliqueraient tout d'abord sur le lien proposé en première position, 16 % sur le deuxième, 10 % sur le troisième et 5 à 6 % sur les liens situés de la quatrième à la sixième place. L'étude est disponible au format PDF à l'adresse suivante : *http://www.cs.cornell.edu/People/tj/publications/joachims_etal_05a.pdf*.

Figure 1-11

Le concept de recherche universelle par Google consiste à mixer, dans les résultats de recherche, des données issues de différentes bases de données, comme ici les images et les vidéos.... et repousse les « liens naturels » vers le bas de la page de résultats.

La recherche universelle de Google

La donne peut encore changer depuis la mise en place du projet « universal search » par Google (*http://actu.abondance.com/2007-20/google-universal-search.php*), qui va amener le célèbre moteur à mixer de plus en plus, dans ses pages de résultats, des liens issus de ses différents index : web, images, actualité, vidéo, cartes, etc. On l'a vu précédemment sur des recherches locales, pour lesquelles Google met en avant Google Maps. Une requête sur le mot-clé « girafe » renvoie également en tête de liste des liens vers des images et des vidéos qui repoussent les liens web organiques encore plus vers le bas de la page et donc « en dessous de la ligne de flottaison ».

Référencement et course à pied...

Bref, si vous désirez réellement être visible dans les pages de résultats des moteurs de la façon la plus optimale possible, ce sont les trois premières places qu'il faudra viser, ce qui n'est pas si simple. D'autant plus que la faisabilité d'un bon positionnement dépendra

de plusieurs critères dont le nombre de résultats, bien sûr, mais également de l'aspect concurrentiel du mot-clé choisi. Pour une requête donnée, la situation sera différente selon qu'une dizaine seulement ou qu'un millier de webmasters ou de référenceurs spécialisés tentent d'être dans ces trois premiers liens. La place n'en sera que plus chère. En effet, la situation ne sera pas la même pour un mot-clé peu concurrentiel, par exemple « matières premières zimbabwe », par rapport à des requêtes comme « hôtel marakkech ». Sur ce type d'expression, de nombreuses sociétés tentent d'être bien positionnées car les enjeux commerciaux qui en dépendent sont énormes. Plus il y aura de concurrence et donc plus on trouvera de candidats motivés sur la ligne de départ, plus la course sera plus rude à gagner. Participer à un marathon dans votre village avec des amateurs offre certainement plus de chances de gagner que le fait de participer au marathon de Paris, où bon nombre de professionnels français et étrangers vont se disputer la victoire. Mais rien n'est impossible !

Le marathon de la première page

La visibilité sur les moteurs de recherche peut donc s'apparenter à un marathon. Il est plus facile de finir dans les dix premiers si on n'est que quelques dizaines au départ plutôt que plusieurs millions car dans ce cas, la difficulté n'en sera que plus accrue, et encore davantage s'il y a des professionnels de la course en face de vous…

Pour continuer cette comparaison entre une course et le référencement, on peut dire qu'en 2009 :

- Il est tout à fait possible d'obtenir des premières pages sur Google en quelques heures sur des mots-clés non concurrentiels, au travers d'une bonne optimisation de vos pages et ce, même si le moteur renvoie plusieurs dizaines de millions de résultats. Si, si...
- En revanche, dès que la requête devient concurrentielle (plus il y a de « pros » de la course qui participent), plus le délai s'allongera. En effet, dans ce cas, l'optimisation seule de la page ne suffira pas. La différence se fera sur l'obtention de « bons » liens, de *backlinks* (liens vers vos pages depuis d'autres sites) efficaces. Et cela prend du temps...

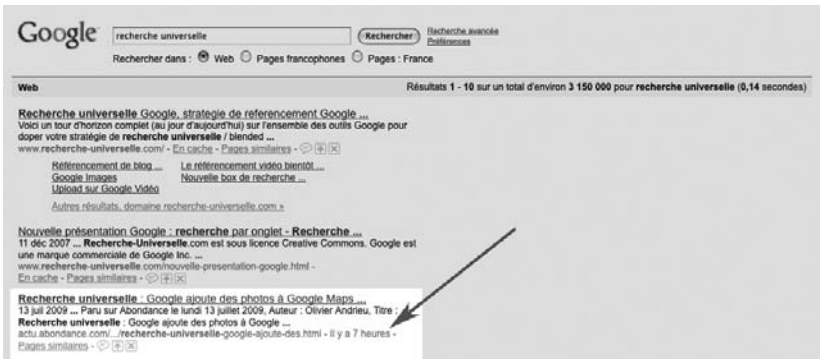


Figure 1-12

Exemple d'une page du site Abondance qui se place en quelques heures sur le podium pour la requête « recherche universelle », qui génère plus de trois millions de résultats sur Google.

Figure 1-13

Autre exemple d'une page du site Abondance sur le podium de la requête « all for good » (4,6 milliards de résultats !). Positionnement obtenu quelques minutes après la mise en ligne de la page pour une requête certes très peu concurrentielle.

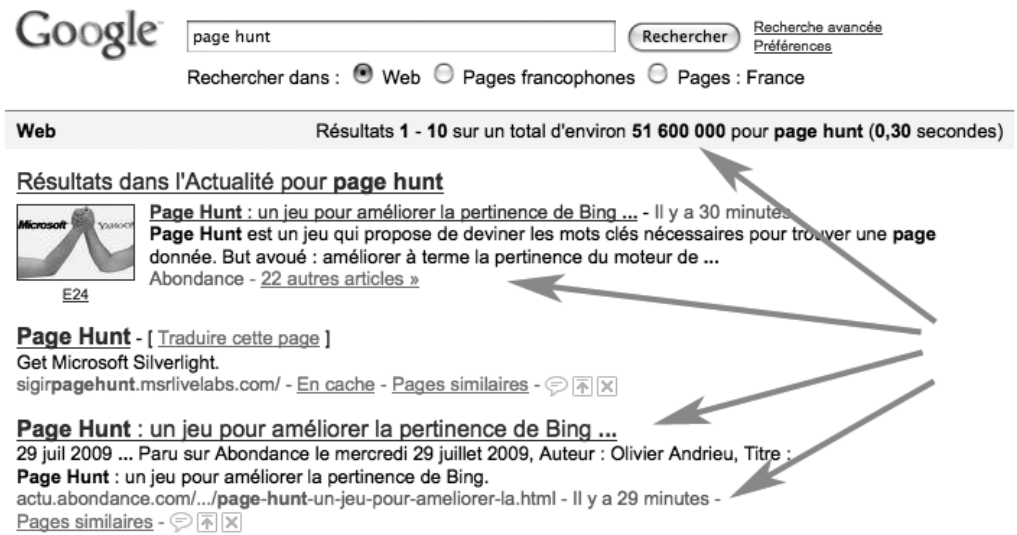


Figure 1-14

Un double positionnement (Google Web et Google News) en moins d'une demi-heure sur la requête « PageHunt » qui génère plus de 50 millions de résultats... mais qui n'est pas très concurrentielle, encore une fois.

Ainsi, vous pouvez tout à fait viser, dans un premier temps, des requêtes non concurrentielles, qui vous permettront d'obtenir très rapidement de bonnes positions et un premier trafic de qualité, même s'il est relativement faible en quantité. En parallèle, il sera possible de travailler à moyen et long terme sur des mots et expressions plus concurrentiels, qui demanderont donc plus de temps pour être profitables. Il faudra certainement une durée d'optimisation plus longue, le temps d'obtenir de bons liens (sachant que ce temps peut être compensé par l'achat de liens sponsorisés en attendant) mais les résultats seront certainement très pérennes...

Pour conclure (provisoirement) cette partie, il est clair que c'est vous qui déciderez jusqu'à quel point vous désirez aller dans le cadre de votre positionnement. Des positions intermédiaires peuvent tout à fait être envisagées : dans le « triangle d'or » pour certains termes, sur la première page pour d'autres et enfin dans les trente premiers résultats pour certaines expressions moins importantes. Mais nous ne pouvons que vous encourager à viser le podium, d'autant plus que ces trois premières places sont souvent assez stables et pérennes dans le temps, contrairement aux suivantes... Dans les chapitres suivants, nous verrons comment faire en sorte que ces ambitions ne soient pas démesurées et comment avoir une idée de la faisabilité et de l'intérêt d'un positionnement sur tel ou tel mot-clé.

Pour résumer

Selon l'ambition de votre site, il vous faudra agir pour obtenir le meilleur positionnement possible (du plus simple au plus complexe) pour vos mots-clés stratégiques :

– dans les 3 premières pages de résultats (les 30 premiers liens organiques) ;

– dans la première page de résultats (les 10 premiers liens) ;

– dans les résultats affichés par le navigateur sans utiliser l'ascenseur (les 4 à 6 premiers liens) ;

– dans le « triangle d'or » du moteur (les 3 premiers liens) de façon idéale.

Le temps pour y parvenir dépendra en grande partie de l'aspect concurrentiel de la requête visée.

Deux écoles : optimisation du site versus pages satellites

Vous l'aurez compris si vous avez lu de façon assidue ce chapitre, la façon dont votre site va être conçu aura une incidence importante sur son classement et donc sa visibilité sur les moteurs de recherche.

Longtemps, deux écoles ont cohabité sur le Web à ce sujet.

- La première consiste à dire qu'il est nécessaire d'optimiser les pages de votre site web : bien étudier leur titre, leur texte, leurs liens, leur URL, éviter les obstacles (voir chapitre 7). C'est donc une optimisation « à la source » du code HTML des pages du site, sans artifice.
- La deuxième école consiste à dire « Développez votre site sans tenir compte des moteurs de recherche ou ne le modifiez pas s'il est déjà en ligne ». Des pages web

spécifiques, appelées « pages alias », « pages satellites », « doorway pages » ou encore « pages fantômes » (toutes ces dénominations désignent le même concept) sont alors créées. Ce sont celles-ci qui seront optimisées pour être bien positionnées sur les moteurs. Ces pages contiennent une redirection (le plus souvent écrite en langage JavaScript mais d'autres voies sont possibles) vers le « vrai » site. Exemple : une page satellite est construite pour l'expression « voyage maroc ». Elle est « optimisée » pour cette requête et contient une redirection vers la page qui traite de ce thème sur le site du client. Si cette page satellite est bien positionnée dans les pages de résultats des moteurs, l'internaute va cliquer dessus et sera donc redirigé vers la « vraie » page du site qui, elle, n'est pas optimisée.

Figure 1-15

Système des pages satellites, utilisées jusqu'en 2006 pour référencer un site web

Ce type de « rustine » a longtemps été utilisée sur le Web, au moins jusqu'en 2005/2006. Disons-le tout de suite : nous ne sommes pas partisans de ces rustines et, plus radicalement, nous vous déconseillons formellement de faire appel à ces pratiques dans le cadre de votre référencement. Allons plus loin : ces pratiques étant aujourd'hui considérées

officiellement comme du *spamdexing* (fraude aux moteurs de recherche) par les moteurs, fuyez une société qui vous proposerait ce type de stratégie !

Définition d'une page satellite

Il est important de bien définir ce qu'est une page satellite : il s'agit d'une page web qui répond à plusieurs critères très spécifiques.

– Elle est créée spécialement pour les moteurs de recherche et le référencement du site web qui les contient. Il s'agit donc d'une page supplémentaire par rapport au site web brut de départ.

– Elle est optimisée pour être réactive par rapport aux critères de pertinence des moteurs de recherche : présence des mots-clés dans le titre, le texte (balises `<h1>` ou autres), l'URL, etc. Ces techniques seront détaillées aux chapitres 4 et 5.

– Elle contient une redirection automatique (le plus souvent en langage JavaScript, mais d'autres méthodes existent) vers une page réelle du site.

Ces trois critères sont essentiels et nécessaires pour bien définir ce qu'est une page satellite. Il suffit que l'un de ces critères ne soit pas rempli pour qu'un document HTML ne puisse être caractérisé par le nom de « satellite ». Depuis que le référencement existe, ce type de page a été mis en place par de nombreux webmasters et référenceurs pour optimiser le référencement d'un site web. Il s'agit d'une solution qui a pu, à une certaine époque, s'avérer intéressante pour certains cas (sites dynamiques, sites techniquement complexes, sites en Flash, etc.). Elle est aujourd'hui clairement à bannir.

Pourquoi faut-il éviter les pages satellites ?

Pourquoi est-ce que nous vous déconseillons d'utiliser des pages satellites ? Nous allons essayer ici de vous donner quelques arguments.

1. Les moteurs de recherche considèrent les pages satellites comme du spamdexing. Relisez les conseils techniques de Google à ce sujet aux adresses suivantes :

   – *http://www.google.com/support/webmasters/bin/answer.py?answer=35769&hl=fr*

   – *http://www.google.com/support/webmasters/bin/answer.py?answer=35291&hl=fr*

   Vous pourrez y lire la phrase suivante : « Une autre pratique illicite consiste à inclure des pages satellites (*doorway*) remplies de mots-clés sur le site du client. Le SEO prétend que cette pratique rendra la page plus pertinente et correspondra à un plus grand nombre de requêtes des internautes. Ce principe est faux, car une page donnée est rarement pertinente pour un grand nombre de mots-clés. Mais ces pratiques peuvent même aller plus loin : très souvent, ces pages satellites contiennent des liens secrets qui pointent vers les sites d'autres clients du SEO. Elles détournent ainsi la popularité du site d'origine vers le SEO et ses autres clients, y compris parfois des sites au contenu douteux ou illégal. » Difficile d'être plus explicite. Lisez attentivement les deux pages proposées aux adresses ci-dessus, elles regorgent de conseils très intéressants.

   D'autres moteurs que Google proposent également dans leur site des *guidelines* assez précises dans ce domaine. Par exemple pour Yahoo! :

   – *http://help.yahoo.com/l/us/yahoo/search/webmaster/*

Et pour Microsoft Bing :

– *http://www.bing.com/webmaster*

2. Malheureusement, de nombreux abus ont été constatés dans ce type de technique, et les moteurs ont, petit à petit, de moins en moins apprécié ce type de page et l'ont fait savoir de façon de plus en plus explicite dans leurs pages d'aide. Pour clarifier ce point, nous avons demandé à Matt Cutts (Google), Olivier Parriche (Yahoo!) et Antoine Alcouffe (MSN) leur opinion sur ce thème. La question était : « Quelle est votre opinion sur les pages satellites, souvent utilisées par les référenceurs ? » Voici leurs réponses, sans langue de bois…

Matt Cutts (porte-parole technique et SEO de Google, dont vous entendrez beaucoup parler dans ces pages) : « Si une page est construite uniquement pour les moteurs de recherche et est remplie de mots-clés, Google considérera qu'il s'agit de spam. Les pages satellites sont le plus souvent créées uniquement pour les moteurs de recherche, ce qui n'est pas réellement utile pour les internautes. Google considère que toute tactique employée avec l'intention de tromper l'internaute par le biais d'artifices est inappropriée. […] La politique qualité de Google (*http://www.google.com/support/webmasters/bin/answer.py?answer=35769#quality*) est en ligne depuis des années et les guidelines que nous fournissons en ligne me semblent clairs. […] Des actions comme le texte ou les liens cachés ou les redirections sournoises sont en totale contradiction avec notre ligne directrice. Plusieurs autres pages, sur notre site, sont disponibles, et notamment une section complète (*http://www.google.com/support/webmasters/bin/answer.py?answer=35291*) traitant des risques encourus dans le cadre de l'optimisation de pages web. Dans de nombreux cas, le fait de lire ces pages répondra à de nombreuses questions pour savoir si une technique est ou non dangereuse. Nous travaillons actuellement à proposer prochainement de nouvelles informations dans ce sens. Nous ne pouvons pas, en revanche, décrire nos algorithmes de détection et expliquer comment nous faisons pour identifier le spam, car ces informations pourraient immédiatement être mises à profit par des fraudeurs… »

Olivier Parriche (Directeur Yahoo! Search France lorsque l'interview a été réalisée) : « En ce qui concerne la lutte contre les techniques de spam, je pense que des préconisations pour un référencement construit sur la base du site lui-même devraient être le réflexe standard. En outre, les pages satellites sont facilement détectables par les moteurs de recherche et peuvent ainsi être repoussées vers le bas des pages de résultats pour préserver la pertinence et l'expérience utilisateur. J'insiste sur le fait qu'il ne s'agit pas de déréférencement… Créer des pages satellites est donc un vain effort. Une fois les actions correctives effectuées, un site spammeur peut reprendre sa place dans les pages de résultats de Yahoo! Search. »

Antoine Alcouffe (chef de produit MSN Search France lorsque l'interview a été réalisée) : « Toute page créée pour les moteurs de recherche dans le but de surpondérer le *ranking* peut être considérée comme du spam par Microsoft Live Search. En effet, certains spams sont aisés à repérer tandis que d'autres le sont beaucoup plus difficilement. Entre une erreur non intentionnelle et la volonté réelle de « jouer » avec le ranking,

il est souvent difficile d'isoler le faux spam du vrai. Afin de faire le meilleur jugement, nous avons établi des règles très fines pour faire face aux pages satellites avec le plus de justesse possible. »

Que faut-il retirer de ces réponses ? Assurément que les pages satellites sont devenues une technique à haut risque, dont il faut désormais envisager le total abandon, puisqu'elles sont clairement considérées comme du spam par les moteurs de recherche.

L'avenir du spamdexing

Il y a fort à parier que les prochaines actions « anti-spamdexing » des moteurs porteront sur le contenu caché à l'intérieur des pages web qu'on voit de plus en plus apparaître sur le réseau. Désormais, les trois techniques d'optimisation à éviter semblent donc être :

– à court terme, les pages satellites ;
– à moyen terme, le texte et les liens cachés (fonctions `visibility:hidden` ou `display:none`, utilisation de balises `noscript` n'ayant aucun rapport avec la balise `script` qu'elles accompagnent, voire sans balise `script` du tout…) dans les pages. Le site du sénateur du Texas a, par exemple, été mis en liste noire par Google en août 2009 pour ces raisons (*http://actu.abondance.com/2009/08/le-site-du-senateur-du-texas-blackliste.html*) ;
– à long terme (mais cette notion reste toute relative sur le Web…), toute manipulation sur des noms de domaine différents pointant, par exemple, vers la même page d'accueil d'un site. Si Google a acquis un statut de *registrar* (société habilitée à délivrer des noms de domaines) (*http://actu.abondance.com/2005-05/google-registrar.php*), ce n'est pas obligatoirement pour vendre ces noms de domaines, mais plutôt pour avoir un accès plus complet aux données les concernant.

La conclusion nous semble donc évidente : n'utilisez plus les pages satellites dans les mois à venir. L'unique bon référencement est bien celui qui repose sur l'optimisation à la source du site web lui-même, sans information cachée dans les pages et sans page satellite. Ceux qui seront *blacklistés* ou pénalisés dans un proche avenir pour avoir abusé des pages satellites ne pourront pas dire qu'ils n'ont pas été prévenus…

Il ne reste plus alors qu'aux webmasters à envisager un référencement basé sur l'optimisation à la source du site web lui-même, sans information cachée dans les pages et sans page satellite. Ce qui donne d'ailleurs d'excellents résultats, comme vous allez pouvoir vous en rendre compte en lisant cet ouvrage. Les webmasters n'adoptant pas cette technique devront inventer de nouvelles possibilités de contourner les algorithmes des moteurs. Et le jeu des gendarmes et des voleurs continuera alors… Jusqu'à quand ? Certainement la pénalisation du site…

Ces arguments nous semblent suffisants pour bien réfléchir avant de mettre en place une stratégie basée sur les pages satellites. La situation, à notre avis, est similaire à celle des balises meta, il y a quelques années de cela, dont le déclin s'est effectué en trois étapes.

1. Les balises meta (`description` et `keywords`, voir chapitre 4) représentaient une solution idéale pour les moteurs de recherche puisqu'elles permettaient de fournir à ces derniers des informations sur le contenu des pages de façon transparente.

Les pages satellites depuis longtemps sur la sellette

Concernant les pages satellites, je vous invite à lire trois *posts* publiés sur le blog du site *Abondance.com*. Ces articles démontrent bien que cela fait de longs mois que la question de la pérennité des pages satellites est posée :

– *Les pages satellites bientôt sur orbite ?* (septembre 2005)

http://blog.abondance.com/2005_09_01_blog-abondance_archive.html

– *Google, SEO et ses bas…* (décembre 2005)

http://blog.abondance.com/2005/12/google-seo-et-ses-bas.html

– *Vers une (r)évolution culturelle du référencement ?* (février 2006)

http://blog.abondance.com/2006/02/vers-une-rvolution-culturelle-du.html

Les pages satellites permettaient également de pallier des problèmes techniques pouvant bloquer un référencement (Flash, sites trop graphiques ou dynamiques, etc.).

2. Certains webmasters sont allés trop loin et on réellement fait n'importe quoi avec les balises meta, les « truffant » notamment de mots-clés n'ayant aucun rapport avec le contenu du site ou indiquant de nombreuses occurrences d'un même terme, etc. Les pages satellites ont connu les mêmes abus, certains référenceurs y ajoutant de façon cachée des liens vers leur propre site, voire, encore pire, vers les sites d'autres clients histoire d'en améliorer la popularité…

3. Que s'est-il passé à l'époque pour les balises meta ? Les moteurs en ont eu assez des excès de certains webmasters et ont, dans leur immense majorité, décidé de ne plus prendre en compte ces champs dans leur algorithme de pertinence (nous y reviendrons...). Les webmasters qui, eux, les géraient de façon propre en ont fait les frais. Les moteurs sont clairs aujourd'hui sur ce point : les pages satellites sont du spamdexing et doivent être abandonnées.

Il est très important de bien comprendre que la page satellite ne doit pas obligatoirement être considérée comme un délit en soi. Il fut une époque où cela marchait très bien et où la communication à ce sujet par les moteurs de recherche était plus que floue… Mais la multiplication des abus a amené les moteurs à supprimer ce type de page de leurs index. Ceux qui auront basé toute leur stratégie de référencement sur ces rustines et auront certainement payé très cher pour cela en seront alors pour leurs frais. Cela est devenu aussi inutile que de baser tout son référencement sur l'usage des balises meta `keywords`, totalement inefficaces aujourd'hui.

Mais ne nous y trompons pas : fin 2009, la majorité des sociétés de référencement en France n'utilise plus les pages satellites comme système de référencement/positionnement et base plutôt leur stratégie sur le conseil et l'optimisation des pages existantes du site voire la création de véritables pages de contenu optimisées. Là est la véritable voie de réflexion pour l'avenir. Comment peut-on penser qu'une société de référencement sérieuse utilise encore des stratégies basées autour des pages satellites en 2009 ?

Mais pour cela, il faut absolument que tous les acteurs de la chaîne de la création de site web soient clairement persuadés que chacun doit et peut avancer dans le même sens.

- Le propriétaire d'un site web doit être conscient que, pour obtenir une bonne visibilité sur les moteurs de recherche, certaines concessions, notamment techniques, doivent être faites (moins de Flash, de JavaScript, plus de contenu textuel, etc.).
- Le créateur du site web (*web agency*) doit être formé aux techniques d'optimisation de site et conseiller, en partenariat avec le référenceur, de façon honnête, le client sur ce qui est possible et ce qui ne l'est pas.
- Le référenceur doit garantir à son client la non-utilisation de tout procédé interdit. Il est possible d'obtenir une excellente visibilité sur un moteur de recherche en mettant en place une optimisation propre, loyale, honnête et pérenne, et sans artifice ni rustine à durée de vie limitée… Le tout est surtout de partir d'une base la plus « saine » possible, c'est-à-dire d'un site web préparé dès le départ pour le référencement.

Alors, si tout le monde y met du sien (et on peut s'apercevoir que, petit à petit, les moteurs de recherche se joignent au cortège en communiquant de plus en plus sur ces domaines), peut-être évitera-t-on le type de problème qu'on voit apparaître aujourd'hui avec le *blacklistage* (mise en liste noire, voir chapitre 8) de certains sites majeurs par les moteurs. Mais cela passera nécessairement par une révolution culturelle et la remise en question d'une certaine approche du référencement. Les sociétés françaises qui se sont perdues dans la voie de la page satellite, ou d'autres techniques, sont-elles prêtes à cette révolution qui n'est peut-être d'ailleurs qu'une évolution ? L'avenir le dira.

Vous l'aurez peut-être compris, l'auteur de ce livre est un fervent adepte de l'optimisation *in situ* des pages constituant un site web. Ce sont ces pratiques d'optimisation loyales, aujourd'hui éprouvées, efficaces, et extrêmement pérennes, que nous allons vous expliquer dans la suite de cet ouvrage.

Ce chapitre introductif est maintenant terminé. Si vous l'avez lu avec assiduité, vous devez logiquement être tout à fait au point sur la stratégie globale à adopter pour optimiser vos pages et donner à votre site une visibilité optimale sur les moteurs de recherche. Vous devez donc être prêts à relever vos manches et à mettre les mains dans le cambouis (terme noble selon nous) ! Cela tombe bien, c'est aux chapitres suivants que cela se passe…

2

Fonctionnement des outils de recherche

Avant d'y référencer votre site, savez-vous ce que l'outil de recherche que vous utilisez au quotidien a « dans le ventre » ? La réponse à cette question n'est pas si évidente. En effet, bien que les moteurs de recherche tels que Google, Yahoo! ou encore Bing semblent très simples d'utilisation, leur fonctionnement « sous le capot » est en réalité très complexe et élaboré. Nous vous proposons dans ce chapitre une analyse globale du fonctionnement des moteurs de recherche ainsi que des processus qui sont mis en œuvre pour traiter les documents, stocker les informations les concernant et restituer des résultats suite aux requêtes des utilisateurs. Bien maîtriser le fonctionnement d'un outil de recherche permet de mieux appréhender le référencement et l'optimisation de son site.

Comment fonctionne un moteur de recherche ?

Un moteur de recherche est un ensemble de logiciels parcourant le Web puis indexant automatiquement les pages visitées. Quatre étapes sont indispensables à son fonctionnement :

- la collecte d'informations (ou *crawl*) grâce à des robots (ou *spiders* ou encore *crawlers*) ;
- l'indexation des données collectées et la constitution d'une base de données de documents nommée « index » ;
- le traitement des requêtes, avec tout particulièrement un système d'interrogation de l'index et de classement des résultats en fonction de critères de pertinence suite à la saisie de mots-clés par l'utilisateur ;
- la restitution des résultats identifiés, dans ce que l'on appelle communément des SERP (*Search Engine Result Pages*) ou pages de résultats, le plus souvent présentées sous la forme d'une liste de dix liens affichés les uns en dessous des autres.

Comme nous l'avons vu dans les pages précédentes, les pages de résultats des moteurs de recherche affichent deux principaux types de contenu : les liens « organiques » ou « naturels », obtenus grâce au crawl du Web et les liens sponsorisés, ou liens commerciaux.

Nous allons nous concentrer ici sur les techniques utilisées par les moteurs de recherche pour indexer et retrouver des liens naturels. Nous n'aborderons pas le traitement spécifique des liens sponsorisés, qui ne font pas partie des objectifs de cet ouvrage.

Technologies utilisées par les principaux portails de recherche

En dehors des trois leaders du marché (Google, Yahoo! et Microsoft Bing), de nombreux moteurs n'utilisent pas leurs propres technologies de recherche mais sous-traitent cette partie auprès de grands moteurs (à terme, ce sera d'ailleurs également le cas de Yahoo!, qui a prévu d'utiliser Bing, la technologie de Microsoft, pour son moteur de recherche, suite à la signature d'un accord entre les deux sociétés fin juillet 2009, voir la page suivante : *http://actu.abondance.com/2009/07/laccord-entre-yahoo-et-microsoft-enfin.html*). En fait, il n'existe que peu de « fournisseurs de technologie » sur le marché : Google, Yahoo! (pour peu de temps encore), Microsoft, Ask.com et Gigablast sont les principaux aux États-Unis, comme sur le plan mondial. En France, les acteurs majeurs sont Exalead, Orange/Voila, qui côtoient d'autres noms moins connus et bien sûr les trois leaders Google, Yahoo! et Microsoft. Voici un récapitulatif des technologies utilisées par les différents portails de recherche en 2009 (au moment où ces lignes étaient écrites, de nombreuses questions se posaient encore sur l'avenir des technologies de recherche de Yahoo!, suite à l'accord avec Microsoft, et leur implantation à l'avenir sur des outils comme AltaVista).

Tableau 2-1 Technologies de recherche utilisées par les principaux portails de recherche francophones en 2009

Sites web / Technologies de recherche	Google	Yahoo!	Bing	Exalead	Ask.com	Voila
Google	X					
Yahoo!		X (2009)	X (> 2009)			
Bing			X			
MSN			X			
Orange/Voila						X
Ask.com France					X	
AOL.fr	X					
Free	X					
Neuf/SFR	X					
Bouygues Telecom	X					
Alice	X					
Exalead				X		
Ujiko (Kartoo)		X				

Tableau 2-2 Technologies de recherche utilisées par les principaux portails de recherche anglophones en 2009

Sites web / Technologies de recherche	Google	Yahoo!	Bing	Exalead	Ask.com
Google	X				
Yahoo!		X (2009)	X (> 2009)		
Bing			X		
MSN			X		
AllTheWeb		X (2009)	X ? (> 2009)		
AltaVista		X (2009)	X ? (> 2009)		
Ask.com			X ? (> 2009)		X
Exalead				X	
Hotbot		X (2009)	X		

Mise à jour

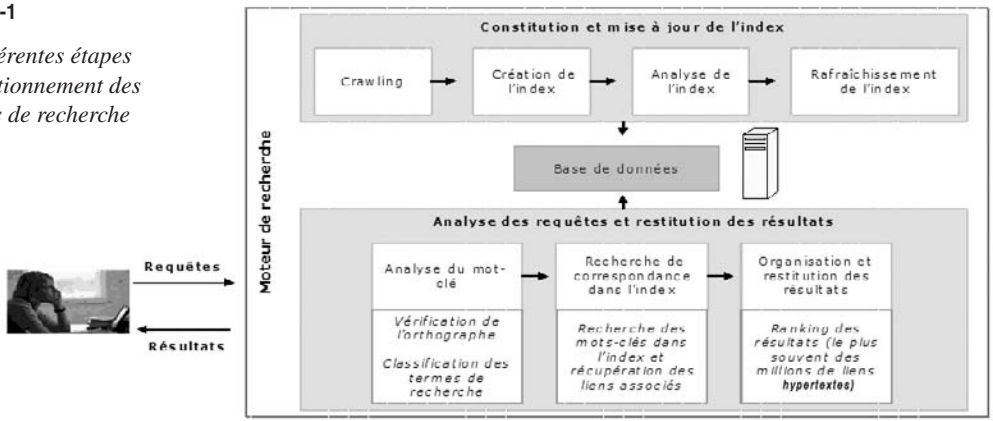
Les données de ce tableau, valables à la fin 2009, peuvent fluctuer en fonction des contrats signés d'une année sur l'autre. Une mise à jour de ces informations est disponible à l'adresse suivante : *http://docs.abondance.com/portails.html*.

Principe de fonctionnement d'un moteur de recherche

Plusieurs étapes sont nécessaires pour le bon fonctionnement d'un moteur de recherche : dans un premier temps, des robots explorent le Web de lien en lien et récupèrent des informations (phase de crawl). Ces informations sont ensuite indexées par des moteurs d'indexation, les termes répertoriés enrichissant un index – une base de données des mots contenus dans les pages – régulièrement mis à jour. Enfin, une interface de recherche permet de restituer des résultats aux utilisateurs en les classant par ordre de pertinence (phase de ranking).

Figure 2-1

Les différentes étapes du fonctionnement des moteurs de recherche



Les crawlers ou spiders

Les spiders (également appelés agents, crawlers, robots ou encore bots) sont des programmes de navigation visitant en permanence les pages web et leurs liens en vue d'indexer leurs contenus. Ils parcourent les liens hypertextes entre les pages et reviennent périodiquement visiter les pages retenues pour prendre en compte les éventuelles modifications.

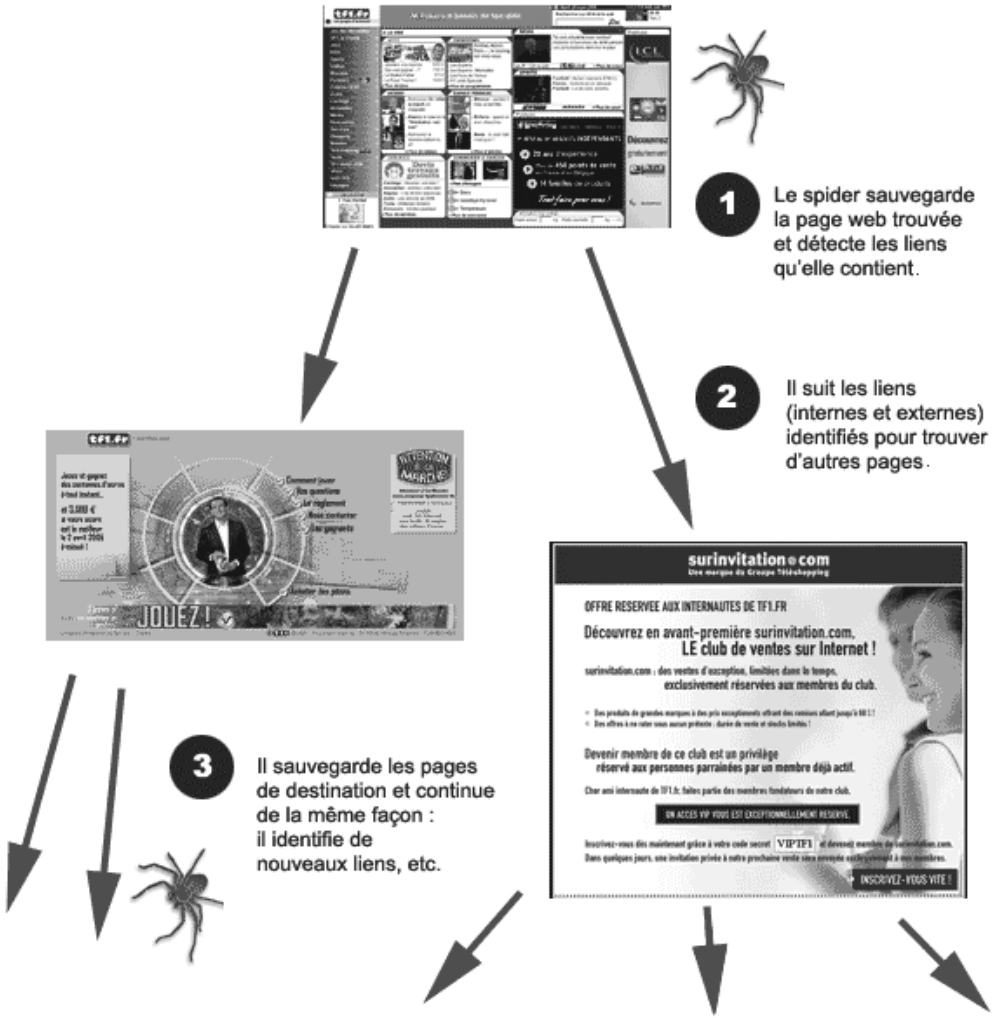


Figure 2-2

Principe de fonctionnement d'un spider

Un spider est donc un logiciel très simple mais redoutablement efficace. Il ne sait faire que deux choses :

- lire des pages web et stocker leur contenu (leur code HTML) sur les disques durs du moteur. L'équivalent de l'option « Sauvegarder sous... » de votre navigateur préféré ;
- détecter les liens dans ces pages et les suivre pour identifier de nouvelles pages web.

Le processus est immuable : le spider trouve une page, la sauvegarde, détecte les liens qu'elle contient, se rend sur les pages de destination de ces liens, les sauvegarde, y détecte les liens, etc. Et cela 24h/24… L'outil parcourt donc inlassablement le Web pour y détecter des pages web en suivant des liens… Une image communément répandue pour un spider serait celle d'un internaute fou qui lirait et mémoriserait toutes les pages web qui lui sont proposées tout en cliquant sur tous les liens qu'elles contiennent pour aller sur d'autres documents, etc.

Parmi les spiders connus, citons notamment Googlebot de Google, Yahoo! Slurp de Yahoo!, MSNBot de Microsoft Bing ou encore Exabot d'Exalead.

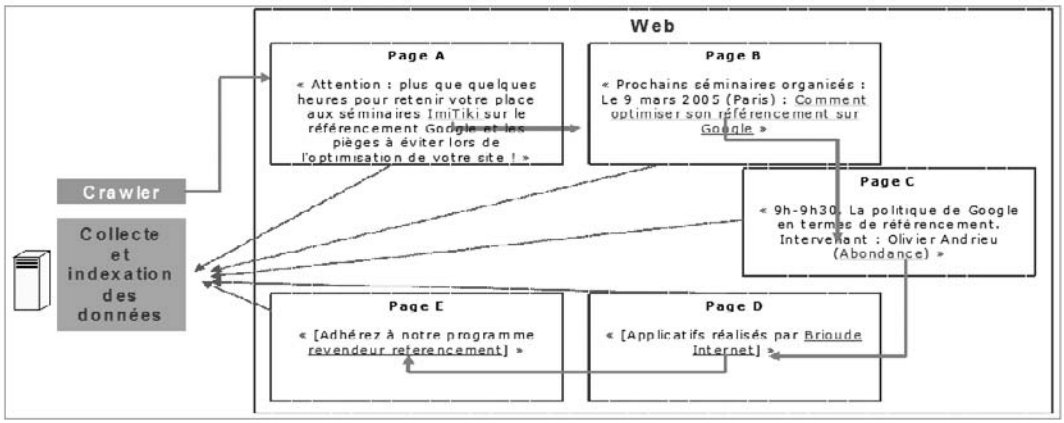


Figure 2-3

Processus de crawl (ou crawling) des robots en suivant les liens trouvés dans les pages web

> **Le fichier robots.txt pour indiquer aux spiders ce qu'ils ne doivent pas faire**
>
> Le fichier `robots.txt` est utilisé par les webmasters pour indiquer aux spiders les pages qu'ils souhaitent indexer ou non (voir chapitre 9).

Mais parcourir le Web ne suffit pas. En effet, lorsqu'un spider arrive sur une page, il commence par vérifier s'il ne la connaît pas déjà. S'il la connaît, il contrôle si la version découverte est plus récente que celle qu'il possède déjà… En cas de réponse positive, il

supprime l'ancienne version et la remplace par la nouvelle. L'index se met ainsi automatiquement à jour.

Quels critères de décision ?

Pour savoir si une page est plus récente qu'une version déjà sauvegardée, le moteur de recherche va jouer sur plusieurs facteurs complémentaires :

– la date de dernière modification du document fournie par le serveur ;

– la taille de la page en kilo-octets ;

– le taux de modification du code HTML du document (son contenu) ;

– les zones modifiées : charte graphique ou contenu réel. Ainsi, certains moteurs pourront estimer que l'ajout d'un lien dans un menu de navigation ne constitue pas une modification suffisante pour être prise en compte… Ils sauront différencier « charte graphique et de navigation » avec « contenu réel » et ne prendre en compte que la deuxième forme de modifications…

En tout état de cause, une page affichant la date et l'heure ne sera pas considérée comme mise à jour de façon continue. Il est nécessaire que le spider détecte une « vraie » modification en son sein pour mettre à jour son index.

De la « Google Dance » au « Minty Fresh »…

Il y a quelques années de cela, les mises à jour des index des moteurs étaient mensuelles. Chaque mois, le moteur mettait à jour ses données en supprimant un ancien index pour le remplacer par un nouveau, mis à jour pendant les 30 derniers jours par ses robots, scrutant le Web à la recherche de nouveaux documents ou de versions plus récentes de pages déjà en sa possession. Cette période avait notamment été appelée la « Google Dance » par certains webmasters. Pour l'anecdote, elle fut d'ailleurs pendant quelque temps indexée (c'est le cas de le dire) sur les phases de pleine lune… On savait, à cette époque, que lorsque la pleine lune approchait, un nouvel index était en préparation chez Google… Nous verrons plus loin que l'expression « Google Dance » désigne désormais tout autre chose.

Ce système de mise à jour mensuelle des index n'a plus cours aujourd'hui. La plupart des moteurs gèrent le crawling de manière différenciée et non linéaire. Ils visitent plus fréquemment les pages à fort taux de renouvellement des contenus (très souvent mises à jour) et se rendent moins souvent sur les pages « statiques ». Ainsi, une page qui est mise à jour quotidiennement (par exemple, un site d'actualité) sera visitée chaque jour – voire plusieurs fois par jour – par le robot tandis qu'une page rarement modifiée sera « crawlée » toutes les quatre semaines en moyenne. De plus, la mise à jour du document dans l'index du moteur est quasi immédiate. Ainsi, une page fréquemment mise à jour sera le plus souvent disponible à la recherche sur le moteur quelques heures plus tard. Ces pages récemment crawlées sont par exemple identifiables sur Google car la date de crawling est affichée. Exemple ici sur une recherche effectuée le 13 juillet 2009.

Figure 2-4

Affichage par Google de la date ou du délai d'indexation de la page. Celui-ci peut être très rapide, parfois de l'ordre de quelques minutes seulement.

Le résultat proposé à la figure 2-4 montre bien que la page proposée a été crawlée (sauvegardée par les spiders) quelques heures auparavant et qu'elle a été immédiatement traitée et disponible dans les résultats de recherche.

Le « Minty Fresh Indexing »

À la mi-2007, Google a accéléré son processus de prise en compte de documents, certaines pages se retrouvant dans l'index du moteur quelques minutes seulement après leur création/modification. Ce phénomène est appelé *Minty Fresh Indexing* par le moteur de recherche. Matt Cutts, dont nous avons déjà parlé au chapitre précédent, explique ce concept sur son blog : *http://www.mattcutts.com/blog/minty-fresh-indexing/.*

On pourra noter que la technique de suivi des liens hypertextes par les spiders peut poser plusieurs problèmes pour :

- l'indexation des pages dites « orphelines » qui ne sont liées à aucune autre et ne peuvent donc pas être repérées par les crawlers qui n'ont aucun lien à « se mettre sous la dent » (si tant est que les robots aient des dents…) pour l'atteindre. Il en est ainsi des sites qui viennent d'être créés et qui n'ont pas encore de *backlinks* (liens entrants) qui pointent vers eux ;
- l'indexation des pages dynamiques de périodiques ou de bases de données (ces pages étant moins facilement prises en compte, nous y reviendrons au chapitre 7) ;
- les pages pointées par des documents proposant des liens qui ne sont pas pris en compte par les moteurs de recherche, comme certains liens écrits en langage JavaScript. Là aussi, nous y reviendrons (chapitre 7).

Le passage des spiders sur les sites peut être vérifié par les webmasters en analysant les fichiers « logs » sur les serveurs (ces fichiers indiquent l'historique des connexions qui ont eu lieu sur le site, y compris celles des spiders). La plupart des outils statistiques comprennent dans leurs graphiques ou données une partie « visites des robots ». Attention cependant, ces outils doivent le plus souvent être spécifiquement configurés pour prendre en compte tous les robots émanant de moteurs français. Les outils statistiques, notamment d'origine américaine, ne prennent pas toujours en compte ces spiders « régionaux »…

Pour tracer les robots…

Plusieurs applications en ligne permettent également d'analyser les visites des robots sur des pages données (voir notamment les solutions gratuites *http://www.robotstats.com/* et *http://www.spywords.com/).* Des « marqueurs » doivent être intégrés par les webmasters dans les pages et les services surveillent si l'un des visiteurs est le robot d'un moteur de recherche.

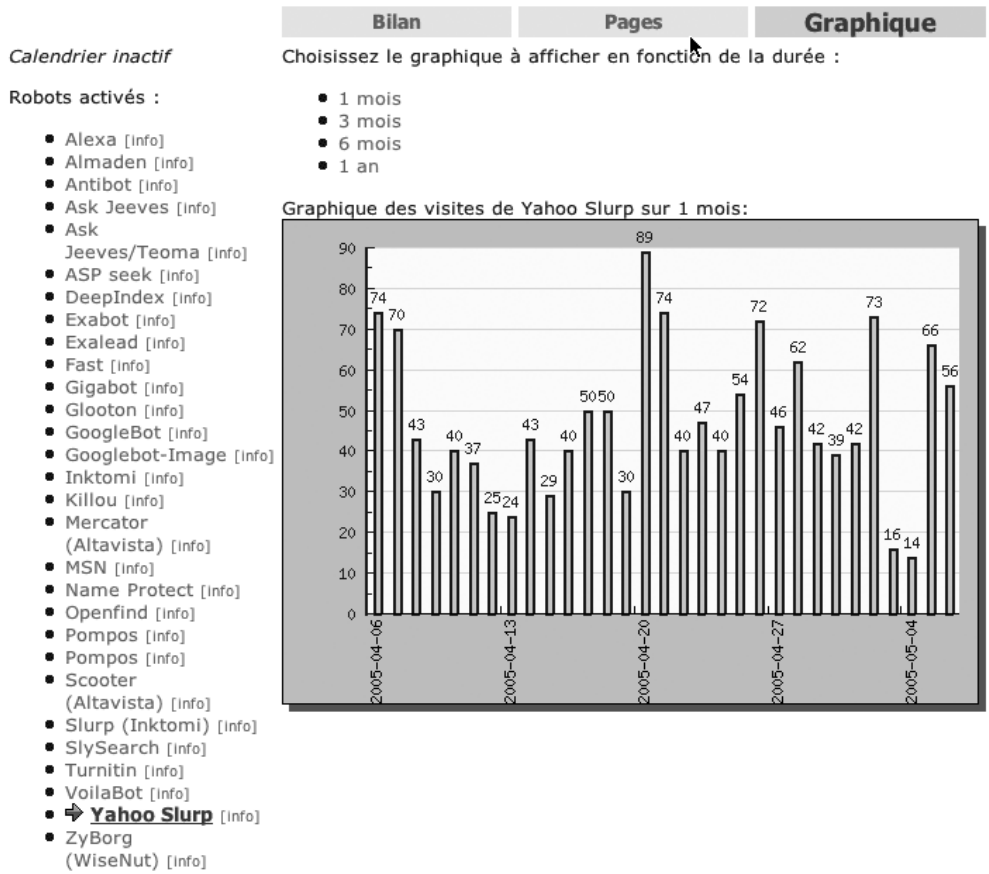


Figure 2-5

Exemple de statistiques fournies par un utilitaire de statistiques en ligne (ici Robotstat, utilitaire français : http://www.robotstats.com/)

Le moteur d'indexation

Une fois les pages du Web crawlées, le spider envoie au moteur d'indexation les informations collectées. Historiquement, plusieurs systèmes d'indexation des données ont été utilisés.

- Indexation des seules balises meta (*meta tags*) insérées par les webmasters dans le code source des pages HTML. Ces balises qui comprennent, entre autres, le résumé et les mots-clés attribués par l'auteur à la page. Très peu de moteurs fonctionnent encore ainsi aujourd'hui.
- Indexation en texte intégral (c'est de loin le cas le plus fréquent). Tous les mots d'une page, et plus globalement son code HTML, sont alors indexés.

Le plus souvent donc, les systèmes d'indexation se chargent d'identifier en « plein texte » l'ensemble des mots des textes contenus dans les pages ainsi que leur position. Certains moteurs peuvent cependant limiter leur capacité d'indexation. Ainsi, pendant de longues années, Google s'est limité aux 101 premiers kilo-octets des pages (ce qui représentait cependant une taille assez conséquente). Cette limite n'est plus aujourd'hui d'actualité. D'autres moteurs peuvent effectuer une sélection en fonction des formats de document (Excel, PowerPoint, PDF…).

Enfin, sachez que, comme pour les logiciels documentaires et les bases de données, une liste de mots « vides » (par exemple, « le », « la », « les », « et »…), appelés *stop words* en anglais, est le plus souvent automatiquement exclue (pour économiser de l'espace de stockage) ou ces mots sont systématiquement éliminés à l'occasion d'une requête (pour améliorer la rapidité des recherches).

Le traitement des stop words par les moteurs de recherche

On a souvent tendance à dire que la plupart des moteurs de recherche ignorent les *stop words* tels que, en anglais : « the », « a », « of », etc., ou en français : « le », « la », « un », « de », « et », etc. Ceci est exact, comme le montre l'explication de Google dans son aide en ligne :

Google ignore les chaînes de caractères dont le poids sémantique est trop faible (également désignés par « mots vides » ou « bruit ») : « le », « la », « les », « du », « avec », « vous », etc., mais aussi des mots spécialisés tels que « http » et « .com » ainsi que les lettres/chiffres d'un seul caractère, qui jouent rarement un rôle intéressant dans les recherches et risquent de ralentir notablement le processus.

On pourrait donc logiquement s'attendre à ce qu'une requête sur les expressions « moteur de recherche » et « moteur recherche » donnent les mêmes résultats. Eh bien non… S'il y a un certain recouvrement entre les deux pages de résultats, elles ne sont pas identiques. Alors, pourquoi cette différence ?

Cela semble venir du fait que Google tient compte de la proximité des mots entre eux dans son algorithme de pertinence. Par exemple, sur la requête « moteur de recherche », Google ne tient pas compte du « de » mais il se souvient tout de même qu'il existe un mot entre « moteur » et « recherche ». En d'autres termes, la requête « moteur de recherche » équivaut pour Google à « moteur * recherche » (l'astérisque (*) étant pour Google un joker remplaçant n'importe quel mot). Alors que sur la requête « moteur recherche », les pages qui contiennent ces deux mots l'un à côté de l'autre seront mieux positionnées, toutes choses égales par ailleurs, que celles qui contiennent l'expression « moteur de recherche »… ☞

Le traitement des stop words par les moteurs de recherche *(suite)*

Pour être plus clair, raisonnons sur un exemple : sur l'expression « franklin roosevelt » (*http://www.google.fr/search?q=franklin+roosevelt*), la majorité des pages identifiées comme répondant à la requête contiennent le nom ainsi orthographié : « Franklin Roosevelt ». Insérons maintenant n'importe quel *stop word* entre les deux termes et lançons la requête « franklin le roosevelt » (*http://www.google.fr/search?q=franklin+le+roosevelt*). Résultat : la plupart des pages contiennent le nom différemment orthographié, sous la forme « Franklin *quelque-chose* Roosevelt ». Google s'est donc souvenu que la requête était sur trois termes, même si le deuxième n'a pas été pris en compte. Et ça change tout au niveau des résultats…

Vous voulez une autre démonstration ? Tapez la requête « franklin * roosevelt » (*http://www.google.fr/search?q=franklin+*+roosevelt*) et vous obtiendrez quasiment la même réponse que pour « franklin le roosevelt ». Là encore, le moteur s'est souvenu que la requête s'effectuait sur trois termes, le premier et le dernier seulement étant pris en compte…

Comment faire, alors, pour que Google prenne en compte le *stop word* s'il vous semble important pour votre recherche ? Il existe deux façons de le faire : soit avec les guillemets, soit avec le signe +.

– Les guillemets vont vous permettre d'effectuer la requête « *"*moteur de recherche*"* » (*http://www.google.com/search?q=%22moteur+de+recherche%22*), les trois mots dans cet ordre et les uns à côté des autres. Dans ce cas, Google prend bien en compte le mot vide dans son algorithme.

– Le signe + vous permet, à l'aide de la requête « *moteur +de recherche* » (*http://www.google.com/search?q=moteur+%2Bde+recherche*), d'inclure de façon obligatoire le mot vide dans la recherche. En revanche, par rapport à l'exemple ci-dessus (avec les guillemets), les mots ne seront pas obligatoirement dans cet ordre et les uns à côté des autres dans les résultats. L'utilisation des guillemets permet donc d'obtenir des résultats plus précis…

L'index inversé

Au fur et à mesure de l'indexation et de l'analyse du contenu des pages web, un index des mots rencontrés est automatiquement enrichi. Cet index est constitué :

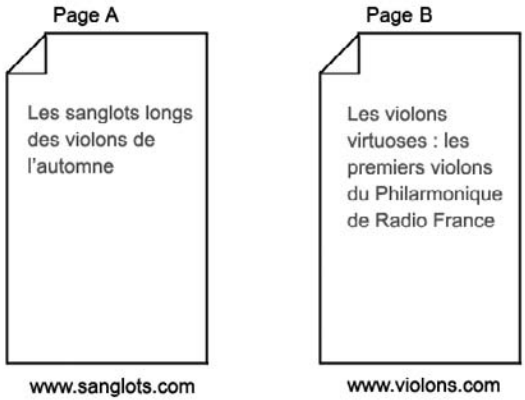
- d'un index principal ou maître, contenant l'ensemble du *corpus* de données capturé par le spider (URL et/ou document…) ;
- de fichiers inverses ou index inversés, créés autour de l'index principal et contenant tous les termes d'accès (mots-clés) associés aux URL exactes des documents contenant ces termes sur le Web.

L'objectif des fichiers inverses est simple. Il s'agit d'espaces où sont répertoriés les différents termes rencontrés, chaque terme étant associé à toutes les pages où il figure. La recherche des documents dans lesquels ils sont présents s'en trouve ainsi fortement accélérée.

Pour comprendre le fonctionnement d'un index inversé, prenons, par exemple, une page A (disponible à l'adresse *http://www.sanglots.com/*) comprenant la phrase « Les sanglots longs des violons de l'automne » et une page B (*http://www.violons.com/*) contenant les mots « Les violons virtuoses : les premiers violons du Philharmonique de Radio France ».

Figure 2-6

Deux pages prêtes à être indexées par un moteur de recherche



Les données suivantes figureront dans le fichier inverse :

Tableau 2-3 Exemple d'index inversé

Terme	Numéro du document indexé	Fréquence	Emplacement			
			Titre	Adresse	Meta	Texte
Automne	A	1	–	–	–	1
France	B	1	–	–	–	1
Longs	A	1	–	–	–	1
Philharmonique	B	1	–	–	–	1
Premiers	B	1	–	–	–	1
Radio	B	1	–	–	–	1
Sanglots	A	2	–	1	–	1
Violons	A	1	–	–	–	1
Violons	B	3	–	1	–	2
Virtuoses	B	1	–	–	–	1

Une requête dans le moteur de recherche avec le mot « violons » sera traitée en interrogeant l'index inversé pour dénombrer les occurrences de ce mot dans l'ensemble des documents indexés. Cette recherche donnera ici comme résultat les deux URL *http://www.sanglots.com/* et *http://www.violons.com/*. La page web *http://www.violons.com/* apparaîtra en premier dans la liste des résultats, le nombre d'occurrences du mot « violons » étant supérieur dans cette page. À noter toutefois, par rapport à cet exemple très simple, que la fréquence des occurrences d'un mot sera pondérée par le processus de ranking des résultats (voir ci-après).

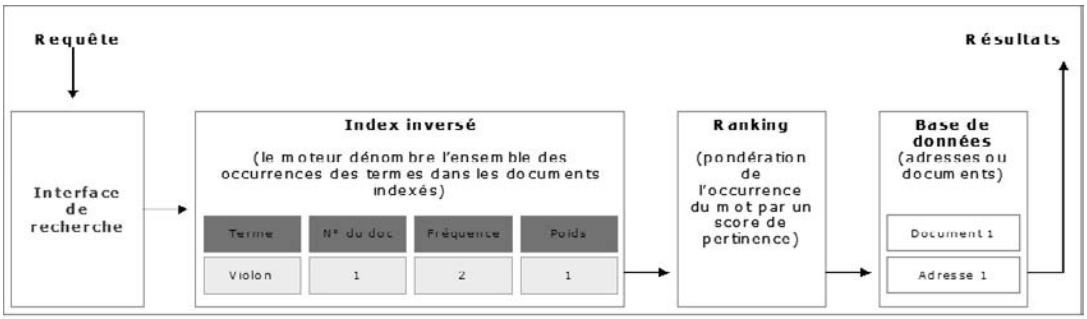


Figure 2-7

Traitement d'une requête grâce à l'index inversé

Notez que Google associe également le contenu textuel des liens pointant vers une page (*anchor text*) – concept de « réputation », sur lequel nous reviendrons largement au chapitre 5 – avec la page pointée (considérant que ces liens renvoyant vers une page fournissent souvent une description plus précise de la page que le document lui-même). Mais un moteur de recherche utilise des dizaines de critères de pertinence différents. Nous y reviendrons...

Le terme « index » peut donc être interprété de deux façons différentes :

- L'index de documents, comprenant toutes les pages prises en compte par le moteur lors d'une recherche. C'est cette base de données que nous appellerons index dans cet ouvrage, par souci de concision.
- L'index inversé, qui comprend en fait les mots-clés potentiels de recherche ainsi que leurs connexions avec l'index de documents. Il s'agit de la partie immergée de l'iceberg, invisible pour l'utilisateur du moteur mais pourtant indispensable à son fonctionnement…

L'index doit être mis à jour régulièrement, en ajoutant, modifiant ou supprimant les différentes entrées. C'est en effet la fréquence de mise à jour d'un index qui fait en grande partie la qualité des résultats d'un moteur et sa valeur (pas de doublons ou de liens morts dans les résultats…), d'où des délais de rafraîchissement relativement courts.

La plupart des moteurs de recherche ont arrêté la « course au plus gros index » depuis plusieurs années. Le premier index de Google comprenait 26 millions de pages. Le moteur de recherche avait arrêté de communiquer sur ce point en 2005 alors que sa base d'URL proposait environ 8 milliards de pages après avoir passé la barre du milliard en 2000. Un post sur le blog officiel du moteur (*http://actu.abondance.com/2008/07/mille-milliards-de-pages-web-selon.html*) estimait en juillet 2008 que le Web détenait au minimum la bagatelle de mille milliards de documents ! Ou plus précisément, mille milliards d'adresses menant à des documents. Parmi ces URL, Google identifiait énormément de duplicate content (même contenu à des adresses différentes), beaucoup de pages totalement inutiles, etc. Ce chiffre représentait donc le nombre de pages web « connues » de Google avant traitement. Mais en 2009, on pouvait estimer la taille de son index dans une fourchette oscillant entre 40 et 100 milliards de pages...

Notons qu'en juillet 2008, le moteur de recherche Cuil (*http://www.cuil.com/*) faisait sensation en annonçant un index de plus de 120 milliards de pages. Malheureusement, la pertinence des résultats ne fut pas à la hauteur des espérances... Mais il est vrai que la taille de l'index n'est pas un critère déterminant dans la pertinence d'un moteur. Encore faut-il avoir les « bonnes » pages et un algorithme de tri efficace pour en extraire la substantifique moelle...

Tableau 2-4 Tailles estimatives de plusieurs index de moteurs (en milliards de pages – fin 2009). La plupart d'entre eux ne communiquent plus sur ces chiffres et rendent très complexe la mise à jour de ce type d'information…

	Cuil	Google	Yahoo!	Live Search	Exalead	Ask.com
Taille de l'index	125	50	20	20	8	8

Le système de ranking

Le ranking est un processus qui consiste pour le moteur à classer automatiquement les données de l'index de façon à ce que, suite à une interrogation, les pages les plus pertinentes apparaissent en premier dans la liste de résultats. Le but du classement est d'afficher dans les 10 premières réponses les documents répondant le mieux à la question. Pour cela, les moteurs élaborent en permanence de nouveaux algorithmes (des formules mathématiques utilisées pour classer les documents). Ces algorithmes sont un véritable facteur différenciant. Ils ne sont donc que très rarement rendus publics. Ils sont dans certains cas protégés par des brevets et font parfois l'objet de « secrets défenses » voire de mythes comparables à celui du 7X (principal composant du Coca-Cola)…

Il existe plusieurs grandes méthodes de ranking des résultats et les moteurs utilisent pour la plupart un mélange de ces différentes techniques.

- **Le tri par pertinence.** Les résultats d'une requête sont triés en fonction de six principaux facteurs appliqués aux termes de la question (toutes ces notions seront revues en détail aux chapitres 4 et 5) :
  - localisation d'un mot dans le document (exemple : le poids est maximal si le mot apparaît dans le titre ou au début du texte) ou son adresse (URL) ;
  - densité d'un mot, calculée en fonction de la fréquence d'occurrences du mot par rapport au nombre total de mots dans le document ;
  - mise en exergue d'un mot : gras (balise `<STRONG>`), titre éditorial (balise `<Hn>`), lien, etc. ;
  - poids d'un mot dans la base de données calculé en fonction de sa fréquence d'occurrences dans l'index (les mots peu fréquents sont alors favorisés) ;
  - correspondance d'expression basée sur la similarité entre l'expression de la question et l'expression correspondante dans un document (un document est privilégié lorsqu'il contient une expression similaire à celle de la question, notamment pour des requêtes à plusieurs mots-clés) ;
  - relation de proximité entre les termes de la question et les termes utilisés dans le document (les termes proches l'un de l'autre sont favorisés).

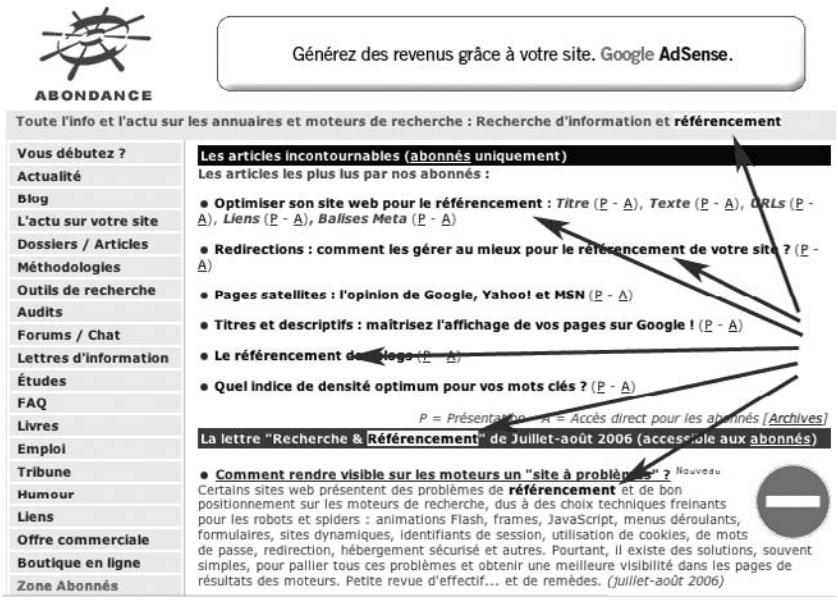


Figure 2-8

La présence et l'emplacement d'un mot dans la page peut avoir une influence sur son degré de pertinence, et donc sur son ranking par le moteur.

- **Le tri par popularité (indice de popularité).** Popularisé – mais pas inventé – par Google en 1998 (pour contrer entre autres les abus possibles des méthodes de tri par pertinence) avec son PageRank, le tri par popularité s'appuie sur une méthode basée sur la « citation » – l'analyse de l'interconnexion des pages web par l'intermédiaire des liens hypertextes – et il est *a priori* indépendant du contenu.

  Ainsi, Google classe les documents notamment en fonction de leur PageRank (nombre et qualité des liens pointant vers ces documents, nous y reviendrons en détail au chapitre 5). Le moteur analyse alors les pages contenant les liens : ceux émanant de pages issues de sites considérés comme importants « pèsent plus lourd » que ceux de pages de certains forums ou de pages perso jugées secondaires, par exemple. Plus une page est pointée par des liens émanant de pages populaires, plus sa popularité est grande et meilleur est son classement.

  Cette méthode de tri des résultats est aujourd'hui utilisée par de nombreux moteurs (pour ne pas dire tous les principaux moteurs).

- **Le tri par mesure d'audience (indice de clic).** Créée par la société DirectHit en 1998, cette méthode permet de trier les pages en fonction du nombre et de la « qualité » des visites qu'elles reçoivent. Le moteur analyse le comportement des internautes à chaque clic, chaque visite d'un lien depuis la page de résultats (et notamment le fait qu'il revienne ou non sur le moteur et au bout de combien de temps) pour tenter de trouver les pages les

plus cliquées et améliorer en conséquence leur classement dans les résultats. Plus une page sera cliquée et moins les internautes reviendront sur le moteur après l'avoir consultée (signifiant ainsi qu'ils ont trouvé « chaussure à leur pied »), et plus cette page sera considérée comme pertinente et sera donc mieux classée à la prochaine requête similaire… Cette méthode semble être utilisée encore aujourd'hui par certains moteurs dont Google.

- **Le tri par catégories ou *clustering*.** Lancé en 1997, Northernlight proposait le classement automatique des documents trouvés dans des dossiers ou sous-dossiers (clustering) constitués en fonction des réponses. Celles-ci, intégrées à chaque dossier, étaient également triées par pertinence. Cette technique de « clusterisation » thématique des résultats est aujourd'hui notamment utilisée, entre autres, par le français Exalead (*http://www.exalead.com/*) et les américains Vivisimo (*http://www.vivisimo.com/*) et Clusty (*http://www.clusty.com/*) ainsi que sur la version américaine de Bing (*http://www.bing.com/*) grâce à la technologie de la société Powerset (*http://www.powerset.com/*), entreprise rachetée par Microsoft en 2008.

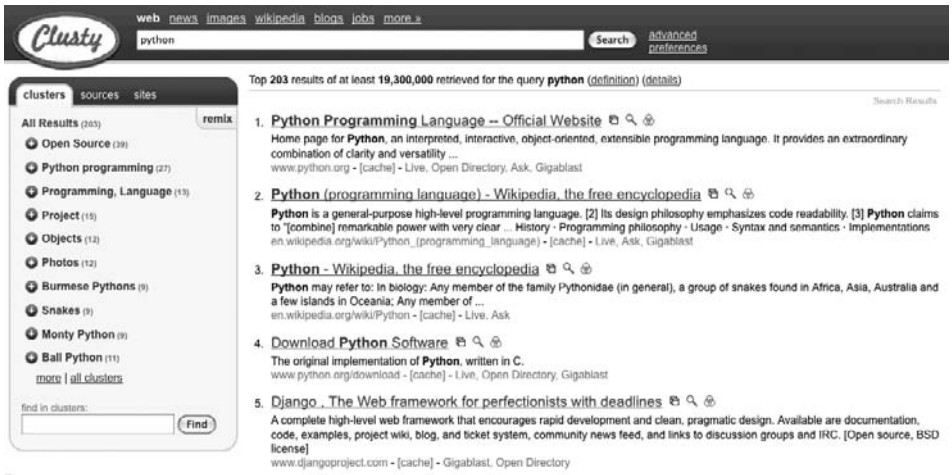


Figure 2-9

Le moteur de recherche Clusty « clusterise » ses résultats : il propose, sur la gauche de l'écran, des dossiers thématiques qui regroupent les résultats par grand domaine.

> **Les méthodes de tri**
>
> Pour plus d'informations, voir l'article de Jean-Pierre Lardy sur les *Méthodes de tri des résultats des moteurs de recherche* (avec tous nos encouragements pour recopier cette adresse sur votre navigateur) : *http://archivesic.ccsd.cnrs.fr/documents/archives0/00/00/00/53/sic_00000053_02/sic_00000053.html.*

Les moteurs sont également amenés à ajuster en permanence leurs algorithmes afin de contrer le *spamdexing*, c'est-à-dire les techniques peu scrupuleuses de spam utilisées par certains webmasters pour tromper les moteurs de recherche et améliorer artificiellement le positionnement d'une page.

Parmi les techniques les plus connues (et pénalisées par les moteurs), dont nous reparlerons au chapitre 8, citons notamment – et par ordre d'apparition chronologique – le fait de multiplier les mots-clés dans les balises meta des pages HTML, qui a certainement amené les moteurs à ne plus prendre en compte ce champ (voir chapitre 4), le fait d'intégrer un texte invisible sur une page (en blanc sur fond blanc, par exemple), la création de sites miroirs ou de liens fictifs ou invisibles pointant vers une page (ce qui permet de détourner l'indice de popularité), les pages satellites, la mise en place de faux portails contenant en fait des liens commerciaux ou le développement de « fermes de liens » (*links farms*), à savoir des listes de liens sans cohérence ayant pour unique objectif de gonfler la popularité des sites inscrits, la création de faux communiqués de presse, l'achat de liens au *prorata* du PageRank des pages, etc. Rendez-vous au chapitre 8 pour tous ces points.

Le logiciel de recherche/moteur d'interrogation

Le moteur d'interrogation (*searcher*) est l'interface frontale (formulaire de recherche) proposée aux utilisateurs. Plusieurs niveaux de requête (interface de recherche simple ou avancée) sont en général offerts. À chaque question, par le biais d'un script CGI (*Common Gateway Interface*), une requête est générée dans la base de données et une page web dynamique restitue les résultats, généralement sous forme de listes ou de cartes de résultats. L'interface CGI permet d'exécuter un programme sur un serveur et de renvoyer le résultat à un navigateur Internet.

Figure 2-10

La recherche avancée de Yahoo! (http://fr.search.yahoo.com/web/advanced) propose de nombreuses et puissantes fonctionnalités de recherche…

Focus sur le fonctionnement de Google

Créé en 1998 par deux étudiants de l'université de Stanford, Sergey Brin et Larry Page, Google (qui s'appelait *Backrub* lors de ses premières versions) s'est rapidement imposé comme le leader mondial des moteurs de recherche.

Le stockage des données et la réponse aux requêtes sont effectués à partir de dizaines de milliers de PC traditionnels tournant sous Linux. Réunis en clusters (grappes), les ordinateurs sont interconnectés selon un système basé sur la répartition des charges entre ordinateurs (un ordinateur distribue les tâches au fur et à mesure vers les autres ordinateurs disponibles).

D'un coût moins élevé que celui des serveurs, les PC traditionnels offrent un avantage au moteur de recherche dans la mesure où il est possible d'agrandir relativement facilement le parc informatique à mesure que croissent le Web et la quantité de documents à indexer.

L'index de Google est découpé en petits segments (des *shards*) afin qu'ils puissent être répartis sur l'ensemble des machines distribuées dans des *datacenters* déployés dans le monde entier, cela afin de réduire au maximum les temps de réponse aux requêtes et les coûts en bande passante. Pour rester disponible en cas de défaillance d'un PC, chaque shard est dupliqué sur plusieurs machines. Plus le PageRank est élevé et plus le nombre de duplicata est élevé (voir *http://www.webrankinfo.com/actualites/200411-infrastructure-google.htm*).

Dévoilée au début des années 2000 (et probablement toujours similaire à l'heure actuelle, même si plusieurs projets, dont le célèbre « BigDaddy » l'ont renouvellée), l'architecture de Google, présentée en figure 2-11, fait apparaître l'interconnexion de plusieurs composants séparés.

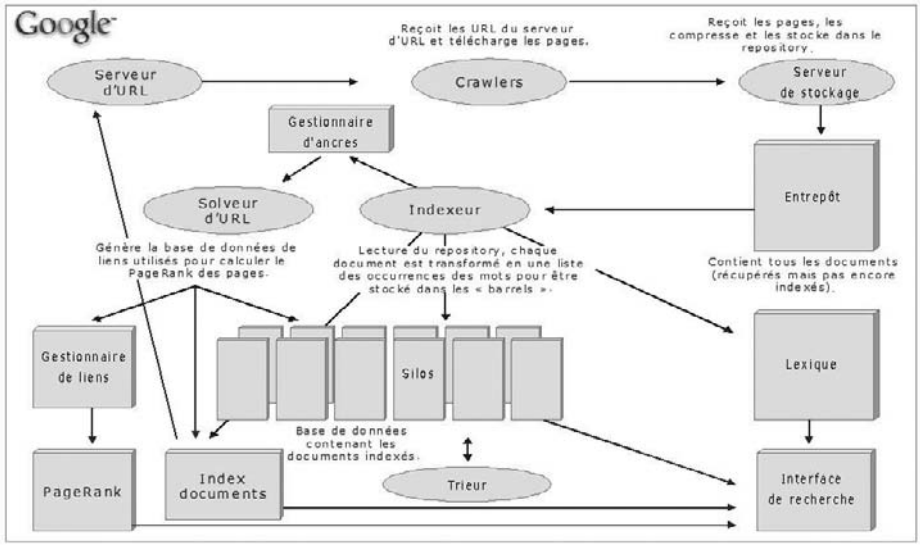


Figure 2-11

Architecture fonctionnelle de Google d'après Sergey Brin et Lawrence Page - The Anatomy of a Large-Scale Hypertextual Web Search Engine *(http://www-db.stanford.edu/~backrub/google.html).*

Chaque composant a un rôle bien défini :

- le serveur d'URL (*URL server*) envoie aux crawlers (Googlebot) toutes les adresses des pages devant être visitées (et notamment les liens soumis via le formulaire de soumission de Google) ;
- le serveur de stockage (*store server)* compresse les pages extraites par les crawlers et les envoie au *Repository* – l'entrepôt - où elles sont stockées ;
- l'indexeur lit et décompresse le contenu du Repository. Il associe à chaque document un numéro d'identifiant, *docID*, et convertit chaque page en un ensemble d'occurrences de termes (chaque occurrence est appelée un *hit*), enregistrant les informations sur le « poids » du mot dans la page (position, taille de police…) ;
- l'indexeur distribue les occurrences dans un ensemble de silos (*barrels*) (organisés par docID) ;
- le gestionnaire d'ancres (*Anchors*) stocke certaines informations générées par l'indexeur, à savoir les liens hypertextes et les ancres qui leurs sont associés (textes des liens) ;
- le solveur d'URL (*URL Resolver*) récupère les informations fournies par l'Anchors et convertit chaque adresse URL pointée par l'ancre en un docID (si cette adresse n'existe pas dans le *Doc Index*, il l'ajoute) ;
- le gestionnaire de liens (*Links*) contient des paires de docID (reçues du solveur d'URL). Il s'agit de paires de liens car chaque ancre appartient à une page et pointe vers une autre page ;
- le PageRank récupère les informations de cette base de données de liens pour calculer le PageRank de chaque document (indice de popularité) ;
- le trieur (*Sorter*) récupère les données stockées dans les barrels, organisées par docID, et les réorganise en *wordID* (identités des mots). Cette opération permet de générer l'index inversé, stocké dans les mêmes barrels ;
- la liste des mots créée par le Sorter est comparée avec celle du *Lexicon* (lexique) et tout mot ne figurant pas dans le lexique y est ajouté ;
- enfin, le *Searcher* (interface de recherche) exécute les recherches pour répondre aux requêtes des utilisateurs. Il utilise pour cela le lexique (créé par l'indexeur), l'index inversé contenu dans les barrels, les adresses URL associées aux mots de l'index inversé (provenant du Doc Index) et toutes les informations du PageRank concernant la popularité des pages.

À chaque requête, le serveur consulte l'index inversé et regroupe une liste de documents comprenant les termes de recherche (*hit list*). Il classe ensuite les pages en fonction d'indices de popularité et de pertinence. Simple, non ?

S'il y a de fortes chances que l'architecture de Google ait grandement changé dans les détails depuis cette présentation datant du début des années 2000, on peut cependant penser que son mode de fonctionnement global est encore aujourd'hui proche de ce qui est décrit ci-dessus...

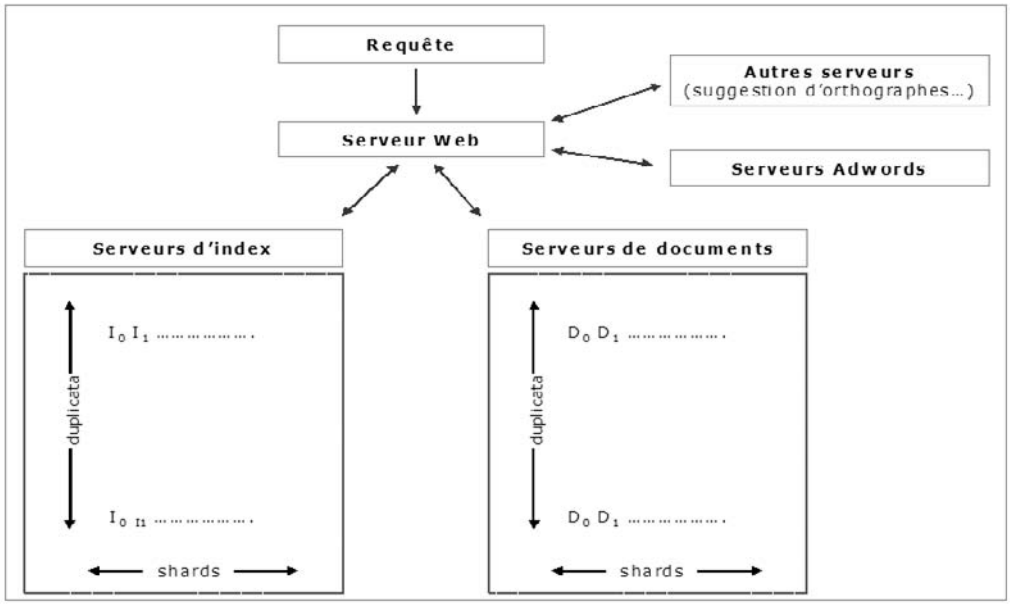


Figure 2-12

Schéma de l'utilisation des serveurs de Google utilisés pour la réponse aux requêtes
Source : http://www.webrankinfo.com/actualites/200411-infrastructure-google.htm

Google : index principal et secondaire

Bien sûr, comme tous les moteurs de recherche, Google utilise un index qui contient les pages web dans lesquelles il va effectuer ses investigations. On l'a vu, selon des sources plus ou moins officielles, la taille de cet index varierait actuellement entre 40 et 100 milliards de pages. Peu importe finalement ce chiffre, à partir du moment où l'index contient les « bonnes pages », c'est-à-dire celles qui répondent à nos requêtes...

Mais Google utilise en fait deux index depuis 2003. Le premier, l'index principal, contient les pages que Google considère comme « essentielles », les plus importantes donc. L'index secondaire, pour sa part, contient ce qu'on pourrait appeler « un deuxième choix », contenant notamment de nombreuses pages considérées comme étant du duplicate content (voir chapitre 7). De plus, les pages présentes dans cet index secondaire sont

« crawlées » (visitées par les robots du moteur) bien moins souvent que celles de l'index principal.

Longtemps, Google a indiqué le fait qu'une page était issue de cet index secondaire au travers de la mention « Supplemental Result » ou « Résultat complémentaire » dans ses résultats.

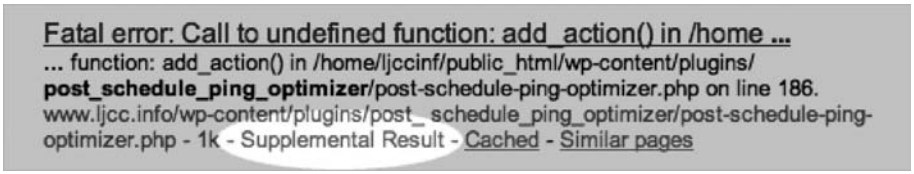


Figure 2-13

Mention « Supplemental Result » indiquant que le résultat est issu de l'index secondaire.

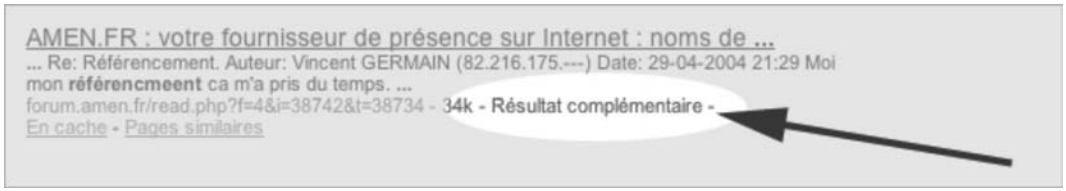


Figure 2-14

Mention identique sur le site en français

Fin 2007, Google a communiqué sur le fait que cet index secondaire n'existait plus (*http://blog.abondance.com/2007/12/rsultats-complmentaires-sur-google-cest.html*) et que le moteur de recherche n'utilisait plus qu'un seul index.

Les deux index cohabitent pourtant encore...

Cependant, il semble bien qu'aujourd'hui encore, cette différence existe toujours, entre deux sous-ensembles de l'index, quelle que soit la forme que prenne cette dichotomie... Peut-être ne s'agit-il pas de deux « index » au sens technique du terme, mais toujours est-il qu'il est clair que toutes les pages ne sont pas placées au même niveau par Google dans son index de départ. Dans ce chapitre, nous resterons donc sur notre vision initiale d'index « principal » et « secondaire », peu importe finalement ce qui se passe sous le capot du moteur...

Le fait est simple à vérifier : effectuez une recherche sur un site web donné, par exemple *actu.abondance.com* sur Google.fr grâce à la requête « site:actu.abondance.com ». Google renvoie ici 4 090 résultats.

Figure 2-15

Requête « site:actu.abondance.com » sur Google

Effectuons maintenant la même requête sur un autre site ayant passé un accord de partenariat avec Google, comme AOL (*http://recherche.aol.fr/*).

Figure 2-16

Requête « site:actu.abondance.com » sur AOL (technologie de recherche Google)

AOL trouve environ 1 500 pages. En fait, ce moteur travaille uniquement sur l'index « principal » de Google (ou tout du moins, Google ne leur fournit que des résultats issus de cet index)...

Il semblerait en fait que :

- L'index principal contienne les pages considérées comme les plus pertinentes par Google.

- L'index secondaire contienne des pages considérées comme moins importantes ou comme dupliquées par Google, et il ne les affichera que lorsque vous le demanderez, en cliquant par exemple sur ce message en cas de duplicate content comme sur la figure suivante.

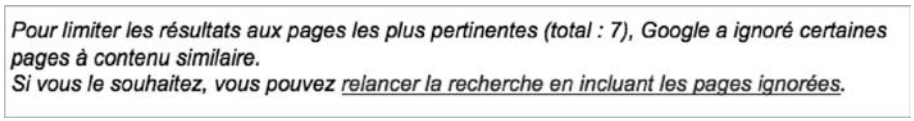


Figure 2-17

Message mentionnant un problème de duplicate content sur Google

En clair, dans l'exemple ci-dessus, avant que ce message n'apparaisse, vous visualisez les pages issues de l'index principal. Une fois le lien « Relancer la recherche en incluant les pages ignorées » cliqué, c'est tout l'index général (principal + secondaire) qui est pris en compte...

La problématique est donc très importante puisqu'une page qui se trouve dans l'index principal conservera ses chances d'être bien positionnée dans les résultats du moteur de recherche, alors qu'une page qui se trouve dans l'« enfer de l'index secondaire » est quasiment perdue pour un bon positionnement... Être « connu de Google » ne suffit donc pas à être bien positionné si la page se trouve dans l'index secondaire il vous faudra alors bien vérifier quelles sont vos pages qui se trouvent dans l'une ou l'autre zone d'investigation du moteur.

Comment vérifier dans quel index sont vos pages ?

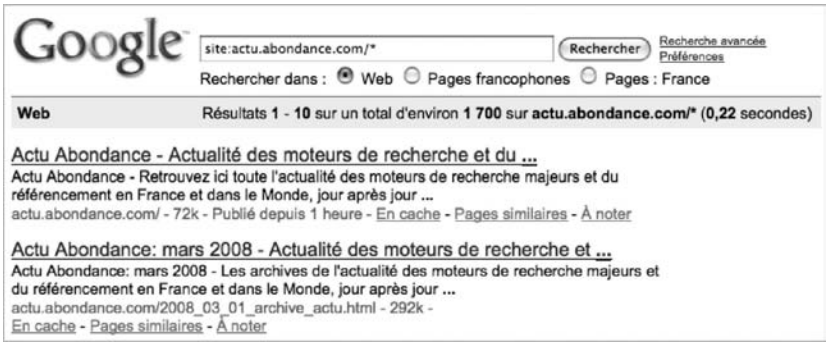
Pour vérifier combien de pages de votre site sont dans chaque index, il existe plusieurs façons, plus ou moins officielles :

- La première, on l'a vu, consiste à utiliser la commande « site: » sur Google, puis sur un site « affilié » comme AOL.fr. Le moteur Google indiquera alors le nombre total de pages indexées (index principal ET secondaire), le moteur affilié ne renverra que les pages de l'index principal. La différence entre les deux chiffres renvoyés donnera le nombre de pages dans l'index secondaire.
- Une autre solution consiste à utiliser la requête « site:www.votresite.com/* ». La mention « /* » après la requête semblerait indiquer à Google qu'il ne doit utiliser que son index principal pour effectuer la recherche.

  Le nombre de résultats renvoyés par cette syntaxe (notez que la requête « site:www.votresite.com/& » semble fonctionner également) s'approche effectivement de celui fourni par AOL. Cependant, Google n'a jamais communiqué sur cette syntaxe spécifique et nous ne vous la fournissons qu'à titre indicatif...

Figure 2-18

Indication des pages web de l'index principal sur Google



- Une solution plus fiable peut être d'utiliser les Google Webmaster Tools (*http://www.google.com/webmasters/tools/?hl=fr*), indispensable espace dédié aux webmasters sur lequel vous vous devez d'avoir ouvert un compte si vous vous intéressez au référencement...

  Dans cet espace, choisissez l'option « Votre site sur le Web » qui vous proposera deux choix importants : « Liens vers votre site » et « Liens internes ».

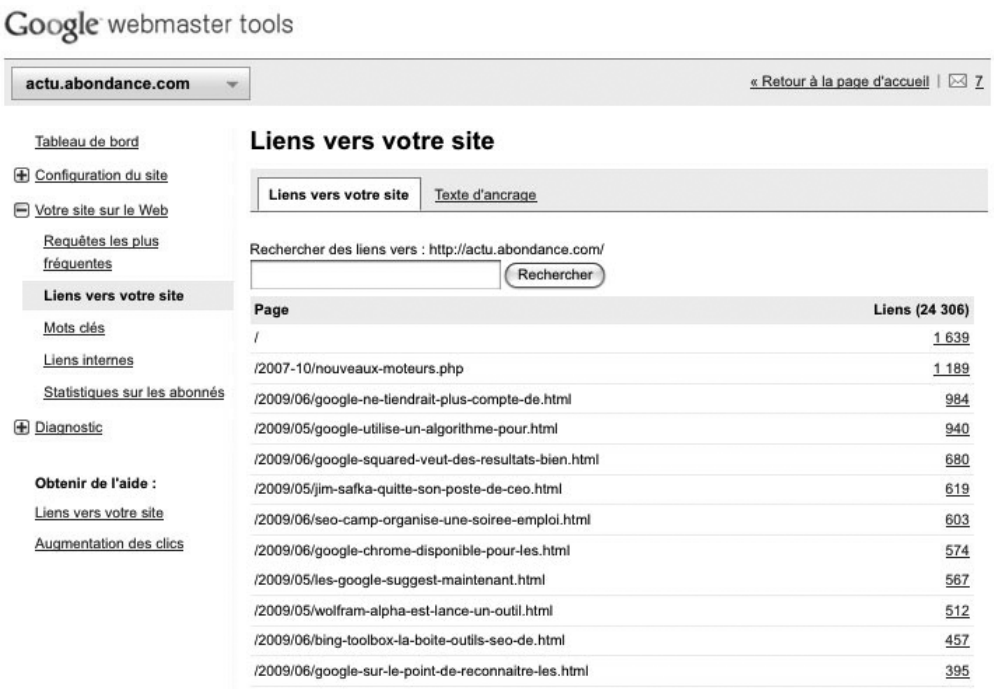


Figure 2-19

Examen des liens externes détectés par les Webmaster Tools de Google

Explorez ces deux zones : il y a de fortes chances pour que les pages qui y sont listées soient celles qui se trouvent dans l'index principal.

En effet, pour répartir les pages d'un site entre ses index principal et secondaire, il semblerait que Google tienne compte des *backlinks* (liens entrants) qui pointent vers elles. Si une page est considérée comme assez « populaire », assez « liée », elle sera plutôt stockée dans l'index principal. Si elle n'a pas assez reçu de liens (internes ou externes), elle sera directement dirigée vers le « purgatoire » de l'index secondaire.

L'avantage des Webmaster Tools de Google est que vous pourrez télécharger la liste de ces pages proposées par l'interface pour webmasters de Google et regarder son contenu de façon exhaustive (alors qu'au travers d'une syntaxe spécifique sur le moteur, comme vu précédemment, Google ne vous renverra toujours que ses 1 000 premiers résultats)... Vous pouvez également utiliser des outils comme Glync (*http://www.thegooglecache.com/glync/*) qui ajoutent des graphiques temporels aux données fournies par l'interface Google afin de suivre dans le temps votre stratégie d'obtention de liens.

Conclusion sur les index de Google

La notion d'index principal et secondaire est importante, voire capitale, en termes de référencement. Le but sera pour votre site web d'avoir le plus possible de pages dans l'index principal. On voit donc ici l'importance du *deeplinking* (stratégie d'obtention de liens externes – ou *netlinking* – vers les pages internes du site) en passant, bien sûr, par des pratiques loyales et honnêtes pour obtenir ces liens... Bien entendu, il faudra aussi que tous vos problèmes de duplicate content aient également été résolus.

Les causes qui font qu'une page se retrouve dans l'index secondaire peuvent être multiples :

- pages n'ayant pas assez de backlinks internes ou externes ;
- pages souffrant de duplicate content (voir chapitre 7) ;
- pages n'ayant pas assez de contenu textuel pour être analysées par Google ;
- pages pas assez visitées (trop peu de trafic).

En clair, pour éviter l'index secondaire, vos pages devront donc :

- Être « assez » liées par des liens internes et externes, si possible depuis des pages populaires. Par exemple, le fait de mettre un lien vers de nouvelles pages pendant quelques jours depuis votre page d'accueil peut être une bonne chose, même si cela peut ne pas suffire...
- Proposer assez de contenu textuel pour être « analysables » par les moteurs (la limite classique des 100 à 200 mots descriptifs – voir chapitre 4 – comme contenu éditorial au minimum).
- Ne pas connaître de problématique de duplicate content (voir chapitre 7).
- Pour ce qui est du trafic, si les trois soucis précédents sont réglés, ce quatrième ne devrait pas poser de gros problèmes...

La problématique du *linking* est importante, comme nous le verrons tout au long de cet ouvrage. Mettre en place une stratégie de liens vers la page d'accueil de son site est une

bonne chose pour accroître la popularité de cette dernière. Cela reste un incontournable du référencement. Mais cette stratégie doit s'accompagner d'un travail important pour gagner des liens vers les pages internes du site également (deeplinking), afin de faire en sorte que le grand nombre de pages web se trouvent le plus rapidement possible dans l'index principal, voyant leurs chances de positionnement optimisées. C'est particulièrement crucial pour les sites de contenu (presse, média...) qui ont pour obligation de voir leur pages internes bien référencées, parfois beaucoup plus que pour leur page d'accueil...

Quelques liens sur la façon dont fonctionne un moteur de recherche

Sur les robots :

– #dS.t Robots – *http://www.robots.darkseoteam.com/*

L'observatoire des robots des moteurs de recherche Google, Yahoo! et Microsoft Bing (ce site fournit des statistiques de passage des robots Google, Yahoo! et Microsoft sur sa page d'accueil).

– The Web Robots Pages – *http://www.robotstxt.org/*

Site fournissant une liste des robots actifs.

Sur Google :

– *http://www-db.stanford.edu/~backrub/google.html*

Article des deux fondateurs de Google, Sergey Brin et Lawrence Page, intitulé *The Anatomy of a Large-Scale Hypertextual Web Search Engine* et publié en 1998.

– *http://www.computer.org/micro/mi2003/m2022.pdf*

Article de l'IEEE Computer Society – Web Search For a Planet, *The Google Cluster Architecture*.

Lexiques sur les moteurs de recherche :

– *http://www.sumhit-referencement.com/savoir-lexique.asp*

– *http://www.webrankinfo.com/lexique.php*

Deux lexiques assez complets dans lesquels vous devriez trouver la signification de la plupart des termes employés dans le monde des moteurs de recherche et du référencement.

Les Sitelinks de Google

Vous l'avez peut-être remarqué, Google propose sur certaines requêtes, en dessous du premier lien-résultat, huit liens (au plus), sous la forme de deux colonnes de quatre, vers des zones internes du site en question. Voici un exemple pour le mot-clé « eyrolles ».

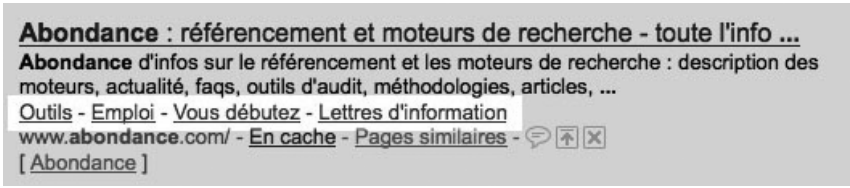
Figure 2-20

Google affiche 8 liens internes pour le site proposé dans ses résultats.

Parfois, ce sont seulement quatre liens, disposés de façon horizontale cette fois, qui peuvent être affichés.

Figure 2-21

Google affiche ici quatre liens horizontalement.



Ces liens sont appelés *Sitelinks* (ou « liens de site » en français) par Google. Le moteur de recherche explique ce concept à l'adresse suivante :

http://www.google.com/support/webmasters/bin/answer.py?answer=47334&topic=8523.

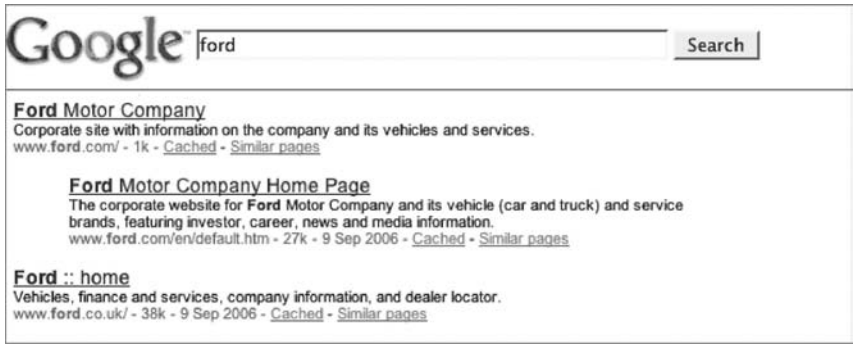
Il est difficile d'en savoir plus sur ces liens, car peu d'informations officielles existent. Cependant, certains points sont clairs car provenant de constatations évidentes ou du peu d'informations fournies par Google :

- Ces liens Sitelinks ne s'affichent, sous la forme de deux colonnes de quatre liens, que pour le premier lien de la page de résultats et sur n'importe quel résultat sous la forme de quatre liens horizontaux.
- Le fait qu'ils s'affichent ou non (présence ou absence de Sitelinks) semble venir de deux critères essentiels : la pertinence (il est vraisemblable que les liens Sitelinks ne s'affichent que si le premier lien est « évident » par rapport à la requête demandée : présence du mot-clé dans le nom de domaine et/ou taux de pertinence – valeur connue de Google mais non divulguée – supérieur à une certaine limite, etc.) et la qualité de la structure du site (voir plus loin).
- Les Sitelinks ne semblent s'afficher que pour des pages d'accueil de sites.
- Le système est entièrement automatisé (pas d'intervention humaine).
- Selon Google, un critère important est la « structure du site », donc la façon dont les liens internes y sont proposés, sans qu'il soit possible d'en savoir plus sur la façon dont le moteur de recherche choisit tel lien plutôt que tel autre...
- Les Sitelinks peuvent pointer vers une page interne du site lui-même (*www.votresite.com/repertoire/pageinterne.html*) ou vers une page d'un sous-domaine (*sousdomaine.votresite.com/repertoire/pageinterne.html*) de ce même site.
- Les Sitelinks peuvent être au format texte, image (l'attribut `alt` fournit alors le texte affiché) ou même parfois JavaScript dans la page d'accueil.

- Attention aux redirections : exemple sur la requête « ford » sur Google.com.

Figure 2-22

Requête « ford » sur Google.com



La première page identifiée (*www.ford.com*), considérée comme très pertinente par Google certainement par le fait que de nombreux liens pointent vers elle, est un document de 1 Ko contenant une redirection JavaScript – donc non détectée par Google – vers la page du dessous (*www.ford.com/en/default.htm*). Comme la première page ne contient pas de liens, il a été impossible à Google de déterminer quels Sitelinks il pourrait proposer. Du coup, rien n'est affiché...

- Il ne semble pas que Google tienne compte des informations incluses dans le code HTML de la page d'accueil qui bénéficie des Sitelinks. En effet, pourquoi, dans ce cas, choisirait-il les liens « Emploi », « Outils » ou « Vous débutez » plutôt que d'autres, traités de la même façon dans la page d'accueil du site Abondance.com ? Ce ne sont pas non plus les liens proposés en premier dans le code source, etc.
- Il ne s'agit pas non plus des résultats qui ressortent en premier sur l'utilisation de la requête « site: » pour n'obtenir que les pages issues du site en question. Exemple avec la requête « abondance site:abondance.com » ou « site:abondance.com ». Les pages proposées par Google dans ces deux cas n'ont pas grand-chose à voir avec celles affichées comme Sitelinks...
- Une hypothèse intéressante et crédible pourrait être la suivante : Google utilise des données de trafic renvoyées par les Google Toolbars de ses utilisateurs ou son navigateur Google Chrome, voire Google Analytics (entre autres sources) pour déterminer quels liens sont le plus souvent cliqués par les internautes lorsqu'ils se trouvent sur la page d'accueil du site en question. Selon le site Social Patterns (*http://www.socialpatterns.com/search-engine-marketing/traffic-determines-google-ui-snippet-links/*), il existerait de fortes similitudes entre les liens affichés dans les Sitelinks et les données de trafic de l'outil Alexa, comparables à ce dont Google pourrait se servir... À la lumière de cette information, observez à nouveau les liens proposés en Sitelinks et visitez les pages d'accueil des sites proposés : n'avez-vous pas l'impression que les Sitelinks sont les liens sur lesquels « vous seriez le plus tenté de cliquer » ? D'autres tests

que nous avons effectués (*http://blog.abondance.com/2008/07/un-exemple-intressant-et-rvlateur-de.html*) semblent corroborer ce fait.

De plus, cette hypothèse apporterait des réponses à de nombreuses interrogations : pourquoi les Sitelinks seraient-ils trouvés par Google lorsqu'ils sont issus de liens JavaScript alors que ses spiders ont encore du mal à suivre ce type de lien ?

Notre hypothèse est donc la suivante :

Le choix d'afficher les Sitelinks ou non est basé pour Google sur quatre critères principaux :

1. Adéquation entre la requête demandée et le site, et notamment son nom de domaine. Si le nom de domaine du premier résultat affiché contient les mots de la requête, il y a de fortes chances pour que ce site soit potentiellement affublé de Sitelinks. Notez bien que ce critère n'est pas vérifié à 100 %, on trouve également des Sitelinks sur certaines requêtes dont les mots-clés ne se retrouvent pas dans le nom de domaine. Mais cela reste assez rare...

2. Le premier résultat proposé doit être une page d'accueil (*www.siteweb.com*).

3. Le taux de pertinence du site par rapport à la requête (donnée non communiquée par Google mais dont nous pouvons attester de l'existence) doit être supérieur à une certaine limite, malheureusement inconnue. Une notion de trafic minimal est certainement également prise en compte.

4. Des liens doivent être détectables au sein de cette page d'accueil.

Le choix des Sitelinks eux-mêmes s'effectuerait alors grâce à deux critères principaux :

1. Liens pointant vers une page interne du site ou d'un de ses sous-domaines (pas de lien externe).

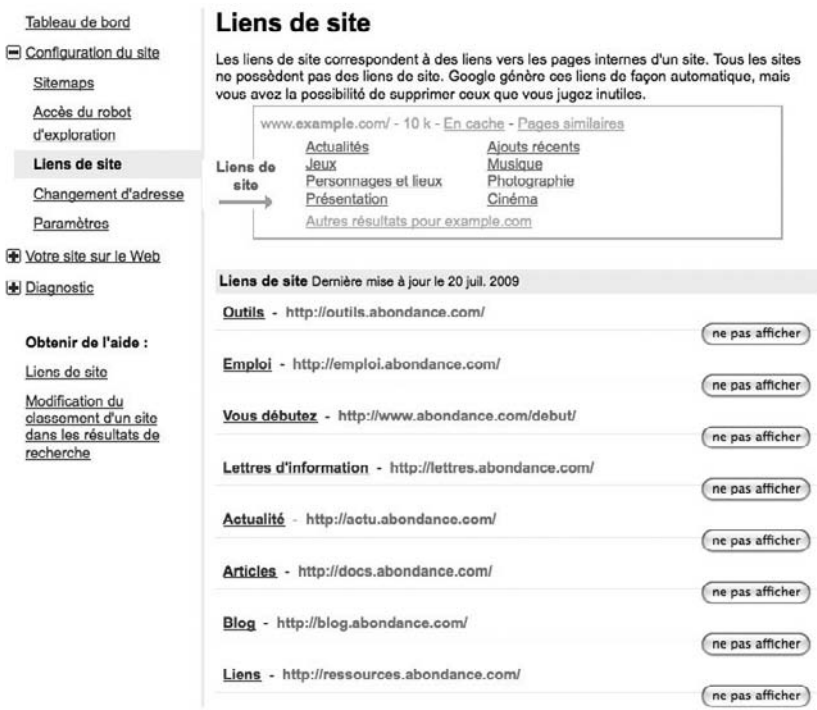
2. Nous pensons fortement, au vu de nombreux exemples analysés, que ces liens sont fournis par des statistiques de trafic et représentent les liens le plus souvent cliqués sur la page d'accueil du site.

Enfin, dernier point (important) sur les Sitelinks : sachez que, dans les Google Webmaster Tools, vous avez la possibilité d'indiquer si vous désirez supprimer un ou plusieurs intitulés. Google les remplacera alors par d'autres, après les avoir recalculés en tenant compte de votre demande. Pour cela, allez dans le menu « Configuration du Site>Liens de site ».

Peut-être vous êtes-vous posé la question en lisant ces lignes : non, il n'est (hélas) pas possible de demander à Google d'afficher tel ou tel lien en Sitelink. Vous ne pouvez que supprimer ceux que Google propose par défaut et ne pouvez pas en proposer d'autres...

Figure 2-23

Google propose de supprimer certains Sitelinks dans ses Webmaster Tools



Comment fonctionne un annuaire ?

Parmi les outils de recherche les plus utilisés historiquement par les internautes, les annuaires ont longtemps eu une place appréciable. Le plus connu d'entre eux a été celui de Yahoo! dont la version française est aujourd'hui disponible à l'adresse suivante : *http://fr.dir.yahoo.com/*. Sa version anglophone peut être trouvée ici : *http://dir.yahoo.com/*.

Ces outils ont un fonctionnement tout à fait différent de celui des moteurs de recherche que nous avons étudiés précédemment. En effet, la principale différence avec des moteurs tels que Google ou Yahoo! Search (la version « moteur » du portail Yahoo!) est qu'ils n'effectuent aucune recherche sur le contenu des documents – des pages – des sites référencés. Ils proposent simplement, ce terme n'étant pas péjoratif, une collection de fiches descriptives des sites qu'ils référencent. Ils présentent, dans une hiérarchie de catégories et sous-catégories diverses, le contenu du Web au travers de ces sites décrits par un nom et un commentaire de quelques mots. Ces outils ressemblent aux Pages Jaunes, qui auraient pris un coup de jeune en se structurant à l'aide d'un thésaurus interactif. La recherche se fait en descendant une hiérarchie qui balaie des thèmes allant du plus général au plus précis, et qui fournit, en dernier lieu, une liste de sites représentatifs du domaine présenté, quel que soit le niveau de l'arborescence atteint. Nous avons donc affaire ici à une base de données de liens pointant vers d'autres sites du Web, ces liens étant classés et décrits de façon hiérarchique.

Figure 2-24

Page d'accueil de l'annuaire francophone de Yahoo! (http://fr.dir.yahoo.com/)

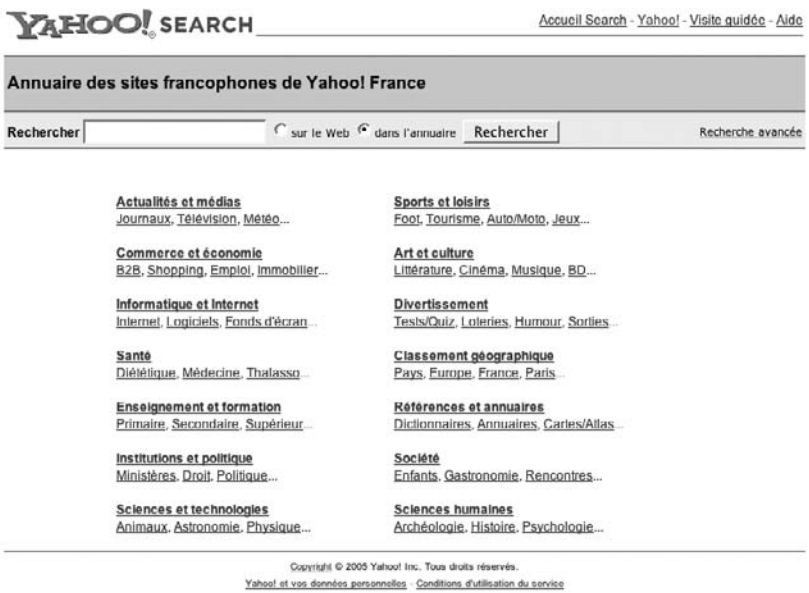


Figure 2-25

La page présentant les moteurs de recherche sur l'annuaire de Yahoo! France, dans la catégorie « Informatique et Internet>Internet>World Wide Web>Recherche sur le Web>Moteurs de recherche »

Ces annuaires sont utilisés pour trouver un site spécifique traitant d'un thème donné. Ils sont très efficaces pour trouver de l'information générale puisqu'ils décrivent les services référencés au moyen de quelques mots seulement.

Globalement, l'annuaire ne dispose que de très peu d'informations sur le site qu'il référence :

- son nom (titre) ;
- son adresse (URL) ;
- un descriptif du site, de dix à vingt mots en général, rédigé par ses documentalistes ;
- la catégorie (ou rubrique, les deux mots sont synonymes dans ce cas) dans laquelle le site est inscrit.

Certains annuaires proposent l'inscription dans plusieurs catégories, d'autres limitent cette soumission à une seule rubrique.

La recherche sur ces outils peut alors se faire de deux façons :

- En descendant l'arborescence afin d'atteindre la bonne catégorie dans laquelle trouver le site adéquat. Si elle a longtemps été utilisée, cette méthodologie de recherche semble aujourd'hui obsolète.
- En saisissant un mot-clé qui sera alors recherché dans les informations que détient l'annuaire : titre, adresse, commentaires et nom de catégorie.

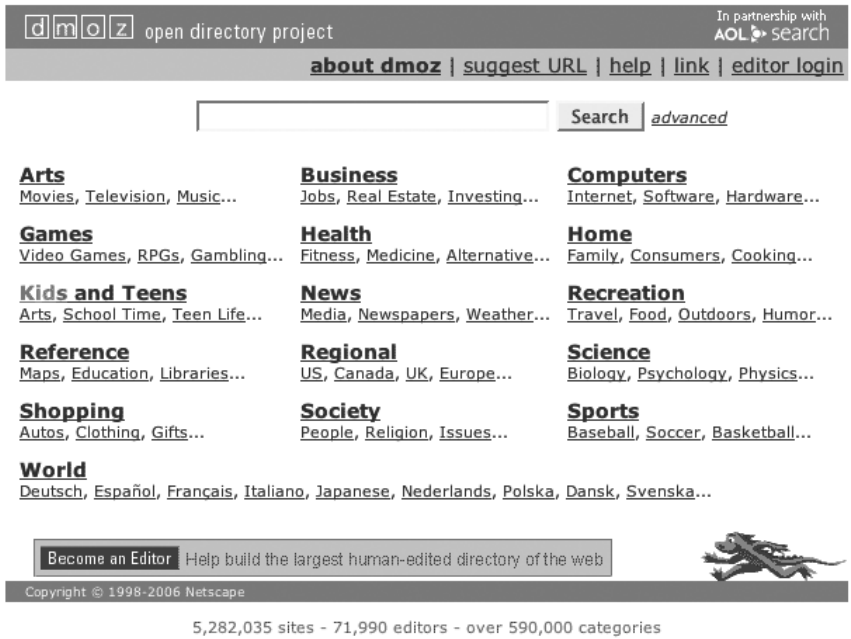
À aucun moment donc, l'annuaire n'effectue une recherche en texte intégral dans le contenu des pages qui constituent le site, comme le ferait un moteur. Il ne prend en compte que les fiches descriptives des sites que contient sa base de données.

La façon dont vos pages sont construites (titre, texte, balises meta, etc.) n'est à aucun moment prise en compte par l'algorithme de classement. Elle ne joue en tout cas aucun rôle lors de l'inscription ou de l'affichage des résultats pour une requête par mots-clés sur le contenu de l'annuaire, que ce soit sur le Guide Web de Yahoo! ou sur l'Open Directory (*http://www.dmoz.org*) ou Dmoz, autre annuaire mondialement connu.

Enfin, sachez, dans un premier temps, que pour être inscrit sur ces outils, il faut le demander. Le processus est exactement l'inverse de ce que nous connaissons pour le téléphone. Lorsqu'un opérateur de télécommunications installe une ligne téléphonique chez vous, vous êtes automatiquement inscrit dans l'annuaire. Si vous ne souhaitez pas l'être, vous devez demander à être inscrit sur une liste rouge. Sur Internet, comme souvent, c'est l'inverse : lorsque vous créez votre site web, vous n'êtes inscrit dans aucun annuaire par défaut. Vous êtes en liste rouge dès le départ. Pour être référencé, il faut soumettre une demande à chacun des différents outils pour que votre requête soit prise en compte, sachant que l'inscription de votre site peut être refusée.

Figure 2-26

L'Open Directory ou Dmoz (http://www.dmoz.org/)



Le référencement sur ces outils se fait donc au moyen d'une action volontaire de la part du responsable du site comme le montre la figure 2-27. La plupart du temps, il est recommandé (voire nécessaire) de trouver d'abord la ou les bonnes catégories dans lesquelles s'inscrire, puis d'effectuer la demande. Ensuite, le documentaliste de l'annuaire va venir inspecter votre site puis l'accepter ou le refuser dans l'annuaire en fonction de la charte éditoriale de l'outil dont il a la charge. Si votre site est accepté, il écrit alors un titre et un descriptif de lui-même et le placera dans les catégories qui lui semblent les plus pertinentes. En clair, vous proposez votre site, l'annuaire (et ses documentalistes) dispose…

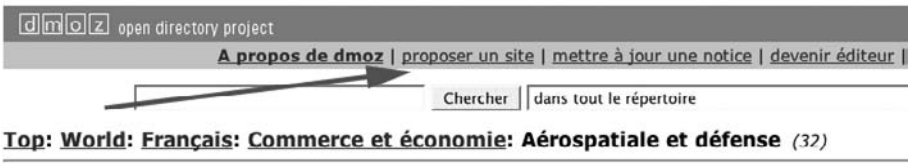


Figure 2-27

Pour être inscrit sur un annuaire (ici Dmoz), il faut tout d'abord choisir la catégorie où l'on veut apparaître (ici : Aérospatiale et défense) et cliquer sur un lien de type « Proposer un site » ou « Soumettre un site ».

Il est également important de préciser que le trafic généré par les annuaires est aujourd'hui très faible, voire quasi nul. La grande majorité du trafic « outils – moteurs + annuaires - de recherche » est à l'heure actuelle généré par les moteurs, qui feront donc l'objet de la majeure partie de cet ouvrage.

Résumé

Il est important de visualiser les différences entres les moteurs de recherche et les annuaires.

– Les moteurs de recherche sont maintenus de façon automatisée par des logiciels (robots, systèmes d'indexation, algorithmes de pertinence) alors que les annuaires sont maintenus par des êtres humains (documentalistes) qui rédigent des fiches descriptives de sites et les classent dans des rubriques (ou catégories) idoines.
– Les moteurs indexent le contenu textuel des pages alors que les annuaires « se contentent » de fiches descriptives de sites (titre, résumé, URL, catégorie(s)).
– Les moteurs indexent des pages web et les annuaires des descriptions de sites web. Différence essentielle…
– L'annuaire consiste en une sélection de sites web, choisis par des humains, référencés dans un thésaurus structuré en différentes catégories. À l'inverse, le moteur de recherche est une énorme collection (index) de milliards de pages web. Son ambition est d'être le plus exhaustif possible…

3

Préparation du référencement

Dans les chapitres précédents, nous avons passé en revue un certain nombre de notions importantes à connaître dans le cadre d'un référencement. L'heure est venue de passer à l'étape suivante qui consiste à définir la stratégie à mettre en place.

Tout d'abord, ayez toujours à l'esprit qu'il est toujours plus simple et plus efficace – vous vous en apercevrez rapidement – d'aborder la question du référencement lors de la création ou de la refonte d'un site. En effet, un travail minutieux et pertinent demande, dans ce cadre, des modifications parfois importantes dans le contenu et la structure du site.

Mais sachez que si votre site est déjà en ligne et que vous ne désirez pas le modifier outre mesure et dans son essence même, il reste possible d'apporter bon nombre de modifications pour obtenir une meilleure visibilité sur les moteurs. Il n'en reste pas moins vrai que le meilleur rendement sera obtenu en conjuguant refonte ou création du site avec les travaux d'optimisation pour les outils de recherche…

Nous allons essayer dans ce chapitre de vous donner des pistes de réflexion sur la façon dont il vous faudra aborder un certain nombre de points cruciaux pour la promotion de votre source d'information.

Méthodologie à adopter

Tout d'abord, il vous faudra mettre en place une méthodologie simple et efficace pour votre référencement. Voici une liste chronologique des actions à mener.

1. Choix des mots-clés.
2. Choix des moteurs et éventuellement des annuaires à prendre en compte.

3. Création ou modification des pages du site en fonction de ces mots-clés et des critères de pertinence des moteurs.
4. Phase de prise en compte des pages par les moteurs à l'aide de liens savamment créés et vérification de la présence des pages dans les index des moteurs de recherche.
5. Vérification du positionnement et/ou du trafic généré par les outils de recherche.
6. Suivi de ces phases de positionnement/trafic et corrections éventuelles pour obtenir de meilleurs résultats.

C'est donc exactement cette trame que nous avons choisie pour les chapitres qui vont suivre dans cet ouvrage.

Choix des mots-clés

Pour mettre en place une stratégie de référencement, la première phase consiste à choisir les « bons » mots-clés pour positionner vos pages web. Contrairement à ce que l'on pourrait croire, ce n'est pas si simple. Il s'agit d'une phase cruciale pour votre référencement : choisir des mots-clés sur lesquels un positionnement est trop complexe peut s'avérer désastreux ; tout comme le fait d'opter pour des termes qui ne sont jamais saisis par les internautes…

Les mots-clés que vous allez choisir sont extrêmement importants et doivent répondre à deux notions essentielles.

- **L'intérêt.** Ils doivent être souvent (le plus possible) tapés par les utilisateurs des moteurs de recherche. Ce n'est pas toujours le cas.
- **La faisabilité.** Il doit être techniquement possible de positionner une page web dans les premiers résultats des moteurs pour ce terme dans des délais acceptables. Ce n'est, là encore, pas toujours le cas, en tout cas dans des délais « raisonnables »…

Bien sûr, les termes choisis doivent décrire votre activité et le contenu de votre site web, cela va sans dire… Nous allons développer toutes ces notions dans les paragraphes suivants.

Le concept de « longue traîne »

L'objectif de cette première partie stratégique sera de déterminer pour quels mots-clés votre site peut et doit être optimisé dans le cadre de la « tête » de la longue traîne. En effet, comme le montre la figure 3-1, on s'aperçoit le plus souvent en regardant les statistiques d'un site web que :

- environ 20 % du trafic « moteurs de recherche » (tête de la longue traîne) est constitué par des mots-clés très souvent saisis sur les moteurs et pour lesquels le site est optimisé et bien positionné. Ceci représente un nombre relativement faible de mots-clés (quelques dizaines), chacun d'eux générant un fort trafic ;

- environ 80 % du trafic « moteurs de recherche » est constitué par la queue de la longue traîne et des requêtes – générées par le contenu des pages web – saisies peu souvent sur les moteurs pour trouver le site. Ceci représente un nombre important de mots-clés, chacun d'eux générant un faible trafic, mais leur somme globale représentant la majorité du trafic « moteurs »…

Principe de la longue traîne

Le principe de la longue traîne est apparu à Chris Anderson lorsqu'il a pu explorer les statistiques de ventes de sites web de commerce électronique. En inspectant une courbe présentant en abscisse les produits vendus et en ordonnée le nombre de ventes, il s'est rapidement aperçu que la courbe pour chacun des sites étudiés ressemblait à ceci.

Figure 3-1

Le concept de longue traîne sur les sites web de commerce électronique, vu par Chris Anderson

La partie rouge (la « tête » de la longue traîne) représente les best-sellers : peu de produits très populaires représentant de très nombreuses ventes pour chaque référence. La partie jaune (la « queue » de la longue traîne) est représentative de produits peu vendus (parfois une à deux ventes par mois) individuellement mais en très grand nombre de références différentes.

Le concept de longue traîne devient très intéressant lorsqu'on s'aperçoit qu'en fait, c'est la « queue » de cette longue traîne qui génère le plus souvent 80 % du chiffre d'affaires de ce type de site (bien que ce chiffre fasse aujourd'hui débat aux États-Unis)... C'est ainsi un très grand nombre de produits vendus très peu souvent qui, par leur masse, représenterait la majeure partie du bénéfice d'un site web tels que ceux inspectés par Chris Anderson.

De plus, si ce concept se vérifie pour, par exemple, des livres papier, il n'en reste pas moins vrai qu'il existe un besoin impératif de stockage de ces exemplaires papier, et ce stockage représente un coût... L'idée de longue traîne est donc encore plus intéressante pour des produits numériques (films, musique, études au format PDF, etc.) pour lesquels le coût de stockage est aujourd'hui quasi nul... Si vous arrivez à proposer en ligne tout ce

qu'il est possible d'écouter en termes de musique sur la planète, soit des millions, voire des milliards de morceaux, le constat de la longue traîne fait que vous pouvez être quasiment sûr que chacun d'entre eux sera au moins acheté une fois sur une année, générant ainsi un chiffre d'affaire considérable...

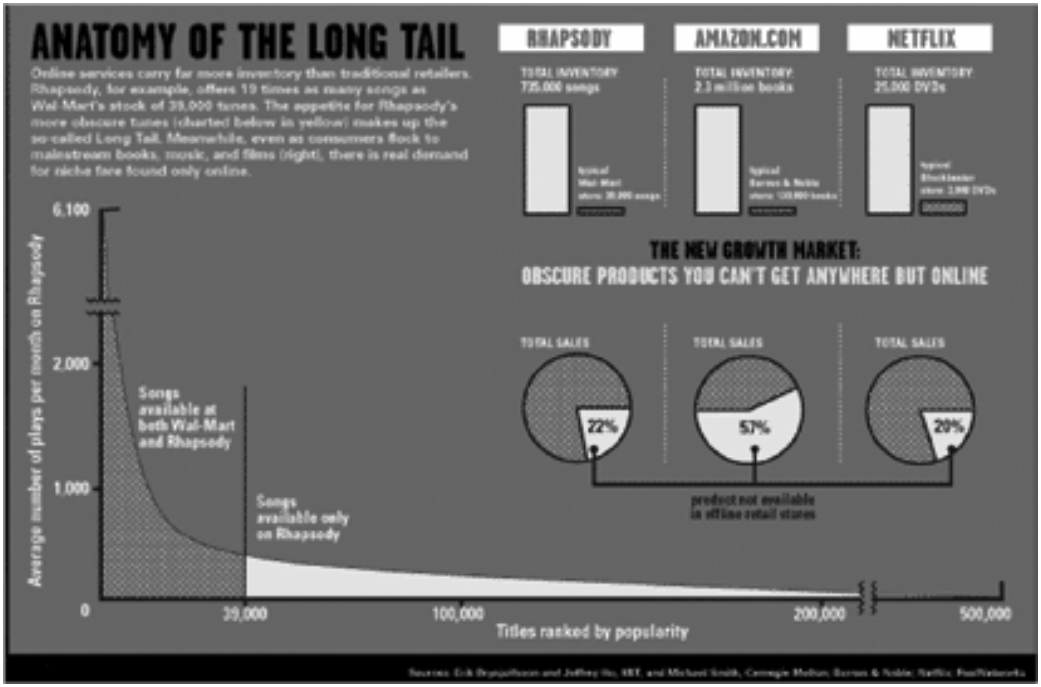


Figure 3-2

Anatomie de la longue traîne

Autre exemple : Chris Anderson, dans son livre *The Long Tail* (paru en 2004 chez Hyperion) consacré à ce phénomène, explore également rapidement le monde des moteurs de recherche, notamment en regardant les statistiques des mots-clés saisis sur le moteur Excite en 2001.

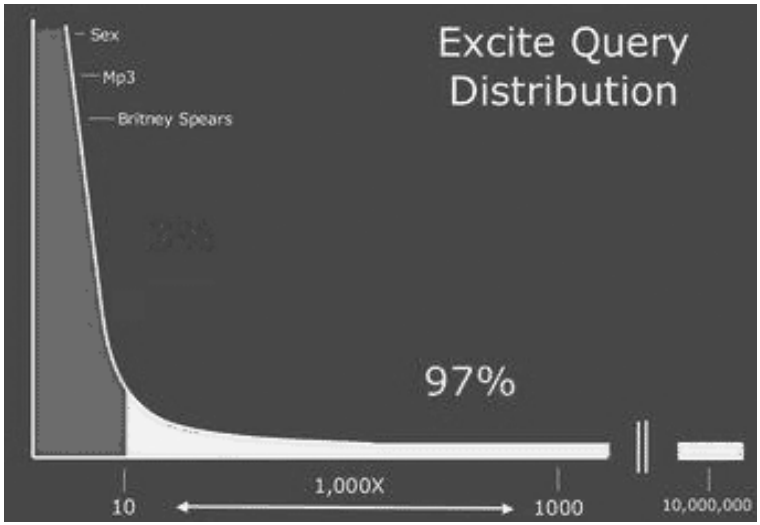
Il s'est rapidement aperçu que, chaque mois, 3 % de toutes les recherches effectuées sur ce moteur se focalisaient sur une dizaine de mots (« sex », « mp3 », « britney spears », etc.), le reste des requêtes était réparti sur des dizaines de millions d'autres termes et expressions... Les moteurs de recherche sont donc typiquement des outils basés sur un concept de « longue traîne ».

D'ailleurs, le marché publicitaire sur les liens sponsorisés, permettant de mettre potentiellement en place des enchères sur des millions de mots, est là aussi un marché type de longue traîne... Et, quelque part, les moteurs de recherche, basés sur l'exploration des

pages web, ne sont-ils pas la longue traîne des annuaires, qui n'inspectent que des fiches descriptives de sites ? Eric Schmidt, PDG de Google, n'a-t-il pas dit, lors de la première assemblée générale des actionnaires du moteur, que la mission de sa société était d'être « au service de la longue traîne » ? On ne saurait être plus clair...

Figure 3-3

Les mots-clés sur un moteur de recherche répondent à une logique de longue traîne.



La longue traîne et le référencement : l'exemple du site Abondance.com

Nous venons d'expliciter rapidement (nous ne pouvons que vous encourager à lire le livre passionnant de Chris Anderson pour en savoir plus) ce concept de longue traîne qui révolutionne actuellement plusieurs concepts de l'économie numérique... Mais il existe un point qui n'est pas abordé dans cet ouvrage : l'application de ce concept au monde du référencement.

Ainsi, nous avons dans un premier temps inspecté les statistiques « mots-clés » du site Abondance (*http://www.abondance.com*). Dans l'immense majorité des outils statistiques disponibles sur le marché, il existe une catégorie d'informations indiquant avec quels mots-clés les internautes ont trouvé votre site sur les moteurs de recherche. C'est le cas de notre outil de statistiques. Nous avons ainsi recoupé des informations sur les 25 296 requêtes qui avaient permis de trouver le site web en question sur les différents moteurs de recherche du Web pendant un mois donné.

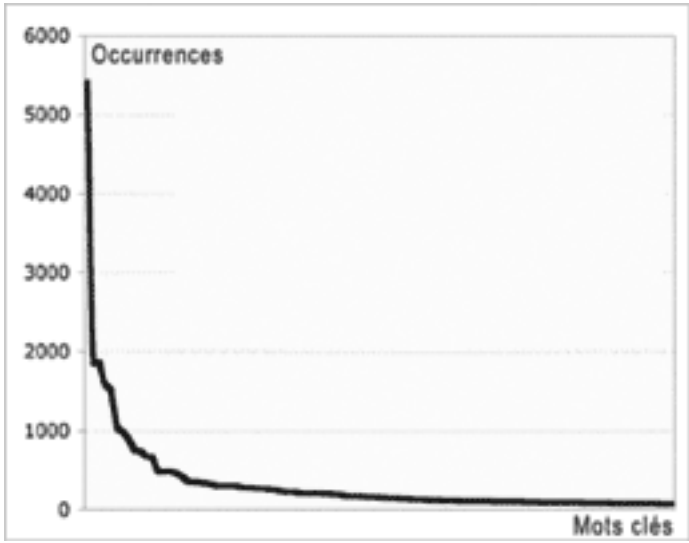
Il s'est rapidement avéré que :

- Moins de 10 requêtes ont généré chacune plus de 1 % du trafic « moteurs » : « abondance » (6,4 %), « miserable failure » (2,2 %), « virtual earth » (2,2 %), « moteurs de recherche » (1,8 %), « robots.txt » (1,7 %), « Google video » (1,2 %), « YouTube » (1,1 %), « humour » (1 %).

- Si l'on ne prend en compte que les 100 premières requêtes les plus populaires, la courbe est déjà typiquement celle d'une longue traîne.

Figure 3-4

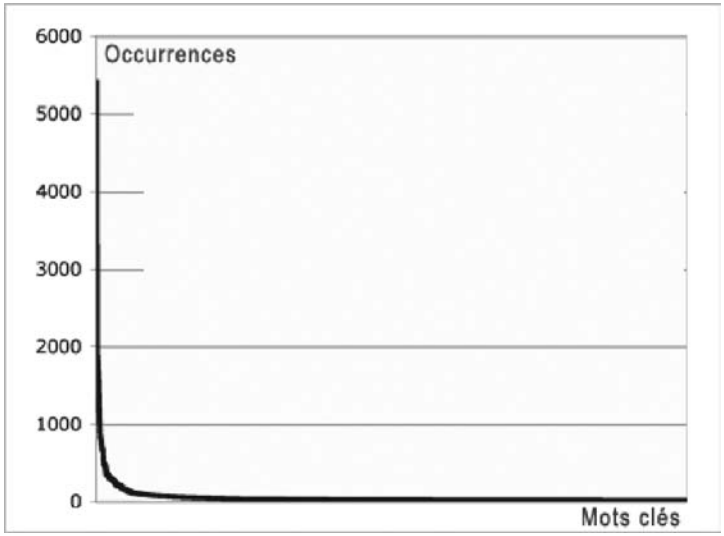
Mots-clés ayant permis de trouver le site Abondance.com sur les moteurs de recherche pendant un mois : 100 premières requêtes



- Si l'on prend, cette fois, les 1 000 premières requêtes, le phénomène est toujours le même.

Figure 3-5

Mots-clés ayant permis de trouver le site Abondance.com sur les moteurs de recherche pendant un mois : 1 000 premières requêtes



- Les 10 premiers mots-clés ont généré 19,77 % du trafic « moteur » (soit 16 637 visites sur les 84 124 amenées par les moteurs de recherche sur le mois testé), les 80,23 % restant étant constitués par 25 286 expressions différentes, représentant chacune moins de 0,8 % du trafic global... Un pur phénomène de longue traîne...
- La dernière requête à générer au moins 0,1 % du trafic (« moteur recherche video ») occupe le 84e rang. Ainsi, à partir de la 85e expression, elles génèrent toutes moins de 0,1 % du trafic « moteur » total...
- La 1 000e requête identifiée (exemples, du 990e au 1 000e rang : « solution paiement en ligne », « recherche en langue arabe », « msn recherche », « mon google », « barre d'outils », « localisé », « moteurs de recherche vidéos », « référencement sur moteur de recherche », « annuaire besançon », « le top », « top mot-clé ») génère encore 7 visites par mois sur le site...
- La 4 000e requête (« adsense rss ») est encore demandée deux fois dans le mois...

Voici le nombre de fois où le site a été trouvé pour des mots-clés sur les moteurs de recherche.

Tableau 3-1 Nombre de visites générées sur le site Abondance.com par des requêtes issues de moteurs de recherche

Nombre de visites générées	Nombre de requetes différentes
1 visite	19 334
2 visites	2 931
3 visites	927
4 visites	499
5 visites	304
...	...

Ainsi, le site a été trouvé, sur un mois, 19 334 fois grâce à un mot-clé qui a généré une visite unique. Une paille qu'il est pourtant difficile de négliger en termes de trafic...

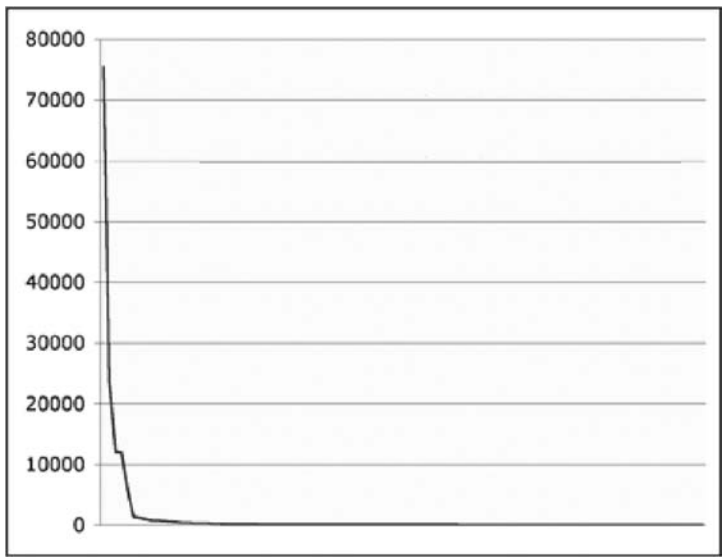
La longue traîne et le site Googlefight.com

Nous avons effectué le même travail sur un autre site du Réseau Abondance, Googlefight.com, un jeu autour de Google, plus orienté « grand public » et qui génère plus de visites (6 millions de pages vues par mois en moyenne) qu'Abondance qui est pour sa part plutôt voué à un public professionnel (bien qu'affichant un trafic de plusieurs millions de pages vues mensuelles)... De plus, ce site contient très peu de contenu, contrairement à Abondance.com. Le phénomène observé est pourtant exactement le

même et les courbes obtenues strictement identiques à celles typiques de la longue traîne.

Figure 3-6

Mots-clés ayant permis de trouver le site Googlefight sur les moteurs de recherche pendant un mois : 100 premières requêtes



À ceci près que les requêtes « googlefight » et « google fight » représentent cette fois près de 70 % des requêtes (on comprend ici l'influence sur ces chiffres du manque de contenu du site)... Une tête de longue traîne très bien fournie donc. En revanche, les chiffres descendent très vite ensuite avec la 200e recherche qui n'est plus demandée que 8 fois dans le mois, puis pour des milliers d'expressions qui ne génèrent qu'une ou deux visites mensuelles... Par ailleurs, le nombre total de requêtes ayant généré au moins une visite sur le site *via* les moteurs de recherche (quelques milliers) est bien plus faible que les 25 296 du site Abondance.com. Absence de contenu oblige...

Ainsi, en étudiant les statistiques des sites Abondance.com et Googlefight.com, nous avons pu nous apercevoir à quel point la théorie de la longue traîne pouvait très facilement s'appliquer au référencement et aux statistiques de trafic sur son site généré par les moteurs de recherche. D'autres recherches et études que nous effectuons depuis de nombreux mois dans un cadre plus large sur d'autres sites web, bien plus connus qu'Abondance, semblent aujourd'hui donner exactement les mêmes résultats statistiques. Une bonne raison pour tenter d'extrapoler ce phénomène dans le cadre d'une stratégie de référencement globale. Nous y reviendrons largement au chapitre 10 lorsqu'il s'agira de mesurer l'efficacité d'une stratégie de référencement...

Extrapolation de la longue traîne dans le cadre d'une stratégie de référencement

Lorsqu'on met en place une stratégie de référencement pour un site web, l'un des premiers réflexes que l'on a est de définir les mots-clés sur lesquels on va tenter de positionner son site. En faisant cela, on va en fait « nourrir » la tête de la longue traîne avec

des mots-clés que l'on désire avant tout voir comme des « best-sellers ». C'est, en revanche, le contenu textuel de votre site qui va nourrir la queue de la longue traîne...

Ainsi, si l'on désire optimiser son référencement en prenant en compte ce phénomène de longue traîne, il faudra :

- Choisir ses mots-clés de départ (ceux qui décrivent votre métier, vos thématiques, votre univers sémantique global) avec soin, sachant qu'il y a de fortes chances pour que ces mots-clés choisis pour le positionnement ne représentent que 20 % du trafic généré par les moteurs de recherche. Cependant, il y a également de fortes chances pour que ce trafic soit bien ciblé et de « bonne qualité » puisqu'il correspondra à des mots représentant votre activité et que vous avez choisis comme étant pertinents pour trouver un site comme le vôtre... À vous donc d'optimiser certaines pages de votre site pour les mettre en valeur. La stratégie à adopter est donc, chronologiquement parlant :
  - choix des mots-clés ;
  - optimisation de certaines pages du site en fonction de ces mots-clés ;
  - mesure du trafic généré spécifiquement sur ces requêtes.

  On parlera ici de « trafic maîtrisé »...
- Optimiser la structure des pages de vos sites pour qu'elles mettent en valeur leur contenu éditorial afin de favoriser la « queue » de la longue traîne. Dans ce cas, vous ne maîtrisez, par définition, pas vraiment les positionnements obtenus, donc un certain pourcentage de ces expressions générera un trafic « stérile ». Mais il y a fort à parier que, dans le lot, certains mots-clés soient très intéressants, même s'ils n'ont pas été imaginés pour cela dès le départ. Dans ce cadre, la stratégie à mettre en œuvre est différente :
  - optimisation du code HTML (voir chapitres 4 et 5), des « masques » ou « templates » de vos pages ;
  - écriture du contenu si possible en tenant compte des contraintes des moteurs (voir, là aussi, la suite au chapitre 4) ;
  - mesure du trafic global généré par ces termes « secondaires », quantitativement et qualitativement.

  Le trafic dans ce cas devient « opportuniste » (ce qui n'a rien de péjoratif...).

On le voit, les deux stratégies sont très complémentaires.

Il est en tout cas certain que le concept de longue traîne ne peut pas, en 2009, être ignoré dans le cadre d'un référencement. Il faut absolument, pour mettre en place une stratégie gagnante de visibilité sur les moteurs de recherche, soigner à la fois les mots-clés qui vont nourrir la tête, comme la structure du site qui va mettre en avant les mots-clés de la queue de cette *Long Tail*. Un travail certes long parfois mais qui peut rapidement porter ses fruits de façon très efficace...

Figure 3-7

Le concept de la longue traîne appliqué au référencement

Les requêtes s'allongent

Deux études très intéressantes nous montrent que les requêtes saisies par les internautes sur les moteurs de recherche ont tendance à s'allonger au fil des années.

– RankStat (*http://findarticles.com/p/articles/mi_hb5243/is_200701/ai_n19665542/*) indiquait en 2007 que ces requêtes s'effectuaient sur un seul mot pour 13,48 % seulement des cas, et le plus souvent sur deux mots (28,38 %) et trois mots (27,15 %). Plus globalement, les requêtes contenaient de trois à cinq mots (51,60 %), voire de quatre à dix mots (30,98 %). Un tiers des requêtes s'effectue donc sur des expressions de plus de trois termes et la moitié sur plus de deux.

– De même, Hitwise (*http://searchengineland.com/search-queries-getting-longer-16676*), en 2009, expliquait que les requêtes sur un et deux mots-clés avaient perdu, en un an, respectivement 3 et 5 % d'occurrences alors que les requêtes sur sept et huit mots ou plus gagnaient dans le même temps 12 et 22 % !

Les requêtes longues, sur de nombreux mots-clés, qui nourrissent le plus souvent la longue traîne, sont donc tout à fait d'actualité dans le cadre d'un référencement.

Comment trouver vos mots-clés ?

Avant de prendre en compte l'intérêt et la faisabilité d'un mot-clé, encore faut-il trouver ce dernier… Il est donc important d'identifier plusieurs moyens qui vous permettront de prendre en compte des termes sur lesquels il vous semble intéressant de positionner votre site. Comment faire ? Voici quelques pistes.

1. **L'intuition.** Certains mots-clés peuvent vous venir automatiquement à l'esprit lorsque vous pensez à votre activité (ne serait-ce qu'en ce qui concerne votre marque...). Notez-les précieusement. Mais rien ne dit que les mots-clés que vous imaginez seront obligatoirement ceux utilisés par les internautes lorsqu'ils chercheront un site tel que le vôtre. Votre vision de votre métier et de votre entreprise, parfois très interne et personnelle, peut être différemment perçue par un internaute *lambda* ou un prospect… On a parfois pas mal de surprises à ce niveau *a posteriori*. Cela dit, la piste intuitive est souvent excellente, ne la négligez donc pas mais ne vous basez pas non plus uniquement sur elle.

2. **Les bases de données.** Il existe des bases de données de mots-clés comme Wordtracker (*http://www.wordtracker.com/*) ou Keyword Discovery (*http://www.keyworddiscovery.com/*) qui peuvent vous aider à identifier les termes les plus intéressants. Certains outils de recherche proposent également en ligne un palmarès des termes les plus souvent demandés sur leur outil, comme le « Lycos Top 50 Searches » (*http://www.lycos.com/*). Mais ces dernières listes ne vous aideront pas vraiment puisqu'elles ne proposent qu'une suite limitée de termes très souvent demandés. Il y a peu de chances que vous y trouviez votre bonheur… En revanche, Wordtracker ou Keyword Discovery sont plus complets, mais payants et assez souvent limités en ce qui concerne les mots-clés en français (bien que Keyword Discovery soit bien plus performant que son concurrent à ce niveau). Ils sont cependant très pertinents pour la langue anglaise. À vous de les tester et de faire votre choix !

   Parmi les outils de type « générateurs de mots-clés » qui vous proposent des termes de recherche contenant un mot préalablement saisi, voici quelques exemples de ce que l'on peut trouver en ligne.

   Tout d'abord, les prestataires de liens publicitaires sponsorisés proposent tous des outils permettant d'identifier des mots-clés souvent saisis sur leur réseau de sites partenaires. Ils fournissent deux types d'informations :

   – Le nombre de fois où la requête a été demandée sur les moteurs de recherche sur lesquels ils affichent leurs liens sponsorisés. Par exemple, Microsoft adCenter indiquera les chiffres de Bing, MSN, etc. En revanche, il ne tiendra pas compte du trafic issu de Google, ce qui est important à savoir… Et le générateur de mots-clés de Google n'affichera pas de statistiques sur ses concurrents, etc.

   – Des expressions connexes contenant le mot initialement demandé. La requête « référencement » proposera ainsi « référencement gratuit », « référencement site », « référencement Internet », etc.

Yahoo! Search Marketing (*http://searchmarketing.yahoo.com/fr_FR/yahoo-search-marketing.php*) et Microsoft adCenter (*https://adcenter.microsoft.com/default.aspx?mkt=fr-fr*) proposent ce type d'outil directement dans leur interface de création et de gestion de liens sponsorisés. En revanche, il faut être dûment enregistré auprès de ces services pour utiliser ces outils. Le générateur de Espotting/Miva n'est quant à lui plus disponible, cette société n'ayant plus d'activité en France. L'outil de Google est libre d'accès, pour sa part. Cela tombe bien, il est, et de loin, le meilleur... C'est donc celui-ci que nous allons décrire dans les pages suivantes.

Le générateur de mots-clés d'Overture

Longtemps, les fans de référencement ont utilisé le générateur de mots-clés d'Overture (Yahoo! Search Marketing) pour identifier les meilleures clés de positionnement. Mais, fin 2007, cet outil a été abandonné pour être intégré à la plate-forme Panama (*http://searchmarketing.yahoo.com/fr_FR/*) de liens sponsorisés, dans une version très décevante, laissant la voie libre à l'outil de Google...

Une autre famille d'outils regroupe ceux qui proposent, lors d'une saisie dans un formulaire de recherche, des expressions connexes à la volée. Il existe un grand nombre de ces outils, tels que :

- Google Suggest, sur la page d'accueil du moteur – *http://www.google.fr/* (voir ci-après)
- KwMap – *http://www.kwmap.com/*
- WikiWax – *http://www.wikiwax.com/*
- Snap – *http://www.snap.com/*

Enfin, certains outils ont été créés pour vous aider à trouver des mots-clés pertinents en partant de vos termes de départ. En voici quelques exemples :

- Outiref – *http://www.outiref.com/*
- Wordtracker – *http://www.wordtracker.com/*
- Keyword Discovery – *http://www.keyworddiscovery.com/*
- SEO Book – *http://tools.seobook.com/general/keyword/*
- Search Combination Tool – *http://www.webuildpages.com/search/*
- Good Keywords (logiciel) – *http://www.goodkeywords.com/*
- TheFreeDictionnar – *http://www.thefreedictionary.com/*
- WebRankInfo – *http://www.webrankinfo.com/outils/semantique.php* ; *http://www.webrankinfo.com/outils/expressions.php*
- Dictionnaire de synonymes – *http://elsap1.unicaen.fr/cgi-bin/cherches.cgi*

Autres ressources

Cette liste n'est bien sûr pas exhaustive. De plus, le Web étant ce qu'il est, certains outils peuvent être devenus inopérants au moment où vous lirez ces lignes… D'autres sites vous sont proposés ici : *http://ressources.abondance.com/generateur-mot-cle.html*. Voir également une liste d'outils intéressants en annexe du présent ouvrage.

3. **Les sondages internes ou externes.** Vous pouvez demander à des connaissances, des amis ou des collègues de bureau quels sont les termes qui leur viendraient à l'esprit pour rechercher une activité ou un produit comme les vôtres sur le Web.
4. **Les résultats sur les moteurs de recherche.** Tapez un certain nombre de mots-clés concernant votre activité sur des outils comme Google, Bing ou Yahoo!. Regardez les résultats proposés par l'annuaire ou le moteur : ils contiennent certainement des termes auxquels vous n'aviez pas pensé au départ.

5. **Les *Related Searches*.** Sur des outils de recherche comme AltaVista (*.com* ou *.fr*), Google, Exalead ou Yahoo.com, le moteur propose, dans ses pages de résultats, des *Related Searches*. Comme vous pouvez le voir en figure 3-8, ce sont des suites de deux ou trois termes contenant – ou non – le mot demandé au départ. Ces expressions sont issues de bases de données statistiques sur les mots-clés les plus demandés par les internautes dans le passé. Ils constituent également des informations très pointues. Nous en reparlerons très bientôt...

Figure 3-8

Exemple en bas de page de résultats de Google pour la requête « moteur » : des synonymes et suggestions connexes sont proposés lors d'une recherche.

6. **L'audit de la concurrence.** Rien ne vous empêche de consulter les balises meta `keywords` de vos sites concurrents (s'ils en utilisent encore...). Au moins, ces champs serviront à quelque chose dans le cadre du référencement… de leurs concurrents. Mais, chut !, nous ne vous avons rien dit…
7. Pensez aux **fautes d'orthographe** et aux **fautes de frappe** sur votre nom ou vos mots-clés essentiels. Cela peut générer un trafic important…

Utiliser Google Suggest pour trouver les meilleurs mots-clés

Depuis l'été 2008, Google propose sur sa page d'accueil l'outil « Google Suggest » qui affiche, au fur et à mesure de la frappe d'un mot-clé dans le formulaire de recherche, des propositions de requêtes (système dit d'« autocomplétion »). Comment ces expressions sont-elles choisies par le moteur ? Est-il le seul outil de recherche à proposer une telle fonctionnalité ? Comment se servir de ces données pour optimiser le référencement de son site web ? Voici quelques pistes d'exploration...

Comme souvent chez Google, les expérimentations ont précédé le lancement officiel d'une nouvelle fonctionnalité. Un article daté du mois d'août 2008 (*http://googleblog.blogspot.com/2008/08/at-loss-for-words.html*) a commencé à parler du déploiement de Google Suggest sur le portail web. En réalité Google Suggest était proposé par Google Labs depuis décembre 2004 (voir *http://actu.abondance.com/2004-51/google-suggest.html*) !

Un autre article (*http://googleblog.blogspot.com/2008/09/update-to-google-suggest.html*), paru en septembre 2008, expliquait le fonctionnement de l'outil, puis en mars 2009, Google annonçait l'arrivée des suggestions localisées dans le champ de recherche (*http://googleblog.blogspot.com/2009/03/local-flavor-for-google-suggest.html*).

Aujourd'hui, Google Suggest est en place et la fonctionnalité est rentrée dans les mœurs. Toute requête tapée dans le champ de recherche s'accompagne d'une liste de suggestions, basées sur le mot-clé tapé.

Figure 3-9

Les Google Suggest sur la page d'accueil du moteur de recherche

Cette fonctionnalité peut facilement être désactivée en allant dans la page Préférences de Google.

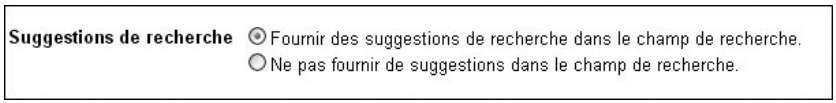


Figure 3-10

Désactivation des Google Suggest dans les préférences du moteur

L'outil Google Suggest est proposé sur le portail de recherche web, mais aussi sur Google Images, Google Vidéo et YouTube.

Outre les suggestions de recherche, Google Suggest est également capable de corriger les requêtes des utilisateurs, ce qui apporte un confort d'utilisation indéniable.

Le nombre de résultats de recherche est purement indicatif et il se révèle le plus souvent différent du nombre affiché après avoir effectué une requête.

Figure 3-11

Google Suggest peut corriger des fautes d'orthographe dans les saisies de mots-clés.

Google Suggest : comment ça marche ?

Pour avoir des informations sur le fonctionnement de ce service, le plus simple est de consulter la documentation disponible sur Google Labs (*http://labs.google.com/suggestfaq.html*) :

« Notre algorithme utilise un large champ d'informations pour prédire les requêtes que les utilisateurs aimeraient voir s'afficher. Par exemple, Google Suggest utilise des données sur la popularité de différentes recherches pour aider à classer les suggestions. Un exemple de ce type d'information de popularité peut être trouvé sur Google Zeitgeist. Google Suggest ne base pas ses suggestions sur votre historique personnel de navigation.

Nous essayons de ne pas proposer de requêtes qui pourraient être offensantes pour une large audience d'utilisateurs. Ceci inclut les mots explicitement pornographiques ainsi que des requêtes qui pointent vers des sites pornographiques, les mots grossiers, les termes de haine et de violence. »

Google Suggest propose donc des expressions en fonction de leur popularité, et en excluant les expressions jugées inappropriées. C'est cette notion de « popularité » qui est difficile à appréhender...

Un article de l'excellent site SEO by the Sea, intitulé *Can Google Read Your Mind? Processing Predictive Queries* (*http://www.seobythesea.com/?p=69*), évoque la fréquence de requêtes utilisateur, la fraîcheur des requêtes et l'utilisation d'un dictionnaire pour sortir les termes suggérés. L'outil aurait été développé après le rachat de Kaltix en 2003, une société qui travaillait sur la personnalisation des résultats de recherche (voir *http://c.asselin.free.fr/french/aout03/Kaltix.htm*).

Il est en tout cas certain que Google Suggest propose des résultats rafraîchis en permanence, en fonction de l'activité des internautes sur le Web. Il s'agit d'un outil dynamique qui montre les recherches les plus populaires en rapport avec un terme donné, de façon purement algorithmique.

Concernant les concepts et thèmes associés à une requête, Google propose également, comme on l'a vu précédemment, des recherches apparentées et des recherches associées qui sont proposées directement dans les résultats de recherche, et qui constituent une bonne source pour identifier des mots-clés pour votre référencement.

La différence essentielle entre les suggestions de recherche et les recherches apparentées ou associées est que ces dernières proposent des concepts liés à la requête utilisateur, tandis que les suggestions utilisent de façon stricte les mots-clés tapés par les internautes.

Sur la requête « référencement » dans Google, on peut, par exemple, avoir les propositions suivantes, respectivement en haut et en bas de page.

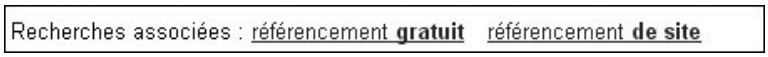


Figure 3-12

Mots-clés associés à la requête, dans les pages de résultats de Google

Figure 3-13

Recherches apparentées, dans les pages de résultats de Google

Les premières associent d'autres mots-clés à celui initialement demandé, alors que les secondes peuvent proposer des termes plus génériques, des synonymes, des concepts, etc. : « comment référencer », « abondance »…

Google Suggest et le référencement

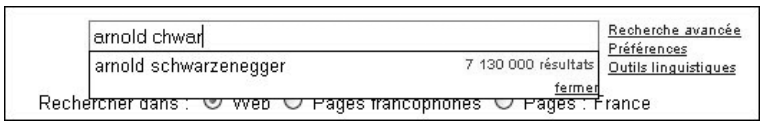
Quel est l'impact de Google Suggest pour le référencement ? Il est clair que ce service largement répandu a des conséquences sur le comportement des internautes.

La correction des requêtes et l'identification de fautes de frappe et d'orthographe

La correction des termes tapés est peut-être l'aspect le plus sympathique de Google Suggest. Si vous ne vous souvenez plus du nom de votre acteur préféré ou si vous êtes fâché avec les complications de l'orthographe française, Google Suggest est votre ami.

Figure 3-14

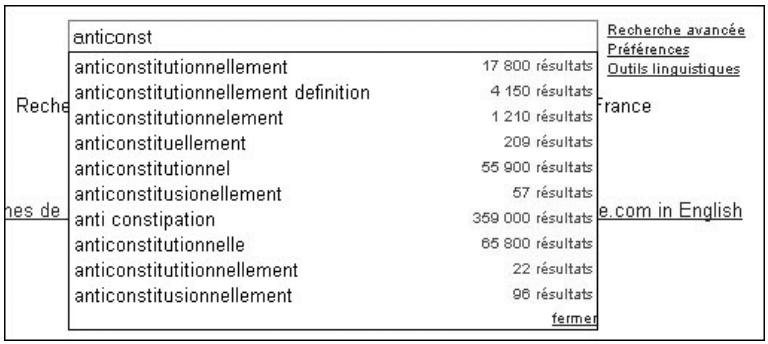
Conan le barbare, c'est plus facile...



Attention néanmoins : Google Suggest n'est pas un correcteur orthographique. Il affiche seulement les requêtes les plus tapées par les internautes. Or, comme il existe de nombreux internautes brouillés avec l'orthographe, il ne faut pas forcément se fier à ce que ressort cet outil... L'exemple suivant est assez parlant.

Figure 3-15

Google Suggest est un reflet de ce que tapent les internautes dans le moteur.



Google Suggest est donc un excellent outil pour identifier des fautes d'orthographe ou de frappe pour votre référencement (voir plus loin dans ce chapitre).

Les internautes auront cependant de plus en plus tendance à cliquer sur le premier résultat Google Suggest, qui est aussi le résultat le plus correctement orthographié (du moins dans la majorité des cas), ce qui peut faire baisser à la longue le trafic sur les « mauvais » mots-clés. Qui s'en plaindra ?

On constate aussi que l'utilisation de majuscules n'a aucun impact sur la suggestion de recherche, Google prenant le parti de tout proposer en minuscules.

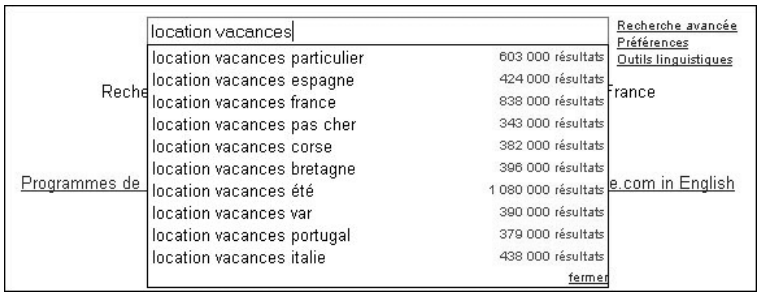
Les requêtes composées

À partir d'un simple mot-clé, Google Suggest propose une liste de mots-clés composés, ce qui a un impact certain sur le référencement. L'aspect positif est qu'il est désormais possible de se positionner sur une requête fortement concurrentielle. Prenons l'exemple du terme ultragénérique « location vacances » (pas moins de 115 millions de résultats de recherche).

Ce terme est difficilement positionnable pour un site *lambda*, mais fort heureusement, Google Suggest propose plusieurs expressions beaucoup plus abordables lorsque l'on tape la requête.

Figure 3-16

Expressions le plus souvent demandées sur Google dans le domaine de la location de vacances



Dans cet exemple, nous pouvons profiter de la forte fréquence de frappe sur « location vacances » pour faire apparaître nos pages optimisées sur « location vacances

particulier » ou « location vacances France », requêtes moins concurrentielles et sur lesquelles des résultats pourront être espérés en moins de temps.

Attention cependant à quelques « effets de bord » de cet outil.

Point numéro 1 : Google Suggest favorise les requêtes complexes, ce qui peut profiter aux sites positionnés sur des mots-clés ciblés. Le revers de la médaille est que la liste de suggestions proposées par Google est très limitée : il s'agit d'un véritable appauvrissement de l'effet « longue traîne ». Le nombre de mots-clés connexes à une requête se limite à une dizaine de termes choisis par Google.

Reprenons l'exemple précédent et imaginons que notre stratégie de référencement porte sur l'expression « location vacances bord de mer ». Cette expression n'est pas affichée par Google Suggest, elle sera donc beaucoup moins réactive que les expressions proposées aux internautes.

Point numéro 2 : Google Suggest diminue la visibilité d'un site sur des expressions longue traîne. On se trouve ici devant un cercle vicieux, car plus une requête est tapée par les internautes, plus elle est populaire et a de chances d'apparaître dans Google Suggest. Les expressions complexes, qui sont peu tapées et donc moins populaires, risquent donc de passer dans les oubliettes du référencement...

On sait heureusement que les résultats Google Suggest sont régulièrement rafraîchis, mais rien ne dit que l'expression « location vacances bord de mer » sera suffisamment visible pour générer du trafic.

Le cas cité dans notre exemple est cependant très particulier : dans la majorité des cas, les internautes ne vont pas utiliser Google pour choisir leur destination de vacances ! On sait qu'ils vont taper des requêtes ciblées (par exemple, « location vacances en Bretagne »), ce qui devrait diminuer les « dégâts collatéraux » sur le positionnement des sites dans Google.

Le contrôle de l'information

Il est absolument impossible de contrôler les informations proposées par Google Suggest, celui-ci utilisant les requêtes les plus tapées par les internautes.

La preuve : dans ce contexte, Google peut se transformer en outil d'incitation au téléchargement illégal.

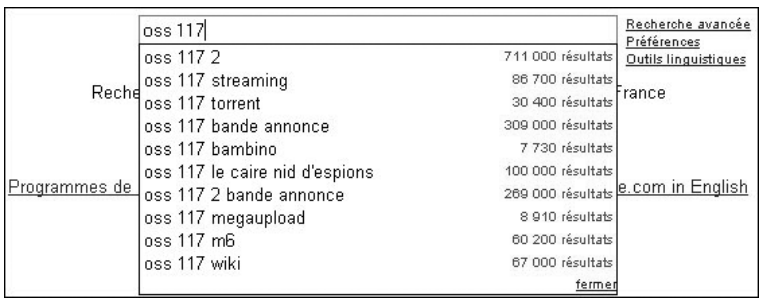


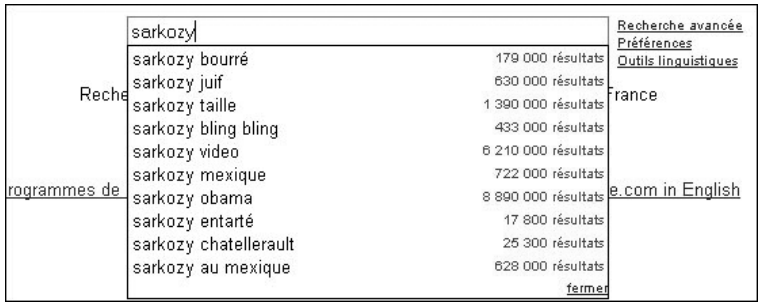
Figure 3-17

Vous avez envie de télécharger un film de Jean Dujardin ?

Ou même en outil de dénigrement de personnalité politique.

Figure 3-18

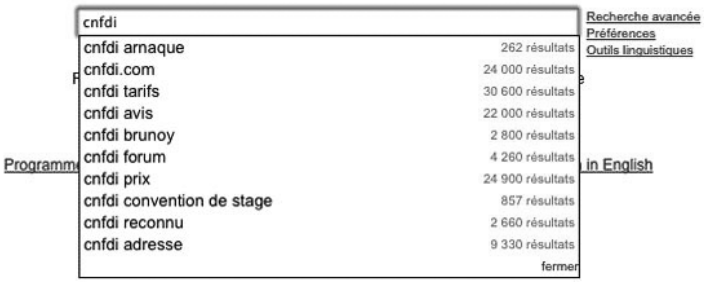
Sans commentaires...



On passe sur d'autres exemples qui ont défrayé la chronique du petit monde des moteurs de recherche en 2009 (*http://actu.abondance.com/2009/07/google-suggest-condamne.html* et *http://actu.abondance.com/2009/07/google-suggest-nouveau-jugement-cette.html*).

Figure 3-19

La preuve que les Google Suggest sont automatisées !



Ces exemples peuvent faire sourire (ou pas, selon de quel côté on se trouve)... Mais il est évident que ce type de requêtes « populaires » peut gravement nuire à l'image d'une personnalité ou d'une marque, et inciter les internautes à consulter des informations peu pertinentes en regard des objectifs d'une marque ou institution.

Comment maîtriser son image sur le Web ? Cela risque d'être de plus en plus difficile étant donné le développement exponentiel du Web social et la diversité des informations proposées sur Internet.

Dans ce contexte, Google ne fait que relayer la tendance générale et donne un aperçu de la façon dont une marque, une personnalité ou encore un objet culturel sont perçus par les internautes. Difficile de lutter contre cela, et Google Suggest se révèle être un cadeau empoisonné pour les chargés de communication. Mais un outil intéressant pour le référencement... L'outil peut ainsi être utilisé comme un baromètre en temps réel de la popularité sur le Web : pour connaître l'image d'une marque auprès des internautes, l'utilisation de Google Suggest peut s'avérer très précieuse.

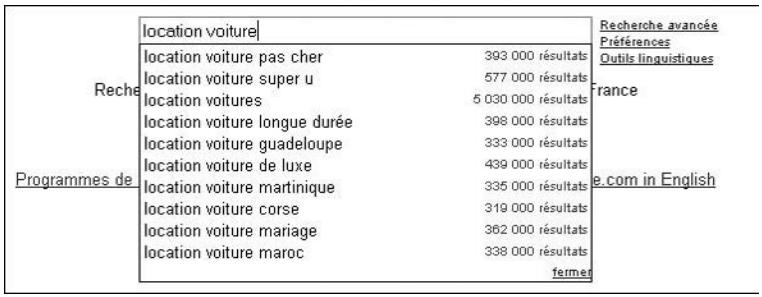
Peut-on donc utiliser les suggestions de recherche pour le référencement ? *A priori*, les requêtes les plus visibles dans Google Suggest ont plus de chances de générer du trafic. Il est donc intéressant de prendre en compte les expressions affichées dans l'interface.

Le problème est le rafraîchissement régulier des suggestions de recherche, qui va impliquer un travail d'analyse sur plusieurs jours, de façon à collecter un maximum de données. Une fois ces données récoltées, il est possible de déterminer des tendances et de connaître les expressions les plus populaires.

Par ailleurs, est-ce que cet outil est plus intéressant que d'autres, tels que le générateur de mots-clés proposé par AdWords (voir plus loin dans ce chapitre) ? Prenons par exemple la requête « location voiture ». Voici ce qui est proposé par Google Suggest.

Figure 3-20

Résultats de Google Suggest pour la requête « location voiture »



Et voici ce qui est proposé par le générateur de mots-clés Google AdWords (sans synonymes, suggestions classées par volume de recherche local décroissant).

Figure 3-21

Résultats du générateur de mots-clés de Google pour la requête « location voiture »

Mots clés	Position prévisionnelle de l'annonce	Concurrence entre annonceurs	Volume de recherche locale : avril	Type de ciblage : Large
Mots clés en rapport avec le(s) terme(s) entré(s) - trié par pertinence				
location voiture	1 - 3		2 740 000	Ajouter
location de voiture	1 - 3		1 220 000	Ajouter
location voitures	1 - 3		165 000	Ajouter
location voiture pas cher	1 - 3		74 000	Ajouter
location voiture a	1 - 3		49 500	Ajouter
location voiture paris	1 - 3		40 500	Ajouter
location voiture aeroport	1 - 3		27 100	Ajouter
location voiture france	1 - 3		27 100	Ajouter
location voiture mariage	1 - 3		27 100	Ajouter
location voiture corse	1 - 3		22 200	Ajouter

Comme on peut le constater, les résultats varient sensiblement, car basés sur des critères de classement légèrement différents. Google Suggest propose en effet des résultats pour un instant *T*, tandis que le générateur Google AdWords repose sur des statistiques mensuelles.

Les deux outils sont donc complémentaires, tout dépend de ce que l'on recherche : un positionnement sur le long terme (générateur AdWords) ou un positionnement « flash »,

dans l'« air du temps » sur les expressions clés les plus populaires du moment (Google Suggest) ?

Un blog ou un portail d'actualité pourrait profiter des données Google Suggest mais pour un site *lambda* peu fréquenté par Google, il vaudra peut-être mieux se focaliser sur les données AdWords.

Autre point à ne pas oublier : l'outil Google Suggest n'est présent (au moment où ces lignes sont écrites) que sur les pages d'accueil des moteurs de recherche de Google. Il n'est pas proposé dans le formulaire présent sur les pages de résultats, qui servent le plus souvent à affiner un résultat. Une donnée qu'il n'est pas inutile de rappeler et qui renforce l'idée que les différentes voies pour trouver des mots-clés pertinents pour un référencement sont encore aujourd'hui très complémentaires...

Fautes de frappe et d'orthographe

Lorsqu'on met en place une stratégie de référencement, il est bien évidemment essentiel de définir au mieux ses mots-clés. Qu'ils soient larges, génériques (comme « audit », « tourisme », « conseil », « DVD », etc.) ou plus précis (« gîte rural auvergne », « expert comptable marseille », « audit hôtellerie agroalimentaire », etc.), l'essentiel sera de bien définir une liste de requêtes représentatives de votre activité, qui pourra être prise en compte dans le cadre d'un référencement naturel.

Mais une stratégie de choix de mots-clés qu'on oublie souvent, et qui rapporte pourtant un trafic loin d'être négligeable, consiste à identifier les fautes de frappe ou d'orthographe, notamment sur le nom de la société (ou organisme) qui édite le site, ou sur des noms de marque ou, enfin, sur des mots-clés importants pour votre activité. Il faut également les prévoir et les prendre en compte car tout le monde ne sait pas obligatoirement comment taper vos patronymes et noms de produits sur les moteurs, ou ne sait pas écrire parfaitement les termes qui correspondent à votre secteur d'activité.

Ainsi, le site Abondance est trouvé sur des mots comme « googol », « googel », « goooogle », « googles », « accona », « cloaking » et « cloacking » (l'auteur de ce livre n'a jamais su écrire ce mot de la bonne façon plus de deux fois de suite...) ou sur le nom « olivier andrieux » (avec un « x »), etc.

Chaque requête représente un trafic relativement faible (certains mots-clés génèrent cependant quelques centaines de visites par mois, voire plus), mais le trafic général que ces fautes d'orthographe ou de frappe représente est loin d'être négligeable (plusieurs milliers de visites mensuelles en tout sur Abondance). Les petits ruisseaux font les grandes rivières... Il faut donc y penser au moment de définir ses mots-clés.

La première difficulté sera d'identifier quelles fautes d'orthographe et de frappe prendre en compte. Une fois ce travail effectué, il vous faudra lancer une optimisation et un référencement idoines pour que les moteurs les prennent en considération. Pas si simple...

Étape 1 – Comment trouver les fautes de frappe et d'orthographe ?

Il existe plusieurs façons, plus ou moins intuitives, pour identifier ces « mauvais » (au sens grammatical du terme) mots. En voici quelques-unes.

- Les générateurs de mots-clés, et notamment celui de Google (*https://adwords.google.com/select/KeywordToolExternal*), peuvent parfois proposer certains mots-clés mal orthographiés, c'est même assez courant sur des fautes classiques (par exemple, « primtemps de bourges »).
- L'analyse des mots-clés *referrers* de votre site web. Si certains mots mal orthographiés sont caractéristiques de votre site, peut-être vous êtes-vous déjà fourvoyé à votre tour et ces mots se trouvent-ils, à votre insu, dans vos pages. Celles-ci peuvent donc être déjà trouvées dans les résultats des moteurs. Vous le saurez vite en regardant vos statistiques, tous les outils de ce type affichent les mots-clés et expressions avec lesquels votre site a été trouvé. Analysez-les, vous y trouverez peut-être (et même sûrement) quelques « perles »...
- L'utilisation de l'outil Google Suggest sur la page d'accueil du moteur de recherche (voir précédemment dans ce chapitre).
- Le sondage ou étude interne : demandez à vos employés, partenaires ou amis quelles sont les fautes qu'ils commettent le plus souvent sur les mots-clés importants pour votre activité.
- Mise en place d'un moteur de recherche interne sur votre site et analyse des mots-clés identifiés.
- Utiliser un générateur de fautes de frappe ou d'orthographe. Il en existe quelques-uns sur le Web, comme :
  - *http://www.seochat.com/seo-tools/keyword-typo-generator/*
  - *http://www.fatfingers.com/*
  - *http://www.fautedefrappe.fr/cgi-bin/typotool/typotool/*
  - *http://aixtal.blogspot.com/2005/07/rcr-pourriss-vos-texte.html*
  - *http://bvwg.actulab.net/fautes-de-frappe.php*
  - *http://www.combell.com/fr/noms-de-domaine/outils-pour-noms-de-domaine/generateur-de-fautes-de-frappe*
  - *http://www.generateur-de-pages.com/generateur-fautes-de-frappe/*
  - *http://www.fautedefrappe.com/old/* (ancienne version...)

    Cette liste n'étant pas exhaustive, vous trouverez également quelques scripts de simulation de fautes de frappe comme *http://www.flashkod.com/codes/SIMULATEUR-FAUTES-FRAPPES_34034.aspx*. À vous de les tester...
- Enfin, le bon sens sera l'un de vos premiers atouts : réfléchissez et notez les fautes « normales » ou « logiques » possibles dans vos mots-clés (inversions, ajouts de lettres, etc.).

Il existe deux types de fautes en langue française :

- les fautes typographiques : omission, addition, substitution, interversion de lettres ;
- les fautes phonographiques, c'est-à-dire sur la base de la prononciation (par exemple, eau => o).

Pour identifier par vous-même des mots erronés, n'hésitez pas à consulter les pages suivantes qui listent les fautes les plus fréquentes en langue française :

- *http://fr.wikipedia.org/wiki/Wikip%C3%A9dia:Fautes_d%27orthographe/Courantes*
- *http://www.osil.ch/eval/node30.html*

Étape 2 – Comment référencer son site sur les fautes d'orthographe et de frappe ?

Vous avez identifié une liste importante de mots erronés pour mieux référencer votre site ? Bravo, c'est une bonne chose. Il faut maintenant les insérer dans vos pages...

Il y a quelques années de cela, cette étape était assez simple, voire élémentaire : les balises meta `keywords` étaient là pour ça ! Il suffisait de remplir cette zone avec les différentes orthographes possibles et hop, le tour était joué, cette zone d'informations étant fournie aux moteurs et lue par ces derniers. Mais ce n'est plus le cas aujourd'hui (voir chapitre 4). Il faut donc trouver d'autres pistes pour les proposer quand même aux outils de recherche. En voici quelques-unes :

- Proposez une page du type « Comment les internautes ont trouvé notre site ce mois-ci »... En plus d'un contenu amusant, vous allez créer une page web dans laquelle vous allez lister les mots-clés – réels bien sûr, mais vous pouvez ensuite faire comme bon vous semble – contenant ou non une faute de frappe ou d'orthographe que vous avez relevée dans vos logs (mémoire de connexion des internautes sur votre site). Cette page, sorte de « Top 100 » des mots-clés qui ont servi à trouver votre site, contenant des « variations inattendues » de vos mots-clés, renforcera votre référencement sur leur intitulé.
- Proposez une page du type « Ce qu'il ne faut pas faire ». Par exemple, un titre du style « Vous ne nous trouverez pas en tapant... » et listant une série de termes mal saisis. Cela laisse ainsi une trace de ces mauvaises orthographes qui seront lues par le moteur et qui lui permettront de vous retrouver (ce qui est certes paradoxal par rapport au titre de la page en question !) si quelqu'un tape ces termes ainsi orthographiés.
- Les URL et attributs `alt` des balises images sont également des zones intéressantes pour proposer des versions non accentuées de vos mots-clés importants.
- Parfois, dans certains des textes de votre site, vous pourrez éventuellement insérer une faute d'orthographe de façon volontaire et bien sûr « à l'insu de votre plein gré »... En d'autres termes, insérez sciemment une faute par-ci, par-là, afin que les moteurs les retrouvent. Certes, c'est « moche » et cela a quelques inconvénients : difficile de réellement optimiser une page pour ce mot dans ce cas et surtout, lorsque les internautes liront votre page et s'apercevront qu'elle contient une faute, votre image de marque en

sera altérée. Cette solution n'est donc pas à privilégier même s'il est difficile de l'écarter de façon définitive (elle a quand même le grand avantage d'être la plus simple...).

On peut imaginer bien d'autres possibilités encore. À vous d'être original et de trouver une façon amusante, sérieuse mais surtout **visible** d'indiquer aux moteurs des « mauvaises formes » de vos mots-clés. En effet, un point très important est à rappeler : **n'essayez jamais de cacher ces mots dans vos codes HTML** ! Proposez-les toujours de façon visible pour les internautes ! Il existe bien des façons de cacher du contenu dans une page web. Ne vous y risquez pas ! Ce qui fonctionne éventuellement aujourd'hui sera pénalisé demain par les moteurs et vous risquez le blacklistage à plus ou moins brève échéance. D'autant plus qu'il existe certainement un moyen « honnête » et, qui plus est, amusant, de fournir ces indications en clair. Un webmaster averti en vaut toujours deux...

Intérêt d'un mot-clé

Étudions maintenant les deux critères qui font qu'un mot-clé sera intéressant pour votre référencement : son intérêt et la faisabilité d'un positionnement sur celui-ci dans les meilleurs délais. Le premier point à voir, une fois que vous avez établi une première liste de termes qui vous semblent intéressants, consiste à s'assurer que les mots-clés identifiés ont un intérêt. En d'autres termes, être sûr qu'ils sont souvent saisis sur les moteurs de recherche par les internautes. Voici comment faire.

Le meilleur outil à notre disposition à l'heure actuelle est certainement le générateur de mots-clés de Google (*https://adwords.google.fr/select/KeywordToolExternal*). Cet outil sert au départ pour les campagnes de liens sponsorisés et vous propose de saisir un mot-clé en vous donnant des statistiques à son sujet.

Par exemple, entrez le mot-clé « référencement » (n'hésitez pas à commencer vos recherches par des mots-clés très génériques) et l'outil vous donnera des résultats tels que présentés sur la figure 3-22.

Le générateur de mots-clés de Google vous indique le nombre de fois où ce terme, ou toute expression le contenant, a été saisi sur le moteur de recherche Google ainsi que sur le réseau des portails partenaires de cette société.

- La colonne « Volume de recherche locale » indique le nombre de fois où la requête a été saisie sur le moteur le mois précédent, avec un filtre à la fois linguistique et géographique appliqué et modifiable (par défaut sur la version française de l'outil : Français, France).

- La colonne « Volume de recherche mensuel global » indique le nombre de requêtes sur le moteur, en moyenne mensuelle calculée sur les douze derniers mois, mais sans le filtre géographique/linguistique. Cela explique que, sur certains mots à orthographe identique en français et en anglais, comme « information », les chiffres sont bien plus

importants dans cette deuxième colonne (*http://actu.abondance.com/2009/05/le-generateur-de-mots-cles-de-google-se.html*).

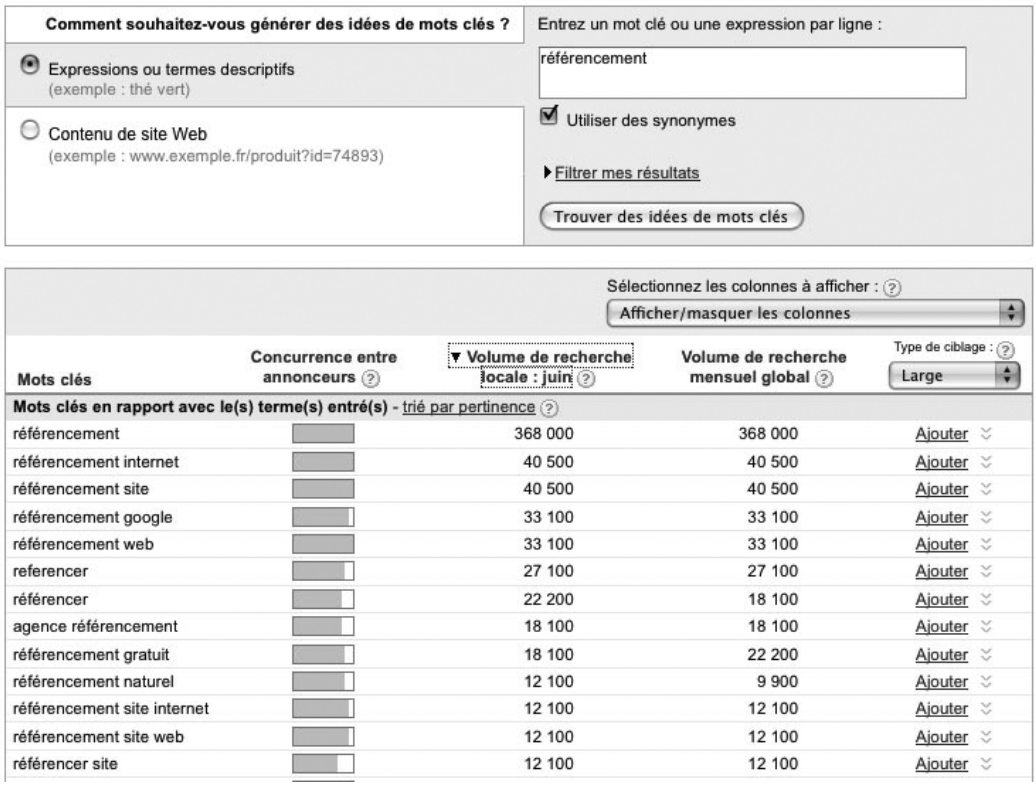


Figure 3-22

Résultats du générateur de mots-clés de Google pour la requête « référencement »

Ces outils sont donc indispensables pour appréhender le potentiel d'un mot-clé. En revanche, il est complexe de dire à partir de combien de requêtes un mot-clé représente un fort potentiel. Tout dépend du domaine dans lequel vous travaillez… Quelques centaines ou milliers de requêtes seront peut-être très intéressantes, voire inestimables, pour votre activité. Évidemment, si un positionnement est possible sur ces termes, plus il y en aura, mieux ce sera.

Ces outils sont plutôt intéressants pour comparer les potentiels de deux termes et savoir lequel est le plus performant. En outre, leur intérêt est également de vous indiquer de façon claire si un terme ou une expression est très peu souvent saisi sur les moteurs… Par exemple, si le résultat est inférieur à 50, réfléchissez bien avant de lancer un positionnement sur cette requête, car le résultat risque fort d'être bien décevant…

Exemple : sur la figure 3-23, la requête « pizzeria arles » n'est pas trouvée par le générateur utilisé sur le mois précédent. Le message « Données insuffisantes » est alors affiché.

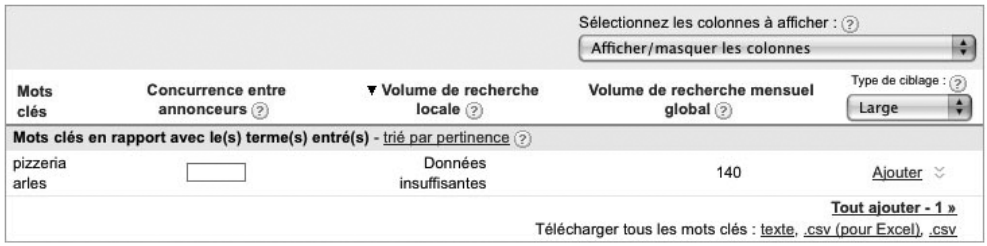


Figure 3-23

Résultat du générateur de mots-clés Google pour la requête « pizzeria arles »

Une fois que vous avez utilisé ces outils, malgré leurs petits défauts, vous devriez avoir les idées plus claires sur le potentiel des termes et expressions que vous désirez prendre en compte pour votre référencement. N'hésitez par à créer un « lexique de mots-clés » que vous aurez sous les yeux, par la suite, lorsque vous aurez à rédiger vos contenus éditoriaux. Vous pourrez ainsi les parsemer de vos termes et expressions importants. Pas négligeable dans une optique de longue traîne...

La faisabilité technique du positionnement

Avoir identifié des mots-clés souvent saisis dans le cadre de votre activité est une première étape essentielle mais cela ne suffit pas. Il faut maintenant vérifier qu'il est techniquement possible de positionner une page de votre site sur ce terme ou cette expression.

Pour ce faire, vous pouvez utiliser Google et taper le mot-clé (ou l'expression) en question dans le formulaire de recherche :

- sur *http://www.google.com/* pour les mots-clés en anglais ;
- sur *http://www.google.fr/* pour les mots-clés en français.

Ensuite, il vous faut regarder le nombre de résultats (ici, plus de 4 millions) retournés par Google (voir figure 3-24).

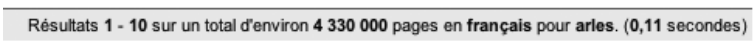


Figure 3-24

Nombre de résultats renvoyés par Google pour le mot-clé « arles »

L'aspect concurrentiel du mot-clé, et donc la faisabilité d'un positionnement sur ce dernier, pourra être fourni par des fourchettes de résultats.

- Jusqu'à 50 voire 100 000 résultats : *a priori* pas de souci à se faire, vous devriez pouvoir bien vous positionner sur ce terme en optimisant de façon professionnelle les pages web de votre site : titre, texte visible, liens, etc. (voir chapitres 4 et 5).
- De 100 000 à 500 000 résultats : la concurrence est plus forte, il sera donc plus complexe de positionner vos pages, mais cela reste possible. Cela prendra peut-être plus de temps et demandera une optimisation plus fine, mais vous avez vos chances…
- Au-delà de 500 000, voire un million de résultats : l'approche est plus aléatoire. Notez bien que rien n'est impossible, mais peu de garanties sont envisageables. Il vous faudra pas mal de travail, beaucoup de temps et un peu de chance pour arriver au Graal des premières positions dans ce cas…

Élaborez vos propres fourchettes

Notez bien que les fourchettes ci-dessus nous ont été dictées par notre expérience. Elles sont donc empiriques. Vous pouvez avoir d'autres idées au sujet de ces nombres, notamment en fonction du domaine d'activité dans lequel vous travaillez.

Bien sûr, il existe un facteur supplémentaire non négligeable : l'agressivité de vos concurrents à ce niveau. Nous en avons déjà parlé précédemment : plus il y aura d'acteurs qui tentent d'atteindre, par l'optimisation de leurs pages, les dix premières places, plus la tâche sera ardue. N'oubliez pas d'en tenir compte, notamment sur le fait que plus le mot-clé sera précis, moins la concurrence sera forte. En d'autres termes, mieux vaut peut-être viser des expressions comme « hôtel sélestat » que « hôtel alsace » voire simplement « hôtel »… De plus, quelqu'un qui saisit le terme « hôtel » vous intéresse-t-il absolument si vous avez un établissement à Sélestat (Bas-Rhin) ? Ne vaut-il pas mieux viser un trafic ciblé (les internautes qui sont intéressés par un hôtel dans votre ville, voire votre région) plutôt qu'un trafic important mais qui risque d'être stérile ? Sans parler de la difficulté d'être bien positionné sur le mot-clé « hôtel » (plus de 30 millions de résultats sur Google avec l'option « Pages francophones » au moment où ces lignes sont écrites), bien sûr…

Comme nous le disions au chapitre 2, imaginez que vous soyez au départ d'une course de fond. Plus il y aura de concurrents, plus il sera difficile de terminer dans les dix premiers. Et plus il y aura de professionnels dans la course, plus la difficulté sera importante… Il en est de même pour votre visibilité sur le Web.

Le référencement prédictif

Il existe de nombreux domaines, dans la vie réelle, qui constituent des événements « prévisibles », appelés parfois les « marronniers » (*http://fr.wikipedia.org/wiki/Marronnier_(journalisme)*)

lorsqu'ils surviennent à date fixe (Saint-Valentin, Noël, fête des Mères, Halloween, etc.). Mais il peut aussi s'agir, par exemple, d'événements sportifs comme une Coupe du monde de football, des championnats de ski, ou encore des manifestations régulièrement organisées comme un festival de musique, un salon automobile ou autres.

Bref, une question se pose souvent au sujet de ces manifestations : si leur date est fixe, à partir de quand faut-il prévoir leur référencement pour être « au top » à la date en question où aura lieu l'événement ? En effet, on voit souvent certains sites web proposer du contenu au sujet d'un événement une ou deux semaines avant celui-ci. Est-ce suffisant ? Ne faut-il pas mettre en ligne du contenu bien avant ? C'est à ces questions que la notion de référencement prédictif va tenter de répondre en observant les courbes temporelles de saisie de mots-clés sur les moteurs de recherche...

Tout d'abord, il est évident que nous allons traiter uniquement ici d'événements prévisibles. Des actualités, comme une tempête, un attentat ou un accident d'avion, ne peuvent bien entendu pas faire l'objet des travaux que nous allons décrire.

Deux outils indispensables

Pour tenter d'y voir plus clair, dans le domaine du « référencement prédictif », nous allons utiliser deux outils qui nous semblent indispensables pour évaluer le délai nécessaire entre le début d'un référencement événementiel et son pic de trafic. Ces deux outils sont :

- Le générateur de mots-clés de Google pour définir l'univers sémantique de l'événement : comment les internautes recherchent-ils l'information au sujet de la manifestation en question sur les moteurs de recherche ? Il s'agit de l'outil que nous avons décrit dans les pages précédentes, disponible à l'adresse suivante : *https://adwords.google.fr/select/KeywordToolExternal.*
- Google Insights for Search pour définir, sur la base des requêtes identifiées dans un premier temps, les tendances temporelles de recherche. Cet outil est disponible à l'adresse suivante : *http://www.google.com/insights/search/#*.

Vous pouvez également utiliser Google Trends (*http://www.google.fr/trends/*) à la place de Google Insights for Search, mais d'après notre expérience, ce dernier nous semble plus complet et mieux adapté à notre quête...

Étape 1 – Définir un univers sémantique simple

Première étape donc, définir un champ sémantique simple (quelques requêtes incontournables) va vous permettre de savoir comment les internautes effectuent des recherches sur les moteurs pour trouver des données sur l'événement en question. Prenons quelques exemples et utilisons le générateur de mots-clés de Google...

- Saint-Valentin : en tapant la simple requête de départ « saint valentin » sur le générateur, on s'aperçoit vite que le champ sémantique principal est limité à quelques requêtes (une fois les résultats triés par « Volume de recherche mensuel global »).

Tableau 3-2 Nombre de saisies mensuelles moyen pour les requêtes autour de la Saint-Valentin

Mots-clés	Nombre de saisies mensuelles en moyenne
saint valentin	90 500
cadeau saint valentin	12 100
carte saint valentin	12 100
cadeaux saint valentin	6 600
cartes saint valentin	5 400
...	...

- Printemps de Bourges : la requête « printemps de bourges » fournit les résultats suivants.

Tableau 3-3 Nombre de saisies mensuelles moyen pour les requêtes autour du Printemps de Bourges

Mots-clés	Nombre de saisies mensuelles en moyenne
printemps de bourges	33 100
printemps de bourges 2008	2 900
...	...

- Salon de l'auto de Genève : testons la requête « salon genève ».

Tableau 3-4 Nombre de saisies mensuelles moyen pour les requêtes autour du Salon de l'auto de Genève

Mots-clés	Nombre de saisies mensuelles en moyenne
salon geneve	27 100
salon genève	8 100
salon auto genève	3 600
salon auto de genève	2 400
...	...

- Coupe du monde de football : simplifions la recherche avec « coupe du monde ». Voici les résultats obtenus.

Tableau 3-5 Nombre de saisies mensuelles moyen pour les requêtes autour de la Coupe du monde de football

Mots-clés	Nombre de saisies mensuelles en moyenne
coupe du monde	246 000
coupe du monde 2006	22 200
coupe du monde 1998	14 800
coupe du monde football	14 800
coupe du monde foot	12 100
...	...

On pourrait multiplier les exemples à l'envi. On s'aperçoit, qu'assez souvent, on peut trouver facilement des « requêtes phares » qui caractérisent rapidement l'événement dont on veut faire la promotion (tout du moins sa prochaine édition).

Une fois ces mots-clés et cet univers sémantique identifiés, nous allons pouvoir passer à la deuxième étape de notre recherche prédictive...

Étape 2 – Effectuer une recherche prédictive

Nous allons utiliser Google Insights for Search (*http://www.google.com/insights/search/#*) pour avoir une idée des « pics de fréquence » des mots-clés identifiés au préalable.

Premier exemple, le plus simple, avec le mot-clé « saint valentin » (avec, pour être plus précis, une recherche uniquement sur la France en utilisant le menu déroulant en haut de page).

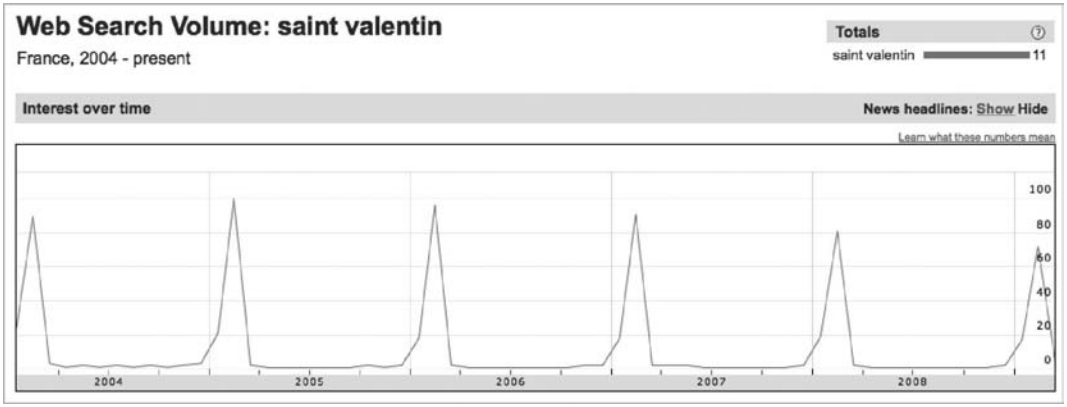


Figure 3-25

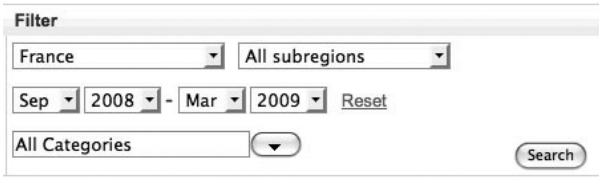
Fréquence temporelle de saisie de la requête « saint valentin » sur Google grâce à l'outil Google Insights for Search

Comme on pouvait s'y attendre, on voit tout de suite un pic chaque mois de février... Prévisible :-)

Approchons-nous maintenant et regardons la courbe (choix « Specific Date Range » du menu déroulant) de septembre 2008 à mars 2009.

Figure 3-26

Choix d'une fourchette de dates spécifique



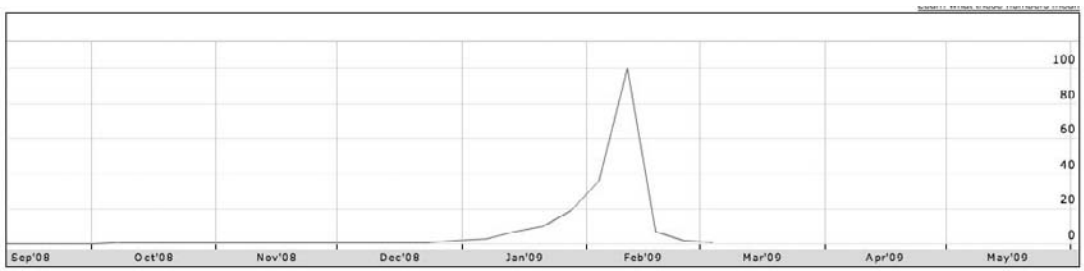


Figure 3-27

Courbe sur neuf mois

Le graphique est clair : si le pic de saisies se trouve, de façon logique, en février, on voit bien que les internautes commencent à chercher des informations au sujet de la Saint-Valentin à partir de début janvier. C'est donc à ce moment-là (et pas obligatoirement avant) qu'il faut être prêt en termes de référencement pour ne pas perdre de trafic. En effet, si vous fournissez un contenu en ligne uniquement début février :

- vous avez perdu tout le trafic généré par les requêtes faites autour de la Saint-Valentin depuis début janvier ;
- les internautes ont commencé à prendre des habitudes sur des sites qui peuvent être vos concurrents et qui s'y sont pris plus tôt. Dommage...

Bref, vous arrivez un peu tard.

Nota : Google Insights for Search propose d'autres informations intéressantes (pays, régions, mots-clés, catégories, etc.). N'hésitez pas à explorer ces possibilités pour éventuellement affiner votre recherche selon l'événement visé.

Étape 3 – Détecter le début du pic des requêtes et commencer à proposer du contenu un mois avant

N'oubliez pas que si vous désirez être prêt en termes de référencement début janvier, cela signifie que vous aurez dû commencer à mettre en ligne du contenu avant, et attendre que les moteurs l'aient « digéré » pour que vous puissiez être bien positionné à ce moment-là. Les moteurs de recherche ayant fait de grands progrès à ce niveau, on peut estimer que si vous proposez du contenu optimisé début décembre, vous aurez certainement de bons résultats en positionnement début janvier.

Cela nous mène quand même deux mois et demi avant la date fatidique du 14 février ! Autant y penser avant...

Bien sûr, il n'est pas nécessaire de proposer dès le départ un contenu très important. Pour reprendre notre exemple sur la Saint-Valentin, vous pouvez tout à fait mettre en place la procédure suivante :

- Début décembre, mise en ligne d'un site dédié (par exemple, à l'adresse *saint-valentin.votre-site.com*) avec un contenu de départ léger et quelques articles pour « amorcer la pompe »...
- Courant décembre, ajout de quelques articles au fur et à mesure (par exemple, un article tous les deux ou trois jours) pour faire vivre le site et montrer aux moteurs qu'il évolue.
- Proposer par ailleurs une page d'accueil mise à jour quotidiennement pour habituer les spiders à venir souvent la visiter.
- En janvier, augmenter la cadence de publication avec un article nouveau par jour.
- À partir de début février, boostez le site avec plusieurs articles nouveaux chaque jour.

Le site a ainsi une cadence d'évolution logique, assez « normale » au départ pour terminer en trombe. Cela devrait plaire aux moteurs. Oui, certes, c'est du travail... Mais on n'a rien sans rien !

Notez bien que le délai à partir duquel il faut proposer du contenu en ligne peut varier en fonction des événements. Par exemple, pour la requête « printemps de bourges », Google Insights for Search donne la courbe représentée à la figure suivante.

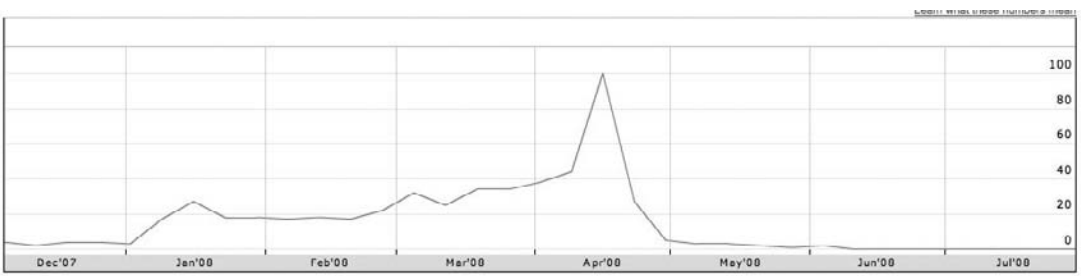


Figure 3-28

Fréquence temporelle de saisie de la requête « printemps de bourges » sur Google grâce à l'outil Google Insights for Search

On voit bien ici que, si la manifestation a lieu en avril (voir le pic évident à cette période), les requêtes commencent « à se réveiller » dès le mois de janvier, soit plus de trois mois avant... C'est normal car dès cette date, on connaît les groupes qui vont jouer lors du festival, et plusieurs annonces suscitent des recherches par les internautes. Autant le prévoir dès le mois de décembre...

En revanche, une compétition comme le Paris-Dakar est surtout recherchée (requête « paris dakar ») sur une période très courte.

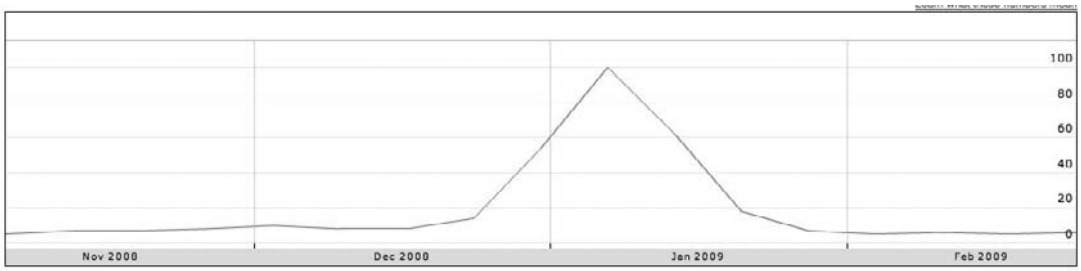


Figure 3-29

Fréquence temporelle de saisie de la requête « paris dakar » sur Google grâce à l'outil Google Insights for Search

Le pic des recherches, ici, commence dans la dernière semaine de décembre pour se terminer la troisième de janvier. Un délai très bref dans ce cas...

Tout dépendra en fait de la manifestation en question. Il vous faut donc, par ordre chronologique :

1. Identifier quelques « requêtes phares », emblématiques de la façon dont les internautes recherchent l'information sur l'événement, grâce au générateur de mots-clés de Google.
2. Chercher dans l'outil Google Insights for Search où se situe le pic de recherches et quand commencent les investigations de la part des internautes.
3. Commencer, *grosso modo*, à proposer du contenu en ligne sur un site ou une rubrique dédié un mois avant cette date de début des requêtes...

Il est en tout cas très clair que trop attendre n'est pas une bonne chose en termes de référencement prédictif. Mais il n'est pas non plus nécessaire d'être prêt trop tôt. Bref, il faut être *just in time*... Si l'avenir appartient à ceux qui se lèvent tôt, une bonne visibilité sur les moteurs appartient à ceux qui proposent du contenu tôt...

Méthodologie de choix des mots-clés

Pour vous aider dans le choix de vos mots-clés, nous vous proposons ci-dessous une petite méthodologie que vous pourrez sans soucis adapter à vos besoins.

1. Dans un premier temps, faites une liste intuitive d'une dizaine de mots-clés qui caractérisent votre activité. Imaginons que vous soyez une société de référencement. Les termes qui vous viendront tout de suite à l'esprit sont les suivants.

Tableau 3-6 Liste de mots-clés de départ pour une société de référencement

Référencement
Positionnement
Moteurs de recherche
Annuaire
Visibilité
Liens sponsorisés
Referencement
Visibilite
Lien sponsorise
Moteur de recherche

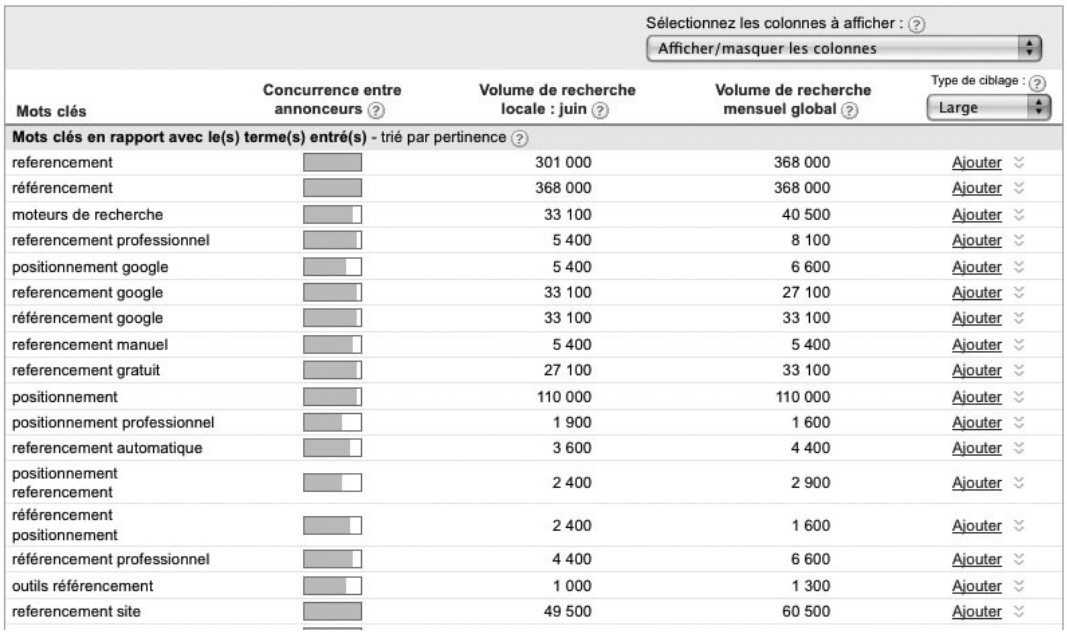
Notez ici les différentes versions d'un même mot (singulier/pluriel, accents, etc.).

2. Comme le montre la figure 3-30, allez sur le générateur de mots-clés de Google (*https://adwords.google.com/select/KeywordToolExternal*) et tapez cette liste dans la zone « Entrez un mot-clé ou une expression par ligne : » tout en cochant l'option « Utiliser des synonymes ».

Figure 3-30

Première saisie sur le générateur de mots-clés de Google

3. Étudiez la liste des mots-clés proposés par l'outil (voir figure 3-31). Elle est exhaustive et comprend certainement des termes et expressions auxquels vous n'aviez pas pensé auparavant. Lister par « Volume de recherche » croissant pour obtenir en premier les requêtes qui sont le plus souvent saisies.



Sélectionnez les colonnes à afficher : ?
Afficher/masquer les colonnes

Mots clés	Concurrence entre annonceurs ?	Volume de recherche locale : juin ?	Volume de recherche mensuel global ?	Type de ciblage : ? Large
Mots clés en rapport avec le(s) terme(s) entré(s) - trié par pertinence ?				
referencement		301 000	368 000	Ajouter
référencement		368 000	368 000	Ajouter
moteurs de recherche		33 100	40 500	Ajouter
referencement professionnel		5 400	8 100	Ajouter
positionnement google		5 400	6 600	Ajouter
referencement google		33 100	27 100	Ajouter
référencement google		33 100	33 100	Ajouter
referencement manuel		5 400	5 400	Ajouter
referencement gratuit		27 100	33 100	Ajouter
positionnement		110 000	110 000	Ajouter
positionnement professionnel		1 900	1 600	Ajouter
referencement automatique		3 600	4 400	Ajouter
positionnement referencement		2 400	2 900	Ajouter
référencement positionnement		2 400	1 600	Ajouter
référencement professionnel		4 400	6 600	Ajouter
outils référencement		1 000	1 300	Ajouter
referencement site		49 500	60 500	Ajouter

Figure 3-31

Listing des requêtes les plus intéressantes sur Google

Rien ne vous empêche de faire cette opération plusieurs fois : vous récupérez dans la première liste des termes intéressants que vous ajoutez ensuite à votre liste initiale de 10 mots-clés et vous relancez une recherche, etc.

4. Copiez-collez la liste fournie (plusieurs dizaines voire centaines de termes) dans un tableur et supprimez les expressions qui ne vous intéressent pas. Ne gardez que celles qui vous semblent convenir le mieux à votre activité (la requête « annuaire allemand », par exemple, n'est pas obligatoirement votre tasse de thé…). À ce stade de la méthodologie, vous avez par exemple en main une centaine de requêtes intéressantes, car elles ont trait à votre activité et elles sont souvent demandées sur Google.

5. Pour chacune de ces requêtes, regardez combien de fois elles sont demandées sur le générateur de mots-clés de Google et combien de résultats sont renvoyés par le moteur de recherche Google.fr (*http://www.google.fr/*) lorsqu'on tape ces mots-clés. Complétez le tableau.

Tableau 3-7 Pour chaque mot-clé, on liste les résultats du générateur de mots-clés (intérêt) et du moteur de recherche Google (faisabilité).

Mot-clé – requête	Résultats du générateur de mots-clés de Google	Résultats du moteur de recherche Google
Annuaire	259 500	103 000 000
Referencement	301 000	27 000 000
Moteurs de recherche	59 000	2 200 000
Positionnement	110 000	15 400 000
Positionnement Web	50 000	2 260 000
Referencement gratuit	18 100	4 810 000
Visibilité	18 100	8 950 000
Positionnement marketing	10 000	1 950 000
Référencement	368 000	27 100 000

6. Pour l'intérêt et la faisabilité, établissez des fourchettes de notes de 0 à 20 en fonction des résultats trouvés. Exemples :

- Intérêt (générateur de mots-clés)
  - moins de 1 000 résultats : 0 point
  - 1 001 à 10 000 résultats : 5 points
  - 10 001 à 50 000 résultats : 10 points
  - 50 001 à 100 000 résultats : 15 points
  - plus de 100 000 résultats : 20 points
- Faisabilité (moteur de recherche)
  - plus de 100 millions de résultats : 0 point
  - de 50 à 100 millions de résultats : 5 points
  - de 10 à 50 millions de résultats : 10 points
  - de 1 à 10 millions de résultats : 15 points
  - moins de 1 million de résultats : 20 points

Reportez ensuite dans votre tableau les notes ainsi attribuées.

Tableau 3-8 Chaque requête reçoit une note d'intérêt et une note de faisabilité.

Mot-clé – requête	Résultats du générateur de mots-clés de Google	Note d'intérêt	Résultats du moteur de recherche Google	Note de faisabilité
Annuaire	259 500	20	103 000 000	0
Referencement	301 000	20	27 000 000	10
Moteurs de recherche	59 000	15	2 200 000	15
Positionnement	110 000	20	15 400 000	10
Positionnement Web	50 000	10	2 260 000	15
Referencement gratuit	18 100	10	4 810 000	15
Visibilité	18 100	10	8 950 000	15
Positionnement marketing	10 000	5	1 950 000	15
Référencement	368 000	20	27 100 000	10

7. Faites la somme des deux notes pour obtenir une note globale :

Tableau 3-9 Une note globale indique quels mots-clés traiter en priorité.

Mot-clé – requête	Résultats du générateur de mots-clés de Google	Note d'intérêt	Résultats du moteur de recherche Google	Note de faisabilité	Note globale (somme intérêt + faisabilité)
Annuaire	259 500	20	103 000 000	0	20
Referencement	301 000	20	27 000 000	10	30
Moteurs de recherche	59 000	15	2 200 000	15	30
Positionnement	110 000	20	15 400 000	10	30
Positionnement Web	50 000	10	2 260 000	15	25
Referencement gratuit	18 100	10	4 810 000	15	25
Visibilité	18 100	10	8 950 000	15	25
Positionnement marketing	10 000	5	1 950 000	15	20
Référencement	368 000	20	27 100 000	10	30

Cette dernière colonne vous donnera ainsi une priorité dans les requêtes à utiliser pour mieux référencer votre site.

Notez bien que vous pouvez adapter cette méthodologie à vos besoins et attentes, et gérer différemment les fourchettes de notes proposées précédemment. Enfin, comme nous l'avons indiqué plusieurs fois auparavant, le choix de vos mots-clés dépendra également de façon importante de l'aspect concurrentiel de ces derniers. Plus il y aura

de webmasters tentant de se référencer sur vos expressions favorites, plus la faisabilité sera aléatoire en termes de délais ! Vous pouvez éventuellement tenir compte de ce paramètre en ajoutant un coefficient spécifique à vos calculs.

Un arbitrage entre intérêt et faisabilité

Bien choisir vos mots-clés pour un référencement consiste donc à trouver un arbitrage entre le potentiel des termes choisis et la faisabilité technique d'un positionnement sur ceux-ci.

N'hésitez pas à y passer le temps nécessaire car cette étape est absolument capitale dans le déroulement de votre référencement. Si vous n'y prêtez pas l'attention nécessaire, vous pourriez avoir de grosses désillusions par la suite… Le tout n'est pas d'être premier sur un mot ou une expression : il faut aussi qu'il ramène du trafic. Et du trafic qualifié, c'est encore mieux !

Sur quels moteurs et annuaires faut-il se référencer ?

Dans les paragraphes qui précèdent, nous avons vu comment fonctionnent les moteurs et annuaires. Mais savez-vous sur quels outils vous allez devoir être référencé et positionné ? Cette donnée est également importante, car il ne sera pas question de perdre du temps à tenter d'apparaître de façon optimale sur un annuaire ou un moteur qui ne ramène aucun trafic. Voyons dans un premier temps ce qu'il en est pour les moteurs de recherche.

Sur quels moteurs de recherche se positionner ?

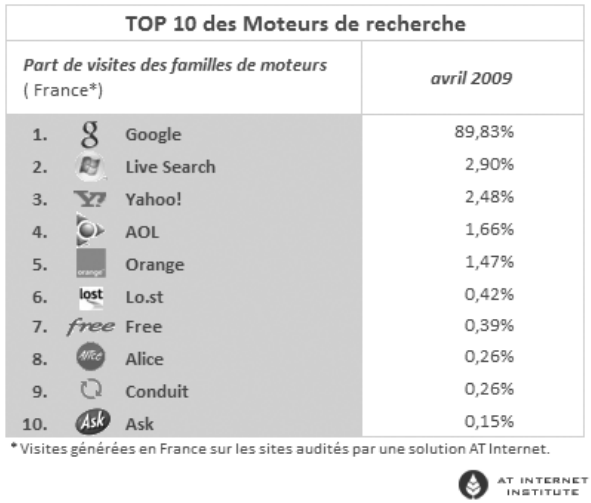
La réponse à cette question est simple : vous devez opter pour ceux qui généreront le plus de trafic sur votre site web. Et ils ne sont pas nombreux… Si l'on en croit les baromètres du référencement disponibles en France (voir figure 3-32), et notamment celui de AT Internet (*http://barometre.secrets2moteurs.com/*), plus de 99 % du trafic est généré par moins de dix outils de recherche : Google (près de 90 % du trafic), Live Search/Bing et Yahoo! (entre 2 et 3 % chacun à la mi-2009), suivis de AOL (qui utilise la technologie Google pour son moteur) et Voila/Wanadoo/Orange. Les autres ne dépassent pas 0,5 % du trafic total.

Et si l'on tient compte du fait que de nombreux moteurs et portails utilisent la technologie de recherche de Google, voire celle de Microsoft, le nombre de technologies de recherche sur lesquelles il va vous falloir être présent est encore plus restreint :

- Google (Google, Neuf/Cegetel, Free, AOL.fr, Alice…)
- Yahoo! (Yahoo!, AltaVista, AlltheWeb, voir ci-après)
- Exalead (Exalead)
- Voila (Voila, Wanadoo, Orange)
- Microsoft Bing (MSN.fr, Bing.com)

Figure 3-32

Baromètre Première Position-Xiti sur les parts de trafic des différents moteurs du paysage franco-phone



TOP 10 des Moteurs de recherche

Part de visites des familles de moteurs (France*)	avril 2009
1. Google	89,83%
2. Live Search	2,90%
3. Yahoo!	2,48%
4. AOL	1,66%
5. Orange	1,47%
6. Lo.st	0,42%
7. Free	0,39%
8. Alice	0,26%
9. Conduit	0,26%
10. Ask	0,15%

* Visites générées en France sur les sites audités par une solution AT Internet.

AT INTERNET INSTITUTE

Soit cinq moteurs seulement, voire quatre puisque suite à un accord signé en juillet 2009 avec Microsoft (*http://actu.abondance.com/2009/07/laccord-entre-yahoo-et-microsoft-enfin.html*), Yahoo! devrait progressivement abandonner sa technologie de recherche d'information au profit de Bing, celle de Microsoft. Il sera donc intéressant de surveiller comment s'effectue cette « passe d'armes » au fil des mois... Gageons qu'au vu de la faible présence, en termes de parts de trafic, de ces deux moteurs en France, la situation évolue peu en 2010 à ce niveau. Cela ne sera peut-être pas le cas aux États-Unis...

Les baromètres du référencement

La France est le pays du monde qui dispose du plus grand nombre de baromètres du référencement, tentant de suivre au mieux les parts de trafic des différents outils de recherche. Voici quelques adresses :

– XiTi/1ère position – *http://barometre.secrets2moteurs.com/*

– adoc – *http://www.barometre.adoc.fr/*

– Wysistat – *http://www.wysistat.net/panorama/*

Cette situation de quasi-monopole de la part de Google est pratiquement identique dans tous les pays d'Europe, les moteurs ou portails typiquement « franco-français » comme Voila, Free ou Exalead étant remplacés par des acteurs locaux comme Search.ch en Suisse ou Yell.com en Grande-Bretagne.

Aux États-Unis, la situation est en revanche légèrement différente avec une hégémonie affirmée mais moins importante de Google. Pour le mois de mai 2009, par exemple, selon le classement de comScore (voir URL ci-après), c'est Google qui s'octroyait la première place mais avec « seulement » 65 % du trafic, devant Yahoo! (20,1 %), MSN/Bing (8,0 %), Ask (3,9 %) et AOL (3,1 %)…

Selon Hitwise, pour avril 2009, le tiercé gagnant était :

1. Google (72,74 %)
2. Yahoo! Search (16,27 %)
3. MSN/Bing (5,68 %)

Enfin, en mai 2009, selon Nielsen Netratings, on obtenait le classement suivant :

1. Google (63,2 %)
2. Yahoo! (17,2 %)
3. MSN/Bing (9,4 %)

Là encore, le *deal* entre Microsoft et Yahoo! pourrait rapidement changer la donne...

Les baromètres anglophones

Les sites ci-dessous publient régulièrement des chiffres sur les parts de marché des outils de recherche dans le monde anglophone :

– OneStat – *http://www.onestat.com/*
– Keynote – *http://www.keynote.com/*
– ComScore – *http://www.comscore.com/*
– Hitwise – *http://www.hitwise.com/*
– Nielsen Netratings – *http://www.nielsen-netratings.com/*

La situation semble donc claire à ce niveau-là.

- Seule une petite dizaine de portails de recherche génèrent du trafic sur les sites web. Et il y a encore moins de technologies.
- Il n'est pas complètement vain de restreindre sa stratégie de référencement au seul Google, qui représente en France près de 90 % du trafic « outils de recherche »… Cette stratégie serait en revanche moins valable pour un site web visant le marché américain, notamment depuis l'annonce de l'accord entre Microsoft et Yahoo!.
- Mais ne prendre en compte que Google, c'est aussi négliger 10 % du trafic des moteurs francophone, ce qui peut être dommage. À vous de faire vos choix…
- Le monde des moteurs de recherche évolue vite et de nouveaux acteurs viennent souvent tenter leur chance sur ce marché. À vous donc de vous tenir au courant, d'effectuer une veille pour sentir les tendances et prendre en compte de nouveaux outils dès qu'ils donnent des signes de vitalité intéressants. Globalement, l'un de ces signes sera une arrivée dans le « Top 10 » des baromètres évoqués dans ce chapitre.

Google n'est pas seul

Si les conseils fournis dans ce livre sont valables en grande partie pour Google, comme les exemples le montreront, ils restent pertinents pour les autres moteurs de recherche, les algorithmes de pertinence utilisés actuellement par les différents concurrents étant très proches.

Sur quels annuaires se référencer ?

Pendant de nombreuses années, le référencement d'un site sur un ou plusieurs annuaires a été partie intégrante d'une stratégie de référencement. Il y a de cela 10 ans, le trafic « outils de recherche » était d'ailleurs divisé en deux sur un site web : la moitié venait des annuaires (et en grande partie du « Guide Web » de Yahoo!) et l'autre partie des moteurs (en France, c'était AltaVista qui se taillait la part du lion du trafic à cette époque-là).

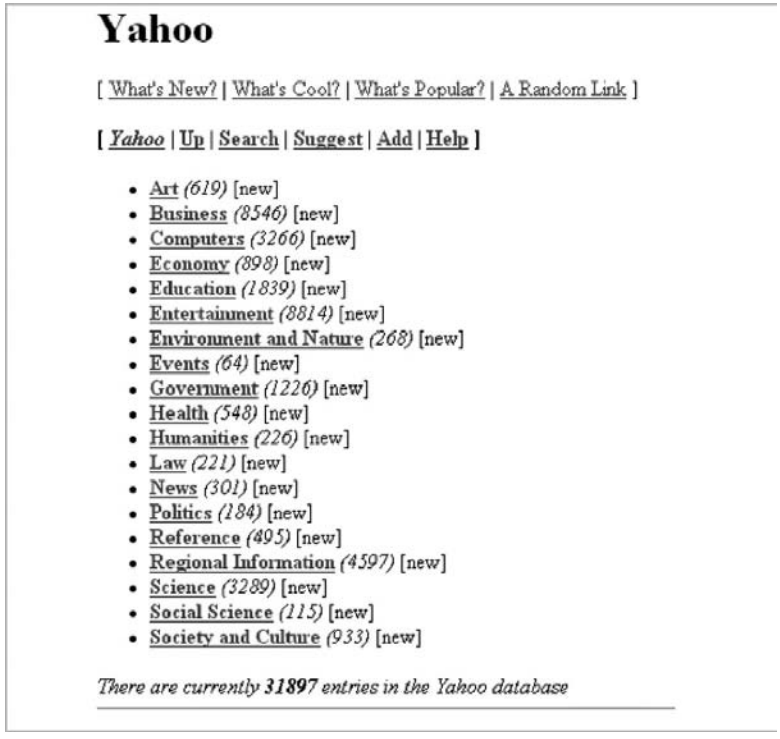
Une stratégie de référencement d'un site web sur les annuaires était donc tout à fait logique et efficace (d'ailleurs, le site Abondance s'appelle ainsi car il commence par les lettres « ab », ce qui le classait toujours en début de liste alphabétique sur ces outils). Mais qu'en est-il aujourd'hui, à une période où Google *truste* plus de 90 % du trafic « outils de recherche » en France ?

Topologie des annuaires

Pour débuter ce chapitre, il nous a semblé important de définir quels types d'annuaires existent aujourd'hui car la topologie de ces outils a grandement été modifiée depuis leur apparition en 1994 et la création par David Filo et Jerry Yang de l'un des premiers annuaires, si ce n'est le premier, sous le nom de Yahoo!.

Figure 3-33

L'une des premières pages d'accueil de Yahoo! en 1994. 31 897 sites référencés...



On peut tout à fait, en 2009, définir plusieurs types d'annuaires :

- **Les généralistes reconnus.** Dans cette catégorie, on trouvera le Guide Web de Yahoo! dans ses versions américaine (*http://dir.yahoo.com/*) et française (*http://fr.dir.yahoo.com/*) ou l'Open Directory ou Dmoz (*http://www.dmoz.org/*). La plupart des autres annuaires de ce type, comme le Guide de Voila, Nomade, Zeal ou autres, ont aujourd'hui disparu ou ne sont plus maintenus. Mais il faut bien dire que les annuaires de Yahoo! ne sont pas en très bon état non plus, très rarement (voire quasi jamais pour la mouture française) mis à jour... Seul l'Open Directory semble surnager encore et être maintenu par ses bénévoles (qui ont bien du mérite...).

Figure 3-34

Dmoz ou Open Directory, l'un des derniers « dinosaures » du domaine...

- **Les généralistes peu connus.** Il s'agit de tentatives de création d'annuaires généralistes qui n'ont pas connu le « succès », tout du moins historiquement parlant, d'un Guide Web de Yahoo! ou d'un Dmoz. Ils sont devenus de plus en plus rares quand ils n'ont pas disparu purement et simplement...
- **Les spécialistes.** Il s'agit d'annuaires spécialisés dans une thématique donnée (le sport, le e-commerce, la politique, l'écologie, etc.). Ils ne s'orientent que vers un domaine précis mais tentent de recenser de la façon la plus exhaustive possible tous les sites les plus utiles et les plus pertinents dans cet univers.

- Les annuaires créés pour le référencement. On trouve dans cette catégorie une multitude (plusieurs centaines, voire plusieurs milliers) d'annuaires pour la plupart totalement inconnus mais qui ont été créés à des fins de référencement, pour « créer du lien » vers les sites qu'ils recensent (et éventuellement toucher de l'argent émanant de l'offre de liens contextuels AdSense de Google).

Cette typologie mise en place, il est temps de passer à l'étape suivante.

Quels avantages les annuaires procurent-ils ?

Tout d'abord, il est important de poser la question de l'utilité de ce type de référencement sur les annuaires. Pourquoi mettre en place une politique d'inscription sur quelques-uns, ou plusieurs dizaines, voire centaines d'annuaires ? Cette motivation peut prendre plusieurs formes.

Motivation numéro 1 : le trafic créé par les clics sur les liens proposés par les annuaires vers les sites web qu'ils référencent. Cela peut s'avérer exact pour certains d'entre eux qui vont créer du trafic, parfois non négligeable, vers les sites indexés. Il sera ici important d'utiliser des outils comme Google Trends for Websites (*http://trends.google.com/websites?q=wikipedia.org*) pour s'assurer que l'annuaire sur lequel vous désirez vous inscrire est souvent consulté par les internautes. Sinon, il n'en vaudra pas vraiment la peine, pour l'argument du trafic direct tout du moins. Sachez cependant que, à l'analyse de nombreux sites web, le trafic direct fourni par les annuaires est quasi nul. Ceci dit, les outils d'analyse d'audience ne sont, pour l'immense majorité d'entre eux, pas configurés pour intégrer les milliers d'annuaires du Web dans leurs chiffres de trafic « outils de recherche » (ils apparaissent alors dans les « sites référents » *lambda*), ce qui fausse obligatoirement l'analyse et les rend complexes à identifier...

Par expérience, on peut dire que, dans l'ensemble, l'inscription dans de nombreux annuaires n'augmente pas (sauf exception rare) de façon notable le trafic sur un site, surtout sur un site professionnel. Il existera toujours de nombreux gestionnaires d'annuaires pour vous dire le contraire mais pour l'instant, l'expérience nous pousse à dire que la génération de trafic direct ne peut pas être une motivation suffisante pour effectuer ce type de travail.

Motivation numéro 2 : un meilleur positionnement dans les moteurs de recherche. La rumeur a longtemps parcouru le Web : être dans le Guide Web de Yahoo! améliore le positionnement du site dans les pages du moteur de recherche de Yahoo!. Autre rumeur : être dans l'Open Directory signifie que l'on sera dans l'annuaire de Google (*http://directory.google.fr/*) et donc qu'on sera mieux positionné sur le moteur de recherche Google ou en tous les cas plus vite indexé.

Disons-le tout net : si la question pour Yahoo! n'est pas totalement tranchée aujourd'hui (mais l'avantage doit être très faible si avantage il y a, et de toute façon, l'annuaire de Yahoo! n'est plus maintenu...), il est clairement inexistant pour Google.

Figure 3-35

L'annuaire de Google, basé sur l'Open Directory

Une bonne gestion de liens fera entrer un nouveau site en quelques heures dans l'index de Google. Ce n'est donc pas la peine de passer par une phase d'inscription dans Dmoz pour être « rapidement » indexé par Google, d'autant plus qu'il est de notoriété publique qu'une inscription dans Dmoz est loin d'être rapide (bénévolat des éditeurs qui le maintiennent oblige)... Là encore, on peut mettre de côté, de façon claire et nette, cet argument.

De plus, Google, par exemple, tout comme Yahoo! avec son annuaire, utilise parfois le descriptif de l'Open Directory dans les *snippets* (résumés textuels) de ses pages de résultats. Mais ceux-ci étant souvent assez courts, voire pas toujours actualisés, on préfère la plupart du temps utiliser la balise meta `NOODP` (voir chapitre 9) pour le lui interdire. On ne peut donc pas dire ici que l'indexation dans Dmoz soit un réel avantage...

Motivation numéro 3 : la popularité d'un site. Ici, on motive l'inscription d'un site dans un annuaire par la création de liens vers la source d'informations, et les liens étant importants pour les moteurs de recherche (notions de popularité et de réputation), cela renforce leur pertinence. En effet, ceci est important, et une inscription dans les annuaires peut avoir un effet bénéfique sur vos liens donc sur votre popularité (PageRank), mais à plusieurs conditions cependant :

- Il faut que l'annuaire soit connu des moteurs de recherche et que ces derniers aient indexé de nombreuses pages de l'outil. Exemple : si un lien vers votre site se trouve

dans la rubrique « Top>Commerce et économie>Assurance>Assurance-vie », il faudra que cette page (par exemple, à l'adresse *www.annuaire-web.com/commerce-economie/assurance/assurance-vie.html*) soit présente dans l'index des principaux moteurs. Sinon ce lien ne servira pas votre référencement.

Notre conseil : utiliser la syntaxe « site: » suivie de l'adresse de l'annuaire en question (par exemple, « site:www.annuaire-web.com »). Elle fonctionne sur la totalité des moteurs majeurs actuels et vous donnera la liste des pages indexées pour ce site. À vous de voir si ce volume vous satisfait.

Encore mieux : pour évaluer l'indexation sur Google, tapez cette requête sur un moteur partenaire de Google comme AOL (*http://recherche.aol.fr/*). Comme ce moteur ne fonctionne que sur l'index principal de Google (voir chapitre 2), ses résultats seront plus proches de la « réalité » en termes d'intérêt en vous dévoilant uniquement les liens émanant des pages réellement importantes pour Google... Si vous vous apercevez que toutes les pages de l'annuaire sont dans l'index secondaire de Google, cela peut ne pas vous apporter grand-chose en termes de popularité.

- Il faut que la page de l'annuaire qui contiendra le lien vers votre site soit un minimum populaire (voir chapitre 5 pour la notion de popularité). Si vous obtenez un lien depuis une page de PageRank 1 ou 2, l'impact sur votre site sera quasi nul (même si vous obtenez de nombreuses pages de ce type). Si le lien est sur une page de PageRank 3 ou 4, voire plus, cela peut devenir intéressant. Notez cependant ici qu'un lien n'étant jamais pénalisant, il ne vous fera pas de mal s'il ne vous fait pas de bien...

  Notre conseil : vérifiez, grâce à la barre d'outils de Google, le PageRank de plusieurs pages internes de l'annuaire, en descendant l'arborescence. Si la majeure partie des rubriques ont un PageRank inférieur à 3, l'annuaire n'a que peu d'intérêt.

- Le lien doit être « en dur », c'est-à-dire qu'il ne doit pas passer par un système tiers (affiliation, système de calcul des clics, etc.).

  Notre conseil : passez la souris sur les liens des sites déjà présents dans les pages de l'annuaire et regardez, en bas à gauche de votre navigateur, vers quelle adresse ils pointent. Si le lien est « en dur » (l'adresse directe du site apparaît), c'est bon. Si une adresse du type *http://fr.srd.yahoo.com/S=2100062373:D1/CS=2100062373/SS=2100797847/SIG=1147tbtdn/*http%3A//www.urma-montpellier.org/* pour aller à l'adresse *http://www.urma-montpellier.org/* (comme sur le Guide Web de Yahoo!, voir figure 3-36) est affichée, votre site ne profitera pas du lien puisque ce dernier n'est pas direct. Aucun intérêt dans ce cas.

- Attention : il peut arriver que l'annuaire demande en retour un lien de l'inscription de votre site. Les moteurs de recherche n'apprécient que modérément les échanges de liens « en direct » (« Je pointe vers toi, tu pointes vers moi »). Les échanges de liens vraiment efficaces se font plutôt en triangle (A pointe vers B qui pointe vers C qui pointe vers A). Voir chapitre 5 pour ce point particulier.

  Notre conseil : évitez les inscriptions sur les annuaires qui demandent de façon obligatoire un lien en retour.

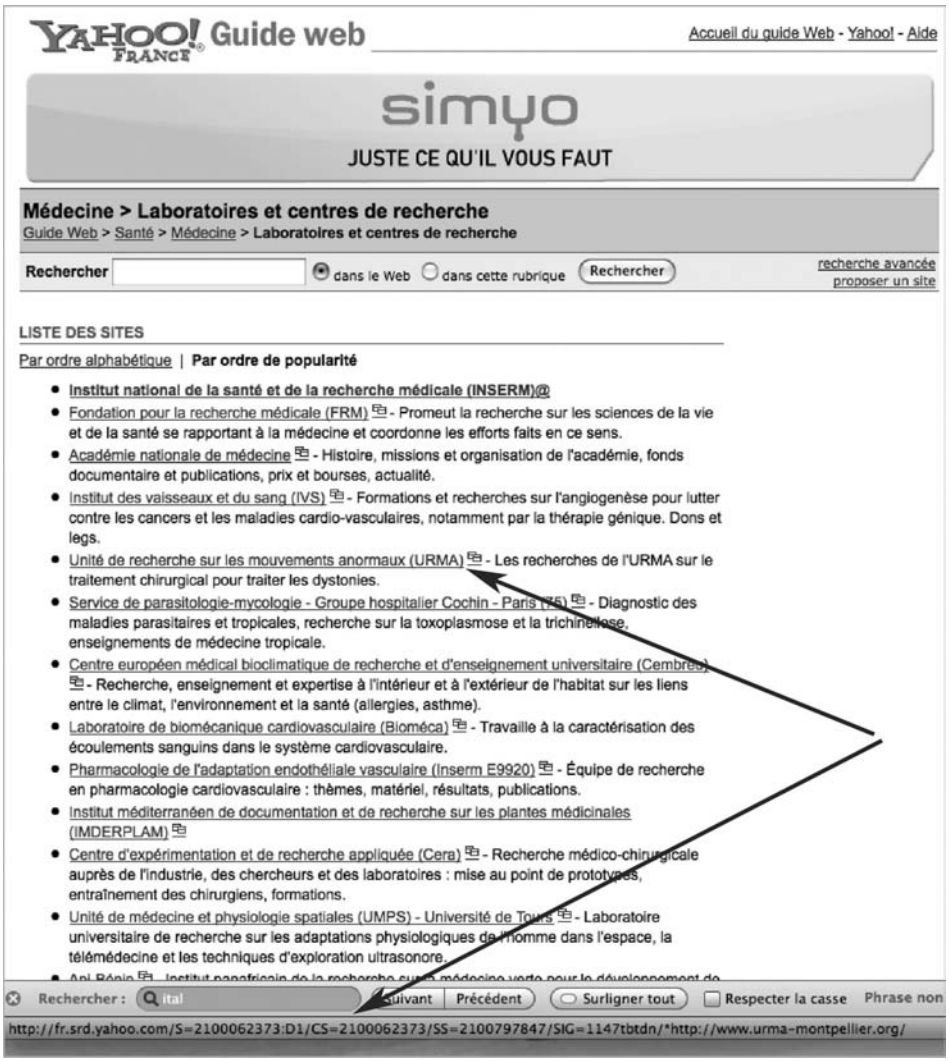


Figure 3-36

En passant la souris sur un lien de l'annuaire, l'adresse du site en direct doit apparaître dans la barre d'état, contrairement à l'exemple ci-dessus.

- N'oubliez pas qu'un annuaire établira, dans ses catégories, un lien vers la page d'accueil de votre site. Seule cette page va donc bénéficier du transfert de popularité offert par ce lien (ce qui est déjà pas mal...). Mais cela ne suffit pas. Dans une bonne stratégie de netlinking, le deeplinking (mise en place de liens vers les pages internes de votre site) est essentiel et dans ce cas, les annuaires ne vous aideront pas. Nous y reviendrons dans le cadre de cet ouvrage...

Notre conseil : les liens émanant d'annuaires servent à renforcer la popularité de la page d'accueil de votre site. Pour vos pages internes, il faudra passer par d'autres voies comme les fils RSS, le *linkbaiting*, etc.

- De la même façon, la transmission de popularité (PageRank) au travers des liens dépend du nombre de liens sortants de la page qui pointe vers vous. Si la catégorie de l'annuaire qui contient votre lien propose des dizaines de liens, voire plus, vous n'en recevrez que des miettes. Peu intéressant...

  Notre conseil : si cela est possible, choisissez l'inscription de votre site dans une catégorie de l'annuaire qui ne contient que peu de sites.

- Enfin, Google a communiqué en 2009 sur le fait qu'il n'appréciait que modérément les annuaires créés spécifiquement pour des besoins de référencement, qu'il assimile a certaines « fermes de liens », et il y a de fortes chances qu'il déprécie, dans son algorithme, les liens qui en sont issus... Autant le savoir.

Les outils de soumission automatique

Il existe plusieurs outils permettant de soumettre de façon automatique votre site à des centaines, voire parfois des milliers, d'annuaires. En voici quatre, classés par ordre alphabétique :

- Linkomatic – *http://www.linkomatic.org/*
- SubmitWolf – *http://www.trellian.fr/swolf/*
- Website Submitter – *http://www.submitsuite.com/products/website-submitter.htm*
- Yooda Submit – *http://www.yooda.com/outils_referencement/submit_center_yooda/*

Bien sûr, dans ce cas, vous n'allez pas vérifier, comme nous vous le suggérons dans ce chapitre, bon nombre de critères (PageRank des catégories, nombre de pages indexées par les moteurs, etc.) pour chacun des annuaires pris en compte par ces logiciels. Vous préparez vos campagnes de soumission et vous lancez l'opération qui s'effectue ensuite de façon automatique, voire semi-automatique.

Est-ce efficace ? Bonne question. En tout cas, cela ne risque pas de faire du mal à votre site (en revanche, utilisez pour vos soumissions une adresse e-mail temporaire si vous ne voulez pas vous voir spammé dans les minutes qui suivent...). Et cela peut même apporter un nombre de liens conséquent à votre page d'accueil si le travail est bien fait. Donc un bénéfice certain. Mais n'oubliez pas non plus que Google horodate toutes les informations qu'il trouve et en tient compte dans ses analyses. Il ne sera donc pas obligatoirement de bon aloi que ce moteur détecte, dans un intervalle de quelques heures ou quelques jours, une forte augmentation de liens vers votre site. Si vous le pouvez, planifiez vos inscriptions sur une durée plus longue !

Bien sûr, tout cela est loin d'être suffisant dans une stratégie de référencement et il faudra de toute façon que vos pages soient optimisées et que vos pages internes reçoivent elles aussi des liens (*deeplinking*), etc. Mais, pour un site qui vient d'être mis en ligne et qui

veut rapidement voir sa popularité croître un minimum, le jeu peut en valoir la chandelle si le travail est fait de façon sérieuse...

Il ne s'agit cependant pas là d'une stratégie essentielle et indispensable pour le bon référencement d'un site mais cela peut apporter un « bonus » qui peut être parfois intéressant. Pratique à tenter donc si vous avez un peu de temps devant vous... Mais n'en attendez pas des miracles...

Et les autres outils de recherche ?

Nous avons évoqué jusqu'ici le référencement de votre site sur les principaux moteurs et annuaires généralistes ayant pour vocation de traiter tout le Web. Cela veut-il dire que les autres outils de recherche (ceux qui ne sont pas considérés comme majeurs) sont négligeables et qu'il ne faut pas les prendre en compte ? Oui et non…

Oui, si la seule chose qui vous intéresse est le trafic du point de vue quantitatif.

Non, si la qualité du trafic obtenu est une chose importante pour vous (et tout nous pousse à croire que c'est le cas).

En effet, en dehors des outils généralistes que nous venons de voir, il existe deux familles de moteurs et d'annuaires qu'il peut être intéressant de prendre en compte dans le cadre d'une stratégie de référencement.

- Les outils de recherche thématiques, qui ne prennent en compte qu'une partie du Web, mais qui tentent de la traiter mieux que les généralistes que sont Google ou autres Yahoo!. Exemples : Rugby Engine (*http://www.rugby-engine.com/*) pour ce sport, Indexa (*http://www.indexa.fr/*) pour les sites professionnels ou Mamma Health (*http://www.mamma-health.com/*), spécialisé dans le domaine de la santé aux États-Unis.
- Les outils de recherche régionaux, qui n'effectuent des recherches que dans une région donnée comme la Bretagne (*http://www.breizhoo.fr/* ou *http://www.breizhat.com/*) ou d'autres…

Ces outils peuvent être soit des annuaires, soit des moteurs, soit un condensé des deux.

Dans ces cas, ne vous attendez pas à voir votre trafic exploser du fait de votre présence sur ces outils de recherche. En revanche le trafic généré sera certainement très bien ciblé. Vous n'aurez donc pas obligatoirement la quantité mais la qualité pourrait bien être au rendez-vous.

Pour trouver ce type d'outils, voici quelques pistes et sites intéressants :

- Indicateur (*http://www.indicateur.com/*)
- Search Engine Colossus (*http://www.searchenginecolossus.com/*)
- Enfin (*http://www.enfin.com/*)
- Abondance (*http://annuaire.abondance.com/*)
- Les annuaires (*http://www.lesannuaires.com/*)

Pour trouver d'autres liens, tentez des requêtes du type « annuaire santé » ou « moteur de recherche santé » sur un moteur généraliste. Les premiers résultats devraient être pertinents (ici dans le domaine de la santé bien sûr).

4

Optimisation des pages du site : les critères « in page »

Comme nous l'avons mentionné dans les chapitres précédents, pendant des années, le référencement (ou positionnement) d'un site web dans les pages de résultats d'un moteur de recherche s'est partagé en deux écoles bien distinctes :

- l'optimisation des pages web du site pour les rendre réactives aux critères de pertinence des moteurs de recherche ;
- la création de pages web spécifiques au référencement, permettant de ne pas toucher aux vraies pages du site. Ces pages étaient appelées « pages satellites », « pages alias » ou encore « doorway pages ».

Longtemps, ces deux écoles ont cohabité, ayant chacune leurs avantages, leurs inconvénients, leurs aficionados… Comme nous l'avons vu dans le premier chapitre de cet ouvrage, l'année 2006 a changé la donne. Il est aujourd'hui primordial d'optimiser les pages web de votre site à la source, sans passer par des rustines d'aucun type. Pour cela, vous vous êtes armé des mots-clés que vous avez définis dans le chapitre précédent. Il vous faut maintenant les placer dans les « zones chaudes » de vos pages : titre, texte, lien, URL, etc., pour que celles-ci soient le plus possible réactives par rapport aux critères de pertinence des moteurs. C'est ce que nous allons voir dans ce chapitre avec l'examen des critères dits « in page », c'est-à-dire concernant l'analyse par les moteurs du code HTML de vos documents. Le chapitre suivant traitera, quant à lui, des critères « off page » qui, sont davantage en rapport avec l'environnement de la page par le biais de notions comme la popularité, la réputation, la confiance, etc. Mais commençons à optimiser vos codes HTML...

Le contenu est capital, le contenu optimisé est visiblement capital !

Avant tout, il nous semble essentiel de souligner un point crucial : s'il est nécessaire de mettre en ligne des pages web optimisées par rapport aux critères de pertinence des outils de recherche, la valeur qualitative du contenu proposé est certainement bien plus importante. En effet, rien ne sert de faire la promotion d'un site web, par le référencement ou par un autre moyen, si ce site ne répond pas aux exigences du public visé au préalable.

La première étape dans la création d'un site web sera donc de réfléchir à son contenu et à l'adéquation de ce dernier aux besoins et aux attentes des internautes qui viendront le visiter. C'est essentiel et vital pour le succès d'un tel projet. Comme on le dit depuis plus de 10 ans sur le Web, *content is king!* – le contenu est le capital !

Ceci étant dit, cela ne suffit pas toujours… Il peut réellement s'avérer opportun de penser « moteurs de recherche » lorsque vous bâtissez vos pages. En effet, non seulement vous proposerez en ligne du bon contenu, mais l'optimisation que vous aurez créée pour vos pages lui donnera une bien meilleure visibilité sur les moteurs. Dans ce cas, *optimized content is emperor!* – le contenu optimisé induit sa visibilité !

L'erreur serait, à notre avis, de travailler sur l'optimisation des titres, textes et autres « zones chaudes » (nous y reviendrons dans ce chapitre) sans avoir auparavant travaillé la qualité du contenu lui-même. Nous insistons lourdement car c'est réellement primordial. En effet, si une bonne optimisation vous permettra de faire venir du monde sur vos pages au travers des moteurs, ce n'est pas cela qui fera rester les internautes sur votre site, y faire une ou des actions concrète(s), y revenir, ou qui fera en sorte que le bouche à oreille fonctionne pour faire venir d'autres personnes, etc. C'est le contenu que vous allez proposer en ligne qui va faire la différence entre trafic efficace, ciblé, et trafic stérile. Faire entrer un prospect dans une boutique vide ne sert pas à grand chose.

C'est aussi la qualité de ce que vous proposez en ligne qui va faire la différence au niveau des liens qui vont se créer vers votre source d'information. Et vous vous apercevrez vite, en lisant ce chapitre – et surtout le suivant –, que le lien est aujourd'hui l'application « qui tue » (nos amis anglophones parlent de *killer application*) du référencement.

Nous sommes persuadés que le meilleur référencement qui soit est celui qui permet, par une bonne et loyale optimisation des pages, de faire connaître le site au mieux, de le mettre en valeur et de donner une visibilité à un contenu de qualité. Et nous allons voir que cela est tout à fait possible !

Zone chaude 1 : balise <title>

La balise `<title>`, correspondant au titre de la page, au sens HTML du terme, est un champ essentiel dans le cadre d'une bonne optimisation puisqu'il est l'un des critères les plus importants (pour ne pas dire le plus important) pour la majeure partie des moteurs actuels et notamment Google.

Lorsque vous consultez un site, le titre d'une page est affiché en haut de la fenêtre de votre navigateur Internet, reprenant le contenu de la balise `<title>`. Sur la figure 4-1 est présenté un exemple sous Windows XP pour le site Abondance.com.

Figure 4-1

Le contenu de la balise <title> d'une page web apparaît dans la zone supérieure de la fenêtre du navigateur.

Dans le code HTML des pages web, le titre se situe entre les balises `<title>` (début de titre) et `</title>` (fin de titre). Exemple, toujours pour le site Abondance.com :

```
<title>R&eacute;f&eacute;rencement et recherche d'info : tout sur le r&eacute;
➥f&eacute;rencement de sites web avec Abondance</title>
```

Notez ici que les lettres accentuées sont codées en HTML. Par exemple, le caractère « é » est ainsi codé `&eacute;`. Nous y reviendrons plus loin.

Premier point important : si vous utilisez un éditeur HTML – et il y a de fortes chances pour que cela soit le cas – comme Dreamweaver ou Frontpage, étudiez le code HTML que l'outil logiciel vous permet de visualiser. Les balises de titre doivent être placées le plus haut possible dans la page. Idéalement, le code source de votre document doit commencer ainsi :

```
<!DOCTYPE html …>
<html>
  <head>
    <title>Titre de votre page</title>
```

Déplacez, après le titre, les balises `<meta>` (voir plus loin) ou toute autre information qui serait ajoutée de façon fortuite par votre éditeur HTML entre `<head>` et `<title>`. Plus la balise de titre sera placée haut dans le fichier HTML, meilleures seront vos chances d'optimiser le classement de votre page.

Libellé du titre

En ce qui concerne le libellé du titre, choisissez une expression qui affiche le plus possible de mots-clés déterminants et caractéristiques de votre activité et du contenu de la page.

Évitez les expressions banales comme « Bienvenue », « Homepage » ou, pire encore, « Bienvenue sur ma Homepage », « Bienvenue sur notre site web », « Welcome », « Accueil », « Page d'accueil », etc. Tous ces titres sont à proscrire car ils ne sont pas du tout assez descriptifs. Le titre d'une page d'accueil, par exemple, doit contenir au moins

le nom de votre entreprise/entité/organisme/association et décrire en quelques mots son activité.

Deux erreurs courantes dans les titres de page

Les deux erreurs le plus souvent commises sur les titres de page sont les suivantes :

– libellé non explicite du contenu de la page : « Bienvenue sur notre site web », « Homepage », etc. ;

– même titre pour toutes les pages du site.

Le simple fait de corriger ces deux points améliore très fortement un référencement…

De même, n'oubliez pas de donner un titre à vos pages. Comme le montre la figure 4-2, le nombre de pages web n'ayant pas de titre (souvent, la mention « Untitled » apparaît dans les résultats des moteurs) est considérable !

Figure 4-2

Le nombre de pages dont le titre n'est pas renseigné est considérable…

On s'accorde à penser le plus souvent qu'un titre bien optimisé propose au plus 10 mots descriptifs dans son libellé. Ne proposez donc pas de titres trop longs (une fourchette entre 5 et 10 termes est un bon compromis). Dans ces dix mots, vous ne compterez pas

les *stop words*, ou mots-clés vides, comme « le », « la », « à », « au », « vos », etc. Retenez donc que les titres de vos pages doivent contenir de 5 à 10 mots descriptifs…

En revanche, rien ne prouve actuellement que l'emplacement d'un mot dans le titre (au début, au milieu ou à la fin) confère à la page qui le contient un avantage en termes de positionnement. Si les mots importants pour décrire le contenu de la page sont dans le titre, c'est déjà une excellente chose.

Pour votre titre, et notamment celui de vos pages internes, n'hésitez pas à mettre en place une structure de ce type :

```
<title>[Contenu] - [Rubrique] - [Source]</title>
```

où :

- `[Contenu]` reprend le titre éditorial de la page (qui sera inséré dans une balise `h1`, comme nous le verrons plus loin).
- `[Rubrique]` est la rubrique dans laquelle la page est proposée sur le site. Cette zone n'est donc pas obligatoire sur une page d'accueil, par exemple.
- `[Source]` est le nom du site, sa marque.

En effet, chaque page de votre site, même – et surtout – les pages internes, peuvent se retrouver dans les pages de résultats des moteurs sur une requête donnée. L'internaute peut donc y accéder directement, sans passer par votre page d'accueil. Il nous semble important que l'internaute sache en un coup d'œil :

- ce qu'il va trouver dans la page (le début de la balise) ;
- qui lui fournit l'information (la fin de la balise).

Vous renforcez ainsi la confiance que l'utilisateur du réseau peut avoir en vous et lui fournissez de nombreuses informations sur ce que vous lui proposez. Il y a fort à parier qu'il vous en sera reconnaissant…

Voici quelques exemples :

```
<title>Nasa : Endeavour décolle, endommagée par ses débris ? - Sciences - LCI</title>
<title>Hadopi 2 : les députés tentent d'atténuer le texte - High Tech - Actualité
➥Challenges.fr</title>
<title>Livre Réussir son référencement Web par O. Andrieu - Informatique et nouvelles
➥technologies - Librairie Eyrolles</title>
```

Attention à ne pas répéter trop souvent certains mots-clés, cela pourrait être pris pour du spam (fraude) par certains moteurs. Si un mot est répété, essayez d'en espacer le plus possible les occurrences. Par exemple, positionnez un mot important deux fois dans le titre en le plaçant en 1re et 7e position, en 2e et 8e, etc.

Si un mot précis caractérise de façon quasi parfaite votre activité (il peut s'agir de votre nom ou d'un terme professionnel, ou autre), essayez cependant de le placer deux fois dans le titre de vos pages. En d'autres termes, tentez de faire en sorte que chaque mot du

titre soit unique (non répété) sauf pour un terme important qui sera proposé en double. Voici un exemple pour la page d'accueil du site Abondance :

```
<title>R&eacute;f&eacute;rencement et recherche d'info : tout sur le r&eacute;
➥f&eacute;rencement de sites web avec Abondance</title>
```

Le titre comporte moins de dix mots et contient seulement deux fois un mot important (référencement) pour lequel le titre sera plus particulièrement réactif. Mais cette « astuce », beaucoup utilisée, aurait tendance à moins bien fonctionner avec le temps...

N'oubliez pas les codes HTML pour les lettres accentuées (`&eacute;` pour la lettre « é », par exemple), car il s'agit là d'un texte à part entière. Sachez qu'il existe d'autres écoles pour le codage des lettres accentuées. Certains référenceurs ne les codent pas (« é » reste « é » et pas `&eacute;`), d'autres les proposent non accentuées, etc. Il ne semble pas exister de recette miracle dans ce domaine. À vous de voir la solution qui vous convient le mieux.

Accentuation et référencement

L'accentuation et le codage des caractères est un sujet complexe et très long à appréhender… Disons, pour résumer, qu'il existe deux codages principaux pour un site web francophone.

– Le codage ISO-8859-1 ou ISO-8859-15 qui est bien adapté aux alphabets occidentaux.
– Le codage UTF-8 qui sera préféré pour des pages web plus « universelles » affichant des alphabets comme le chinois, le russe, le grec, etc.

Dans les deux cas, une balise meta spécifique devra être indiquée en début de page pour indiquer le codage utilisé (exemple : `<meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1">`).

Ensuite, on peut résumer la situation en disant que les caractères devront être codés en HTML (`&eacute;`) si le codage est ISO-8859-15 et pourront être éventuellement laissés « tels quels » (« é ») dans le code HTML pour l'UTF-8. Que les spécialistes du domaine nous pardonnent ici ces approximations…

Sachez cependant que les lettres accentuées non codées en HTML s'affichent parfois mal sur certains navigateurs et que les lettres non accentuées (alors qu'elles devraient l'être) peuvent choquer certains internautes qui pourraient croire à des fautes de frappe ou d'orthographe. Eh oui, n'oubliez jamais que le titre est lu par les internautes !

En effet, comme le montre la figure 4-3, le titre est la première information lue par l'utilisateur du réseau lorsqu'un moteur lui présente l'une de vos pages dans ses résultats (la plupart du temps, le titre est ce qui est représenté en plus gros, donc le plus visible). La formulation doit ainsi être explicite et présenter le contenu du document. L'internaute aura le plus souvent le choix entre dix pages – dix liens – présentés par le moteur comme résultat de sa requête. Il choisira peut-être celle qui proposera le titre le plus clair, le plus descriptif, mais aussi le plus « sexy » par rapport à sa demande. Un bon compromis est à trouver entre la lisibilité et l'efficacité, et donc un placement optimisé des mots-clés.

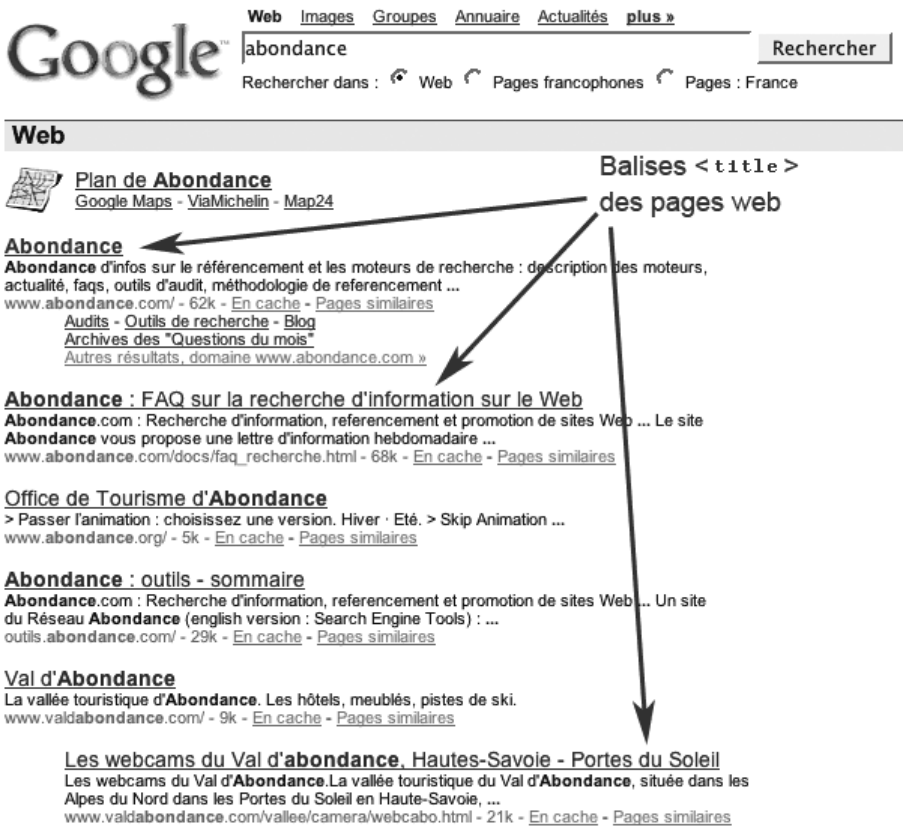


Figure 4-3

Le contenu de la balise <title> est repris par les moteurs de recherche dans leurs pages de résultats pour désigner les pages en question.

N'oubliez pas que le titre est aussi l'information qui est affichée en premier sur le navigateur lorsqu'on appelle la page. Parfois bien avant que le contenu ne s'affiche !

De plus, lorsqu'un visiteur placera un signet (favori, *bookmark*) sur votre page, c'est le titre positionné entre les balises `<title>` et `</title>` qui sera pris en compte en tant qu'intitulé dans le menu des marque-pages. Faites donc en sorte que cet intitulé rappelle à l'internaute le contenu proposé.

Il est important de bien prendre en compte un compromis entre lisibilité (un titre qui veut dire quelque chose) et optimisation (intégration d'un maximum de mots-clés pertinents et descriptifs). Une suite de mots-clés séparés par une virgule, par exemple, pourrait être considérée comme très optimisée, mais elle sera, en revanche, très peu lisible :

```
<title>abondance, annuaire, référencement, moteur de recherche…</title>
```

Pour ce qui est de la casse des lettres (minuscules ou majuscules), cela n'a pas réellement d'importance aujourd'hui : les moteurs de recherche traitent indifféremment des mots comme IBM, Ibm ou ibm. Choisissez donc l'occurrence qui semble la plus logique pour les termes que vous choisirez. En revanche, n'oubliez pas que si votre titre est réactif au mot-clé « chaussures » ou à l'expression « chaussures de tennis », comme celui-ci :

```
<title>Chaussures de tennis pour terre battue, dur et herbe. Des chaussures de
➥marque !</title>
```

Il y a fort peu de chance pour que ce titre soit optimisé pour le singulier : « chaussure »… La plupart des moteurs traitent différemment le singulier/pluriel, masculin/féminin, etc. Nous y reviendrons très prochainement lorsque nous parlerons du « texte visible ». Le choix initial des mots-clés est donc prédominant.

La création de titres efficaces est une phase essentielle de la promotion de votre site. À vous de vous entraîner en vous aidant des mots-clés que vous avez répertoriés dans votre phase de réflexion préalable sur le référencement (voir chapitre 3). Comme nous l'avons déjà dit pour la quasi-totalité des moteurs de recherche, le titre est le critère de pertinence numéro un ! Raison de plus pour le soigner du mieux possible.

Supposons que votre activité consiste à vendre des chaussures de sport, notamment des « tennis ». Essayez pour votre page d'accueil un titre comme :

```
<title>Chaussures de tennis Stela : terre battue, dur et herbe. Stela créateur à Paris,
➥France</title>
```

ou :

```
<title>Stela, fabricant de chaussures de tennis pour terre battue, dur et herbe &agrave;
➥Paris, France</title>
```

qui sera préférable aux exemples suivants :

```
<title>Stela : bienvenue</title>
<title>Bienvenue sur le site web de Stela</title>
<title>Chaussures Stela</title>
<title>Chaussures de tennis</title>
```

Le titre « Bienvenue sur le site web de Stela » est un bon exemple de « cyber hara-kiri ». Seul le mot « Stela » pourra faire l'objet d'une recherche par un internaute dans l'intitulé.

N'oubliez pas également de donner des informations géographiques (ici Paris et France) si vous pensez qu'elles peuvent être importantes dans le cadre d'une recherche.

Certaines sociétés insèrent également leur slogan dans le titre de la page d'accueil. Exemple pour un site d'optique (fictif) :

```
<title>Être bien lu, c'est être bien vu</title>
```

Ce titre réjouira le service Communication de l'entreprise puisqu'il affiche son slogan. Mais pour ce qui est du référencement, il est catastrophique car il ne contient ni le nom de l'entreprise, ni de mots décrivant son activité (optique, opticien, lunettes, etc.). Ce type de problème, il est vrai, cause parfois quelques frictions entre le service

Communication et les gens responsables de la promotion du site sur Internet… Une solution : laisser la page d'accueil telle quelle et optimiser plutôt les pages internes. Ce n'est pas une solution parfaite, loin de là (la page d'accueil est très importante pour les moteurs) mais faute de grives…

Titres multilingues

Si votre site s'adresse à plusieurs communautés linguistiques, les pages bilingues ou trilingues sont à déconseiller. En règle générale, il vaut mieux scinder votre site web en plusieurs entités distinctes, avec des pages différentes, des titres différents, et donc des mots-clés différents. Voir le chapitre 5 dont une partie est entièrement consacrée à cette problématique.

Utiliser des pages qui contiennent du texte en deux langues nécessiterait de créer des titres également bilingues. D'une part, cela induirait des répétitions qui pourraient passer pour du spam. D'autre part, la présence conjointe de mots en français et en anglais risque de désorienter certains internautes, et d'altérer la lisibilité du titre. Enfin, les moteurs n'aiment pas les pages bilingues, car ils ne peuvent y reconnaître une langue unique. Et pour un moteur, lorsqu'il n'y a pas une langue unique, il n'y a pas de langue ! Ainsi, une page en anglais ET en français ne sera peut-être pas trouvée sur Google France avec l'option « Pages francophones ». Nous en reparlerons…

En conclusion, on peut dire qu'un titre ne doit être rédigé que dans une seule langue !

Un titre pour chaque page !

Toutes les pages de votre site doivent recevoir un titre différent et résumant leur contenu en 10 mots au maximum. Si le titre de la page d'accueil peut être assez généraliste et contenir des mots-clés plutôt génériques par rapport à votre activité, n'hésitez pas à proposer des titres de plus en plus précis, contenant des termes de plus en plus pointus au fur et à mesure que l'internaute descendra l'arborescence de votre site.

Ceci est assez logique. La page d'accueil d'un site est souvent très généraliste. Elle présente les grands thèmes qui y sont développés. Logiquement, plus on navigue dans le site, plus on descend dans l'arborescence, et plus l'information doit y être précise et pointue. Le titre devant résumer en quelques mots le contenu de vos pages, il est normal qu'il suive la même évolution.

Faites en sorte de passer du temps sur les titres de chacune de vos pages : c'est très important. Osons un chiffre : un site web qui affiche des pages ayant chacune un titre optimisé résumant le contenu proposé dans chaque document a déjà fait quasiment 40 % du travail de référencement.

Même si votre site est réalisé en *frames* (fenêtres ou cadres distincts, voir chapitre 7) – sachez à ce sujet que de nos jours, l'utilisation de frames est fortement déconseillée par le W3C –, chacune de vos pages (« pages mères » descriptives des frames et « pages filles » de contenu) doit avoir un bon titre. En effet, pour un moteur de recherche, chaque

page web est considérée comme un document à part entière, que ce soit une page « fille » ou une page « mère » dans le cas des frames.

L'idéal, pour chaque page, serait donc que le titre résume en 10 mots au maximum ce qui est proposé dans ladite page, le tout en contenant les mots-clés importants par rapport à ce contenu. En résumé, il faut un titre :

- contenant 10 mots au maximum ;
- qui résume le contenu de la page en question ;
- qui contient les mots-clés important par rapport au contenu de cette page ;
- qui, pour les pages internes, reprend le titre éditorial de la page (balise `h1`) au début, puis continue avec la rubrique de la page et se termine avec la mention de la source.

Dernier conseil : il vaut mieux parfois mettre en ligne plusieurs petites pages qui proposent une thématique unique, décrite par un titre performant (émaillé de mots-clés bien ciblés) qu'un seul grand document qui traite de sujets divers et qui possède donc un titre plus vague car devant s'adapter à de nombreux thèmes. Plus le sujet traité dans la page sera précis, plus il vous sera possible de créer un titre explicite et donc efficace en regard des critères des moteurs de recherche. Ne l'oubliez pas lors de l'élaboration de l'arborescence de votre site !

Le titre des pages de votre site demandera donc à être particulièrement soigné lors de la phase de (re)construction de vos documents web. N'hésitez pas à suivre les quelques conseils donnés dans ce chapitre, cela devrait grandement aider votre future visibilité. Cela semble vite dit et très simple, mais vous vous apercevrez assez rapidement que c'est loin d'être le cas et qu'il s'agit surtout ici de prendre le temps d'optimiser chaque titre de chaque page… Cela dit, le jeu en vaut réellement la chandelle, car un site web proposant un contenu de qualité avec des titres bien pensés a fait une bonne part du chemin qui le mène à un bon référencement.

Pour résumer

Voici quelques conseils pour bien optimiser les titres de vos pages :

– placez la balise `<title>` le plus haut possible dans votre code HTML ;
– un titre de page web est avant tout descriptif du contenu de la page en question ;
– insérez le plus possible de mots-clés déterminants et caractéristiques du contenu de la page ;
– ne dépassez pas 10 mots par titre (hors « mots vides ») ;
– doublez éventuellement un mot important ;
– proscrivez les titres bilingues ou trilingues, etc. ;
– le titre doit parfaitement résumer le contenu de la page ;
– le titre d'une page d'accueil est souvent assez générique et se précise au fil de l'arborescence ;
– chaque page de votre site doit avoir un titre qui lui est propre (et qui doit être optimisé).
– Les pages internes doivent être dotées d'un titre commençant par la reprise du titre éditorial (`h1`) du document, suivi par le nom de la rubrique, puis le nom de la source.

Insérer des codes ASCII dans le titre : bonne ou mauvaise idée ?

Pendant quelques temps, on a vu fleurir, dans les pages de résultats de Google, des codes ASCII ou Unicode utilisés dans les balises `<title>` et meta `description` (voir plus loin) de certaines pages, permettant de « mettre en avant » certains liens. En voici quelques exemples.

Figure 4-4

Des codes ASCII ou Unicode sont insérés pour mettre en avant certaines pages dans les résultats des moteurs.

On pourrait multiplier ces exemples à l'envi, car les caractères de ce type sont légion. Il est clair que, visuellement parlant, l'œil est attiré par ces résultats « qui sortent de l'ordinaire » et donc par les snippets qui utilisent ce type d'« artifice »...

Bien entendu, il ne s'agit en rien ici d'astuces permettant de mieux positionner une page, l'avantage – qui nous semble bien réel – n'est donc que visuel et permet d'assurer un meilleur taux de clic sur une page déjà optimisée – et donc positionnée – par ailleurs.

La question qui se pose porte donc sur la « légalité » et l'éthique de ces méthodes au sens des prérogatives de Google et de ses règles de « bon référencement ».

Pour en avoir le cœur net, Sébastien Billard, spécialiste SEO bien connu et basé dans le Nord de la France, avait posé en 2008 la question sur le groupe de discussion Google pour webmasters (*http://s.billard.free.fr/referencement/?2008/06/26/491-les-symboles-ascii-ou-unicode-dans-la-balise-title-sont-ils-du-spam*). Il ressort de cette requête (*http://groups.google.com/group/Google_Webmaster_Help-chit-chat-fr/browse_frm/thread/11ae14da97b590bb*) que Google ne voit pas d'un très bon œil ce type de « manipulation » et que les sites qui les utilisent peuvent être pénalisés (même si on peut penser que la pénalisation sera faible, la « faute » étant loin d'être grave, voir chapitre 8...). Selon Google, ce type de pratique serait assimilable à son conseil énoncé ainsi dans ses « guidelines » : « Évitez les « astuces » destinées à améliorer le classement de votre site par les moteurs de recherche » (*http://www.google.com/support/webmasters/bin/answer.py?answer=35769#quality*).

Aussi, le conseil que nous pouvons donner aux webmasters qui voudraient tenter ce type de manœuvre est de « raison garder » et de ne pas aller trop loin dans ce type d'affichage. On peut penser qu'un ou deux caractères Unicode dans le snippet (`title` + meta `description`) passeront sans problèmes, s'ils sont en rapport avec le contenu du site (un cœur pour un site de rencontres, un soleil pour la météo, par exemple).

Mais si leur utilisation commence à « sentir la manipulation visuelle » et n'a pour seule motivation que la volonté de mettre en avant un résultat par rapport à ses «concurrents », il se peut que la pénalisation ne soit pas loin. Bref, comme souvent dans le monde du référencement, votre optimisation sera surtout question de bon sens...

Zone chaude 2 : texte visible

Optimiser les titres de ses pages ne suffit pas. On dit souvent des moteurs de recherche qu'ils sont des « obsédés textuels » : seul le texte d'une page leur importe, pas les images ou autres animations. Ils adorent « comprendre » le contenu d'un document en lisant son « texte visible ». Tout d'abord, définissons clairement ce que nous entendons par cette expression :

- « texte » car ce dont nous allons parler ici est le contenu textuel de vos pages, c'est-à-dire tout le contenu que vous pouvez sélectionner avec votre souris, copier puis coller dans un traitement de texte comme Word. En d'autres termes, tout le texte affiché dans le navigateur et qui peut être identifié dans votre code source HTML. Ainsi, un texte inclus dans une image n'est pas considéré comme étant au format textuel. Il en est de même pour un texte inséré dans une animation Flash, etc.

- « visible » car nous ne parlerons ici que du texte proposé loyalement sur les pages web, sans traiter des cas de spam datant du paléolithique inférieur du Web et consistant à insérer, par exemple, du texte en blanc sur fond blanc (voire en jaune très clair sur fond blanc, quelle misère…), invisible pour l'internaute mais soi-disant pris en compte par le moteur. S'il est vrai que ce type de spam marche parfois sur certains outils de recherche, et pas des moindres (même sur ceux qui disent combattre ce type de pratique), il s'agit clairement de fraude de bas étage et nous n'en parlerons

donc pas. De plus, lorsqu'un internaute découvre la supercherie (et ce n'est pas vraiment très compliqué), la perte de crédibilité envers le site coupable est telle que cela devrait décourager tout webmaster sérieux de commettre ce type de « méfait ». De la même façon, toute tentative visant à cacher du texte dans un code HTML (et Dieu sait si techniquement parlant il existe de nombreuses possibilités de le faire) sera considérée comme du spam et ne pourra donc pas bénéficier de l'adjectif « visible »…

La notion de texte visible signifie également que nous ne traiterons pas ici, par exemple, de l'attribut `alt` des balises images, qui sera étudié plus loin, à la fin de ce chapitre. Même si ce texte est visible lorsqu'on passe la souris sur l'image (sur certains navigateurs), il ne l'est pas sans action précise sur la page. Idem pour les commentaires (non pris en compte par les moteurs) et tout contenu du code HTML non visualisé à l'écran du navigateur.

Regardez vos pages avec l'œil du spider !

Lorsque vous allez consulter un site web, vous utilisez bien sûr l'œil de l'internaute « humain » qui regarde l'écran de son ordinateur. Mais les spiders des moteurs de recherche ont, pour leur part, une vision toute autre de vos pages. Voici deux façons astucieuses de vous mettre à la place d'un robot et de visualiser votre site sous un œil nouveau et... parfois assez surprenant.

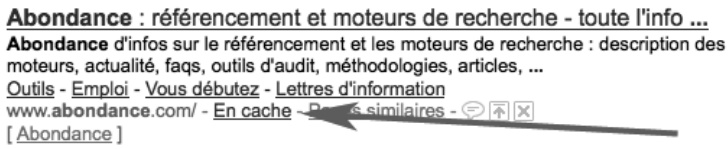
Tout d'abord, il est important de bien comprendre que les robots des moteurs n'ont qu'une vision parcellaire de vos documents. On les compare souvent, avec raison, à un utilisateur aveugle qui utiliserait un système de reconnaissance vocale pour comprendre ce que contiennent vos pages. Bref, pour les spiders, l'essentiel est le texte, le texte et encore le texte ! De vrais « obsédés textuels » on vous dit !

Le cache de Google

La première façon de « vivre dans la peau d'un spider » et de visualiser ce si important contenu textuel est d'utiliser le cache de Google. Pour ce faire, recherchez votre site sur le moteur et visualisez le résultat obtenu, comme nous l'avons fait sur la figure suivante avec le site Abondance.com sur Google.

Figure 4-5

Le site Abondance présenté dans les résultats de Google



À ce stade, cliquez sur le lien « En cache » proposé à droite de l'URL. Vous obtenez la page représentée par la figure 4-6.

Figure 4-6

Version en cache de la page d'accueil du site Abondance

En haut de la page, dans la zone textuelle grisée proposée par Google, vous trouverez cette phrase : « Version en texte seul ». Cliquez sur le lien ainsi proposé. Vous obtenez alors une vision *spider friendly* de votre page, proposant uniquement le texte, donc ce que voient les spiders des moteurs.

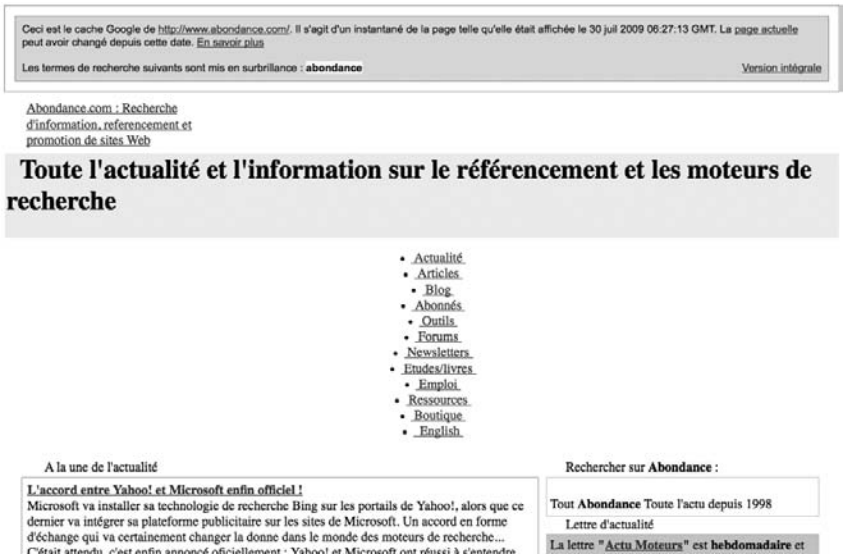


Figure 4-7

Version textuelle de la page, lue par les spiders des moteurs

Ce n'est clairement pas la même chose. Faites le test avec votre site et vous risquez peut-être d'être surpris... Notez également que les feuilles de styles (CSS) ne sont plus appliquées non plus, ce qui peut tout changer... C'est donc ainsi que Google et ses compères voient votre site !

Les simulateurs de spider

Le petit inconvénient du fait d'utiliser le cache de Google est que ce dernier affichage, s'il donne un excellent aperçu des données prises en compte par le moteur, respecte la majeure partie des informations de mise en page, notamment pour ce qui est des tableaux (`<table>`) et, plus globalement, de la façon dont les « blocs de texte » sont agencés dans la page web (si ces informations ne sont pas présentes dans les CSS). Or, un spider a le plus souvent une vision beaucoup plus « linéaire » des données : il lit les codes HTML de haut en bas sans réellement tenir compte de la mise en page proposée dans la fenêtre d'un navigateur.

Il est alors possible, pour obtenir une vision encore plus réaliste de la façon dont un moteur lit vos documents, d'utiliser un « simulateur de spider » comme Spider Simulator (*http://www.spider-simulator.com/*), du réseau Abondance, qui donnera alors, dans sa page de résultats, la vision linéaire présentée sur la figure 4-8, plus proche de celle des robots.

Figure 4-8

Spider Simulator donne une version plus linéaire du texte lu dans la page.

Seul inconvénient, le texte est fourni « tel quel » donc assez difficilement digérable et facile à lire s'il est... abondant ! Mais les spiders n'ont-ils pas de gros estomacs ?

Toujours est-il que cette vision est certainement la plus proche de celle d'un spider, et notamment en ce qui concerne son ordre de lecture. Aussi, dans le cadre de votre référencement, il vous faudra tout particulièrement soigner les 100 premiers mots ainsi détectés par ces outils, comme nous allons le voir ci-après. Ils doivent parfaitement définir l'univers sémantique de votre page. Est-ce le cas pour votre site web ?

Autres simulateurs de spider

Voici quelques outils similaires pour effectuer vos tests :

– Spider Simulator (réseau Abondance) – *http://www.spider-simulator.com/*

– Search Engine Spider Simulator – *http://www.webconfs.com/search-engine-spider-simulator.php*

– SEO Tools - Spider Simulator – *http://www.seochat.com/seo-tools/spider-simulator/*

– Webmaster Toolkit - Search Engine Spider Simulator – *http://www.webmaster-toolkit.com/search-engine-simulator.shtml*

Testez-les pour trouver celui qui vous convient le mieux. Sachez cependant qu'ils renvoient tous à peu près la même information...

Autres possibilités

Il existe d'autres possibilités pour regarder votre site avec les mêmes yeux que ceux d'un robot de moteur. Vous pouvez, par exemple, utiliser un navigateur tel que Lynx (*http://lynx.isc.org/*) qui ne lit que le texte des pages web. La version qu'il fournira de vos pages sera donc très proche de celle d'un spider. Vous pouvez également installer un *add-on* pour Firefox comme Web Developer (*https://addons.mozilla.org/fr/firefox/addon/60*) qui permet de désactiver le JavaScript, les CSS, etc., pour afficher une version « brute » de votre page. Il existe de nombreux outils très utiles pour les référenceurs, n'hésitez pas à vous reporter aux annexes du présent ouvrage.

Ceci étant dit, nous allons enlever notre habit de spider et reprendre celui de référenceur pour commencer notre exploration des différentes possibilités d'optimisation du texte visible des pages…

Localisation du texte

L'un des critères importants pris en compte par les moteurs lors du calcul de pertinence d'une page par rapport à un mot donné est la présence de ce terme au début du document plutôt qu'à la fin. Plus le mot en question sera placé haut dans la page (le plus proche possible de la balise `<body>` dans le code source en langage HTML), plus sa présence sera jugée pertinente.

Globalement, on a tendance à dire que le premier paragraphe des pages est le plus important. Une page optimisée doit contenir au minimum 100 mots, si possible plus, mais les 30 à 50 premiers termes (les deux à trois premières phrases) seront primordiaux.

Vous devez donc insérer dans ce premier paragraphe, le nom de votre entreprise/association/entité et les mots-clés importants pour votre activité, décrivant parfaitement le contenu du document en question.

Un exemple est donné en figure 4-9 d'après le site Abondance.com.

Figure 4-9

Exemple de texte mis en valeur sur une page web. Le logo (à gauche) et le bandeau publicitaire sont des images. Elles ne sont donc pas prises en compte textuellement.

Juste en dessous, la phrase suivante est indiquée : « Toute l'info et l'actu sur les annuaires et moteurs de recherche : Recherche d'information et référencement ». Cette phrase n'est pas là par hasard. Elle contient de nombreux termes importants pour décrire le site : « annuaires », « moteurs de recherche », « référencement », « recherche d'information ».

Puis, si vous regardez à l'intérieur du code HTML de la page, cette phrase est suivie par les termes situés à la gauche de l'écran et qui décrivent les grandes rubriques du site : « Actualité », « Dossiers/Articles », « Outils de recherche », « Audits », « Forums/Chat », etc.

La page a ainsi été conçue pour que tout le texte situé « en haut du code » contienne les mots-clés importants pour décrire son contenu.

Une conséquence directe et importante est qu'il sera difficile de référencer les sites en Flash et que ces sites rencontreront des problèmes pour être bien positionnés. Si le moteur n'y trouve pas, au format texte, vos mots-clés favoris, il sera difficile pour vous d'être bien positionné dessus. Un bon titre seul ne suffit pas toujours pour être bien classé. La présence d'un contenu textuel qui allie la qualité à la quantité est importante. Nous y reviendrons…

La mise en exergue du texte

Les moteurs de recherche privilégient également les mots qui sont mis en exergue dans les pages web.

Premier exemple : la mise en gras. N'hésitez pas à placer vos mots importants en gras (balise `<strong>` en HTML), comme ici :

```
Nous vendons des <strong>amortisseurs</strong> et toutes les <strong>pi&egrave;ces
➥d&eacute;tach&eacute;es</strong> pour votre voiture.
```

Ce qui donnera l'affichage suivant :

Nous vendons des **amortisseurs** et toutes les **pièces détachées** pour votre voiture.

Une page qui contient le mot « amortisseurs » en gras sera donc mieux positionnée, toutes choses égales par ailleurs, qu'une page contenant ce même mot en romain.

Sachez également que, pour séparer le fond de la forme, le W3C préconise l'utilisation de `<strong>` par rapport à `<b>`.

N'en profitez pas non plus pour mettre tout votre texte en gras, ce qui donnerait un résultat assez horrible et pénible à la lecture. Seuls les mots-clés importants doivent être ainsi mis en exergue. N'oubliez jamais qu'une page web est avant tout créée pour les internautes.

Autre point pour mettre en avant votre texte : les liens. Aujourd'hui, pour tous les moteurs, le fait qu'un mot soit cliquable est important. Il est à la fois important par rapport au poids du mot en question dans la page mais également pour le positionnement de la page cible. En effet, si vous insérez le code suivant dans votre page :

```
Voici des informations sur l'<a href="http://www.votresite.com/assurance.html">
➥assurance-vie</a>
```

Le résultat sur un navigateur sera le suivant :

> Voici des informations sur l'assurance-vie.

Ceci aura deux conséquences :

- la page contenant ce code sera mieux référencée pour l'expression « assurance-vie » car ce mot y est cliquable ;
- la page pointée par le lien (ici `assurance.html`) le sera aussi pour cette même expression (nous y reviendrons au chapitre 5). Il s'agit ici de la fameuse notion de « réputation ».

Vous faites donc d'une pierre deux coups. Google, notamment, est très sensible au texte des liens pour classer ces pages de destination. Tenez-en compte.

Voici, par exemple, un exemple de mauvais lien :

> Pour avoir des informations sur l'assurance-vie, cliquez ici.

A priori, l'expression « cliquez ici » n'est pas vraiment importante pour votre activité et il y a peu de chances pour qu'un internaute la saisisse sur un moteur de recherche. Aussi n'est-il pas vraiment nécessaire de la mettre en valeur dans vos pages en la rendant cliquable.

En revanche, le texte suivant sera très bien optimisé :

> Voici des informations sur l'**assurance-vie**.

Les mots importants sont « cliquables » et en gras. Bravo !

Les moteurs prennent également en compte les balises `<hn>` (`<h1>` à `<h6>`) pour donner un poids aux pages web sur une requête donnée. Si un mot est compris entre des balises `<h1>` et `</h1>` (plus grande taille de titre en HTML), cela a un poids important pour le classement du document sur ce terme.

Un exemple avec la phrase en haut de la page d'accueil du site Abondance.com :

```
<h1>Toute l'info et l'actu sur les annuaires et moteurs de recherche :Recherche
➥d'information et r&eacute;f&eacute;rencement</h1>
```

Utilisez les balises <hn> à bon escient

Les balises <hn> ont été conçues, au départ, pour indiquer un niveau de titre dans un document HTML, h1 étant le niveau le plus haut. Pour en savoir plus sur cette balise, nous vous conseillons de consulter le site du W3C (*World Wide Web Consortium*) : *http://www.w3.org/MarkUp/html-spec/html-spec_5.html#SEC5.4/.*

Les mots-clés à l'intérieur de cette balise ont alors un poids plus fort que s'ils étaient en gras ou, pire, en romain.

L'inconvénient historique majeur de cette balise <hn> est qu'elle est tombée au fur et à mesure en désuétude, car il n'a longtemps pas été possible, par défaut, de maîtriser la façon dont son contenu était affiché (police de caractères, couleur, taille, etc.).

Mais la situation a changé. La solution consiste à utiliser les feuilles de styles (CSS ou *Cascading Style Sheet*) pour redéfinir ces balises afin de les faire apparaître comme bon vous semble. Exemple avec la feuille de styles utilisée pour le site Abondance.com :

```
h1
{
font-family     : Verdana,Helvetica;
font-size       : 10px;
color           : #3366cc;
font-style      : normal;
font-weight     : bold;
text-decoration : none;
}

h2
{
font-family     : Verdana,Helvetica;
font-size       : 1.4em;
color           : #000055;
font-style      : normal;
font-weight     : bold;
text-decoration : none;
}

h3
{
font-family     : Verdana,Helvetica;
font-size       : 1.2em;
color           : #000055;
font-style      : normal;
font-weight     : bold;
text-decoration : none;
}
```

Toute autre version est évidemment possible, en fonction de la charte graphique de votre site. Une fois cette balise ainsi redéfinie, vous pouvez afficher dans votre page la phrase clé, descriptive de son contenu et contenant vos termes importants.

Notons que vous pouvez bien entendu utiliser les balises `<h1>` à `<h6>`, mais `<h1>` étant prévue pour les titres de plus haut poids, elle sera plus intéressante pour mettre en exergue vos mots-clés.

Enfin, cette astuce n'est pas spécifique à Google. La plupart des moteurs de recherche connus prennent davantage en compte les termes soulignés par une balise `<h1>`.

Dans une page interne, présentant un produit, par exemple, il sera essentiel que la balise `<h1>` contienne le titre éditorial de votre page. Exemple :

```
<h1>Produit P1, la peinture à l'eau indispensable pour tous vos travaux de bâtiment</h1>
```

De plus, nous avons vu auparavant que le contenu de cette balise `<h1>` sera repris dans la balise `<title>` de cette même page :

```
<title>Produit P1, la peinture à l'eau indispensable pour tous vos travaux de bâtiment -
➥Peintures professionnelles - Vospeintures.fr</title>
```

Pour terminer, sachez qu'il vaut mieux éviter d'afficher, par exemple, tous les textes de vos pages entre des balises `<h1>`. Il n'est pas certain que le résultat soit à la hauteur de vos espérances. Réservez cet emploi à des titres et structurez vos contenus à l'aide de ces balises : c'est quand même pour cela que cette balise a été créée. Encore une fois, restez dans les limites du bon sens pour optimiser vos pages…

Dans le domaine de la presse, par exemple, on voit souvent cette pratique pour une page qui traite d'une actualité :

- Titre éditorial : en `<h1>`.
- Chapô (résumé en début d'article qui en décrit le contenu) en `<h2>`.
- Sous-titres en `<h3>`.
- Mots importants en gras (`<strong>`).

Il s'agit d'une optimisation très intéressante et efficace. Elle peut tout à fait être utilisée dans un autre contexte que celui de la presse...

Le titre éditorial h1 : la base (et la balise) d'un bon référencement

On ne dira jamais assez l'importance capitale de vos titres éditoriaux, affichés dans la balise `<h1>`. En effet, cette zone va nous servir, dans la suite de ce chapitre et dans le chapitre suivant, pour être recopiée dans la balise `<title>`, dans l'URL, etc., dans le cadre d'une optimisation la plus cohérente possible de votre site. Si votre titre éditorial `h1` est « mauvais » (trop court, trop long, pas assez descriptif, etc.), c'est toute votre optimisation qui sera « bancale » par la suite. Pensez-y dès le départ !

Les moteurs prennent-ils en compte les feuilles de styles ?

Voici une question qui revient souvent dans le petit monde du référencement : à la mi-2009, il semblerait que la plupart des moteurs de recherche ne prennent pas en compte les feuilles de styles (CSS). Donc, si un texte est paramétré en gras dans la feuille de styles correspondante, il y a de fortes chances qu'il ne soit pas considéré comme tel par le moteur.

Une ruse pour contourner cela consiste à créer une feuille de styles pour le texte en question et d'y indiquer un style « roman » (`font-weight:normal`), puis de proposer la mise en gras dans la page elle-même.

Exemple de feuille de styles :

```
.exemple
{
font-family     : Verdana,Helvetica;
font-size       : 12px;
color           : #3b3b3b;
font-style      : normal;
font-weight     : normal;
text-decoration : none;
}
```

Exemple de texte dans la page :

```
<span class="exemple">Ceci est un texte en roman. <strong>Ceci est un texte en gras.
➥</strong></span>
```

Ce qui donnera dans la page :

> Ceci est un texte en roman. **Ceci est un texte en gras**.

On perd, bien sûr, le fait d'indiquer la mise en gras directement dans la feuille de styles, ce qui est la fonction première des CSS. En revanche, on est sûr que cette mise en exergue sera prise en compte par les moteurs de recherche.

Pour information, il semblerait que Google, notamment, mettait en place en 2009 de nouveaux robots capables de lire, du moins en partie, les feuilles de styles. Leur prise en compte ne serait donc qu'une question de temps (mais cela fait belle lurette que cette rumeur court le Web...). Évitez donc de les remplir avec n'importe quoi et notamment de tenter de frauder par ce biais, cela pourrait se retourner rapidement contre vous. À bon entendeur…

Nombre d'occurrences des mots et indice de densité

Pendant longtemps, le nombre d'occurrences d'un mot (c'est-à-dire le nombre de fois où le mot est présent) dans la page a été très important pour les moteurs de recherche. Même si cette notion revêt encore une certaine importance, elle semble moins critique aujourd'hui. En effet, les moteurs actuels ont davantage basé leurs algorithmes de pertinence sur la notion d'indice de densité (ou *keyword density* en anglais). Peu importe le

nombre d'occurrences d'un mot dans une page, c'est sa « densité » qui est prise en compte.

En clair, pour un mot donné, son indice de densité (IDM, indice de densité d'un mot-clé) dans une page web est égal au nombre d'occurrences du mot dans la page divisé par le nombre total de mots du document.

Exemples :

– Dans une page contenant 100 mots, un terme est répété 3 fois. Son IDM est alors de 3/100 = 3 %.

– Même nombre d'occurrences dans une page de 200 mots : IDM = 3/200 = 1,5 %.

En tout état de cause, le chiffre que l'on voit le plus souvent comme limite maximale d'IDM d'un mot dans une page oscille dans une fourchette allant de 2 % à 5 %. Vous pouvez pour cela utiliser des sites spécialisés qui vous permettent de calculer automatiquement cet indice :

- Outiref – *http://www.outiref.com/*
- WebRankInfo – *http://www.webrankinfo.com/outils/indice-densite.php/*
- Keyword Density Analyzer – *http://www.keyworddensity.com/*
- Keyword Density – *http://www.ranks.nl/tools/spider.html/*
- SEO Tools – *http://www.seochat.com/seo-tools/keyword-density/*

Et bien d'autres…

Pour modifier l'IDM dans vos pages, deux solutions s'offrent à vous : soit vous proposez assez de texte autour d'un mot donné si la page est courte, soit vous répétez le mot plusieurs fois si la page est longue.

Dans le premier cas, vous pouvez proposer des pages courtes, mais très denses (tout en tenant compte du fait qu'une « bonne » page propose au moins 100 mots descriptifs).

Sachez cependant qu'il n'existe pas réellement d'IDM parfaitement idéal pour toutes les pages. De plus, en pratique, personne n'a le temps de surveiller l'IDM de tous les mots de toutes ses pages web, ce serait utopique. Vérifiez donc sur votre site l'IDM de certains mots-clés très importants pour votre activité mais n'allez pas plus loin dans ce domaine, c'est un conseil... De plus, on rentre ici dans un domaine où on commence à « écrire pour les moteurs », ce qui n'est pas obligatoirement une bonne chose. N'oubliez pas que vos contenus sont avant tout destinés à être lus par des internautes de chair et d'os... Ne vous laissez pas obnubiler par l'indice de densité des mots de votre site, vous pourriez y perdre beaucoup de temps au détriment d'autres critères plus importants.

Les différentes formes, l'éloignement et l'ordre des mots

Si vous en avez la possibilité, n'oubliez pas d'insérer dans vos pages les féminins ou les pluriels de vos mots-clés importants, ainsi que certains termes qui auraient la même

racine (poisson, poissons, poissonnerie, poissonneries, poissonnier, poissonniers, poissonnière, poissonnières, etc.). Rappelez-vous qu'une page bien positionnée sur « chien » ne le sera pas obligatoirement sur « chiens ».

Pensez donc à indiquer dans vos pages les différentes occurrences des termes susceptibles d'être saisis sur un moteur de recherche. N'oubliez pas que chaque page de votre site peut être optimisée en fonction de certains mots-clés. Ne tablez pas que sur votre page d'accueil pour être bien positionné ! On entend souvent dire, dans le domaine du référencement, qu'il est complexe de voir une page très réactive (donc bien positionnée) sur plus de 2 ou 3 mots-clés ou expressions. Ce n'est certainement pas faux.

En revanche, la casse des lettres n'a pas aujourd'hui d'importance, comme nous l'avons mentionné auparavant. Des mots ainsi orthographiés : ibm, IBM ou Ibm, seront pris en compte de la même manière par les moteurs.

Privilégiez également les couples de mots et les expressions (« chaussures de tennis », « société de service », etc.), c'est ce qui fera peut-être la différence sur des recherches plus précises émanant d'un internaute.

Par ailleurs, si vous désirez être positionné, par exemple, sur l'expression « Paris Dakar », faites en sorte que les deux mots soient présents dans la page l'un à côté de l'autre et non pas éloignés. En d'autres termes, une page contenant « Paris Dakar » l'un à côté de l'autre sera plus réactive qu'une page contenant « Paris » au début et « Dakar » à la fin.

L'ordre est également important, notamment sur Google. Une page contenant « Dakar Paris » sera moins réactive sur l'expression « Paris Dakar ». Tenez-en compte !

Une thématique unique par page

Privilégiez les pages qui traitent d'un thème unique plutôt que de longs documents qui abordent de nombreux concepts très différents. Beaucoup de moteurs tentent de faire ressortir d'un document l'idée principale qu'il contient et en tiennent compte dans leurs classements. Facilitez-leur la tâche…

Proposez sur votre site de nombreuses pages à thème unique plutôt que de longs documents traitant de plusieurs domaines différents. Ceci peut se révéler indigeste pour les moteurs… et l'internaute, d'ailleurs !

Langue du texte

Évitez également les pages bilingues ou trilingues, comme nous l'avons déjà vu précédemment pour les titres : les moteurs auront du mal à bien traiter une page contenant des termes dans plusieurs langues différentes. Privilégiez les pages monolingues.

Une fois que vous aurez effectué ce travail de base sur votre texte visible, et si vous avez bien optimisé les titres de vos pages, dites-vous bien que vous avez fait un travail majeur dans le cadre de votre stratégie de référencement. L'essentiel est – presque – fait !

Zone « Pour en savoir plus »

Nous y reviendrons dans le chapitre suivant, mais n'hésitez pas à compléter vos contenus textuels avec une zone de liens internes et externes de type « Pour en savoir plus ». Cela renforcera l'univers sémantique de votre page en indiquant aux moteurs de recherche des liens connexes parlant de la même thématique que celle abordée dans la page en question.

Un contenu en trois zones

Plus globalement, pensez vos contenus textuels en trois zones distinctes :

- **La zone 1** présente le titre (`h1`), ainsi qu'un éventuel sous-titre ou sur-titre, un fil d'Ariane, un chapô, etc. Bref, tout ce qui permet, en quelques mots-clés, d'indiquer de façon factuelle de quoi parle la page.
- **La zone 2** représente le corps textuel du contenu. C'est dans cet espace que les liens ont le plus de poids pour les moteurs.
- **La zone 3** propose des liens internes et externes sur des sujets similaires et connexes.

Ces trois zones (voir figure 4-10), sur 200 mots descriptifs au minimum, feront en sorte que votre contenu sera bien optimisé pour les moteurs de recherche (tout en étant intéressant pour l'internaute).

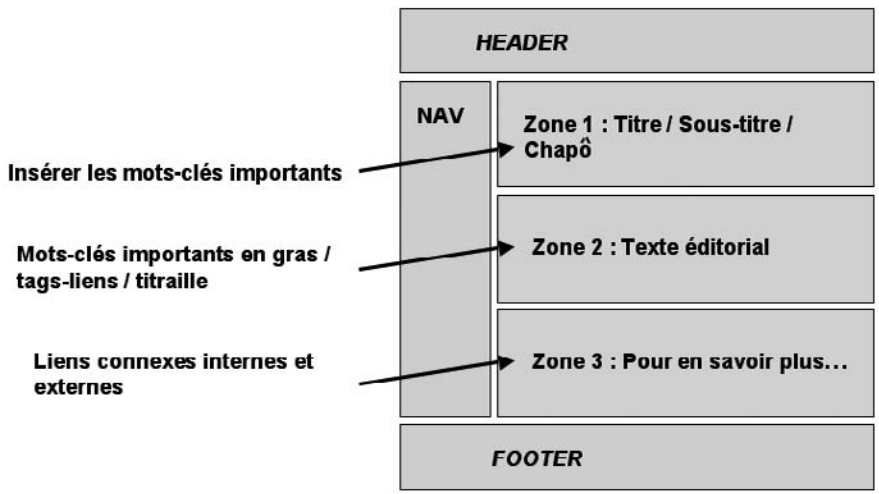


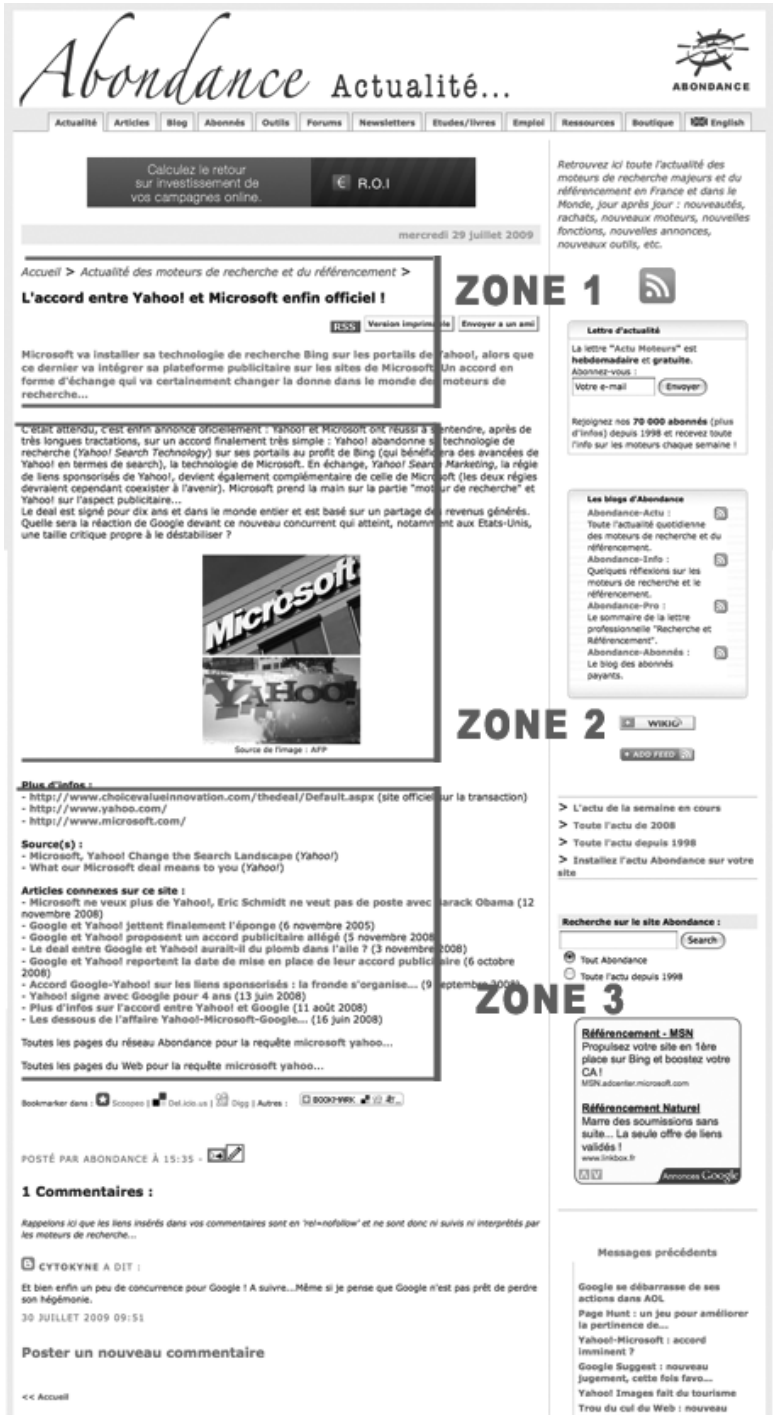
Figure 4-10

Le contenu éditorial d'une page web se subdivise en trois zones distinctes.

La figure 4-11 montre un exemple pris sur un article du site Abondance (*http://actu.abondance.com/2009/07/laccord-entre-yahoo-et-microsoft-enfin.html*).

Figure 4-11

L'article est divisé en trois zones rédactionnelles distinctes, chacune ayant son importance.



Bibliographie sur le contenu

La notion d'écriture pour le Web se développe petit à petit et de plus en plus d'articles et d'ouvrages y sont consacrés. En voici quelques-uns, souvent indispensables.

- Bien *rédiger pour le Web... et améliorer son référencement naturel* de Isabelle Canivet (un livre indispensable paru en 2009 aux éditions Eyrolles, à lire absolument !)

  http://www.action-redaction.com/couverture-bien-rediger-pour-le-web-et-ameliorer-son-referencement-naturel.htm

- *Référencement : la revanche du contenu*, Dixxit (livre blanc)

  http://www.dixxit.fr/livre-blanc-referencement/

- *Écrire pour Google : un destin de feuille morte…* de Emmanuel Parody

  http://www.zdnet.fr/blogs/ecosphere/ecrire-pour-google-un-destin-de-feuille-morte-39601073.htm

- *Créer le contenu qui plaît aux moteurs de recherche* de Emmeline Ratier

  http://www.journaldunet.com/solutions/0607/060721-referencement-ecrire-pour-moteur.shtml

- *Améliorer son référencement grâce au contenu rédactionnel* de Isabelle Canivet

  http://www.action-redaction.com/le-referencement-par-le-contenu-redactionnel.htm

- *Écrire pour le Web : quand vos lecteurs sont des moteurs* de Jean-Marie Le Ray

  http://adscriptum.blogspot.com/2006/04/ecrire-pour-le-web-quand-vos-lecteurs.html

- *Optimisation du contenu : travaillez votre text appeal* de Sébastien Billard

  http://s.billard.free.fr/referencement/index.php?2006/11/30/319-optimisation-du-contenu-travaillez-votre-text-appeal

- *Les crevettes de Madagascar* de Sébastien Bailly

  http://bailly.blogs.com/pro/2006/09/les_crevettes_d.html

- *Écrire pour le Web : la règle de G + 2H + 5W* de Joël Ronez

  http://blog.ronez.net/?p=354

- *Comment écrire des titres qui tuent ?* de Jan

  http://bvwg.actulab.net/6-ecrire-des-titres.seo

- *Rédiger pour être référencé… et lu* de François La Roche

  http://www.bloguemarketinginteractif.com/rediger-pour-etre-reference-et-lu/

Pour résumer

Voici quelques conseils pour bien optimiser le texte visible de vos pages.

- Soignez au maximum le contenu du premier paragraphe de vos pages (les 2 ou 3 premières phrases).
- Créez des pages à thème unique, monolingues, mais contenant au moins 100 à 200 mots descriptifs.
- Mettez en exergue les mots-clés importants pour votre activité : gras, titre (balise `<hn>`), texte des liens, etc.
- Prenez en compte les différentes formes des mots : féminin/masculin, singulier/pluriel, etc.

– Prenez en compte la proximité des mots entre eux ainsi que leur ordre.
– Ne recherchez pas obligatoirement à truffer vos pages de mots-clés identiques, cela n'aura que peu d'effet sur le poids de vos pages.
– Restez sur un indice de densité oscillant entre 2 et 5 % pour vos mots-clés importants.
– N'oubliez pas que chaque page de votre site peut être optimisée pour un mot-clé ou une expression. Ne misez pas tout sur la seule page d'accueil.
– Une page web aura du mal à être réactive à plus de 2 ou 3 mots-clés ou expressions. Jouez sur le nombre de pages optimisées plutôt que sur le nombre de termes dans une même page.
– Concevez vos pages en trois zones : une pour le titre et la description, une pour le texte lui-même et une dernière sous la forme d'un « Pour en savoir plus ».

Zone chaude 3 : adresse (URL) des pages

Après le texte et le titre des pages, nous allons maintenant explorer leur URL ou adresse web, du type :

http://www.votresite.com/produits/gamme/article.html/

Premier point très important : il est essentiel d'avoir son propre nom de domaine (votre-site.com ou votre-entreprise.fr) pour imaginer obtenir un bon référencement pour son site web. Des adresses comme *perso.votre-fournisseur-d-acces.com/votre-entreprise/* ou autres sont parfaites, dans un premier temps, pour tester le réseau sans réelle stratégie d'entreprise et créer un site web « pour voir », mais si vous avez une quelconque ambition sur Internet, achetez au plus vite votre nom de domaine, le prix (entre 10 et 20 euros par an, parfois moins) en vaut vraiment la chandelle. N'hésitez pas une seconde, la différence est radicale pour le référencement, vous vous en apercevrez bien vite.

La question du choix du nom de domaine d'un site (votresite.fr, votresite.com…) est très souvent l'objet de multiples discussions, réflexions et décisions qui ressemblent parfois à des compromis pas toujours très heureux. Que faut-il faire ? Choisir un *.fr* ou un *.com* ? Un nom composé avec ou sans tiret ? etc.

Cela n'est nullement un secret des Dieux : le nom de domaine d'un site est important pour son référencement. Il doit, idéalement, contenir un ou plusieurs mots décrivant au mieux ce qu'il propose dans ses pages : nom de l'entreprise, activité principale, etc. La présence d'un mot-clé de recherche dans le nom de domaine d'un site est bien souvent un critère déterminant pour son classement.

Pour en avoir le cœur net, tapez le mot-clé « référencement » sur Google. Voici quelques-uns des noms de domaine affichés sur les deux premières pages :

- *www.referencement-2000.com/*
- *www.referencement-team.com/*
- *www.referencement-fr.com/*
- *www.referencement.tv/*
- *www.forum-referencement.net/*

- *www.referencement-gratuit.net/*
- *www.referencement-gratuit.com/*

Cela se passe de commentaires. Autre exemple : tapez des mots-clés génériques comme « auto » ou « finance » et vous verrez que, très souvent, ce terme se retrouve dans le nom de domaine sur le moteur de recherche leader et notamment dans les premières positions. Le phénomène se répète sur de nombreux moteurs.

Sur Google, tapez « livre référencement » ou « baromètre référencement » et vous comprendrez pourquoi nous avons choisi ainsi les noms de domaine de ces sites du Réseau Abondance. Ils se sont placés premier sur Google en quelques semaines malgré une concurrence importante, ce qui montre bien l'importance du nom de domaine dans les algorithmes actuels des moteurs. Ce nom de domaine est également important pour l'affichage de Sitelinks sur Google (voir chapitre 2).

Bien sûr, l'intitulé du nom de domaine ne suffit pas. Une optimisation complète du site est nécessaire (titre, texte, etc.) mais le nom de domaine jouera un rôle souvent complémentaire dans vos positionnements. Il n'est pas essentiel, cependant, comme nous le verrons, mais il peut apporter un plus non négligeable.

Le problème principal du nom de domaine vient du fait que celui-ci doit rester court, lisible et mnémotechnique. Par conséquent, vous aurez le choix entre « nom-de-votre-entreprise.com » ou « caracterisation-de-votre-activité.com » et c'est tout. En d'autres termes, soit « stela.com » soit « chaussures-de-tennis.com » (si Stela est le nom d'une société fictive qui vend ce type de produits). C'est un peu court et vous n'avez pas vraiment le choix de placer beaucoup de mots-clés dans cette zone. Autant les choisir au mieux donc.

Sachez enfin que toute stratégie reposant sur des « galaxies de noms de domaines », visant à acheter de nombreux noms de domaines différents et contenant chacun des mots-clés pertinents pour votre activité, est assimilée à du spam et est absolument à proscrire en 2009. Pénalité et/ou liste noire assurée ! Rien ne vous empêche d'acheter plusieurs noms de domaines, notamment pour éviter qu'un petit malin ne vous les « pique », et de tous les rediriger vers une adresse « canonique », mais il n'est absolument pas recommandé de bâtir une stratégie de référencement sur cette base. Nuance... Nous reviendrons sur ce point très bientôt au sein de ce chapitre.

Quel domaine choisir ?

Première interrogation bien souvent posée : faut-il choisir un « *.fr* » ou un « *.com* », voire toute autre extension (*.org*, *.eu*, *.info*, *.biz*, etc.) pour son site ? Question épineuse en soi mais dont la réponse ne dépend pas réellement de la façon dont le moteur de recherche le prendra en compte. En effet, il n'existe aujourd'hui aucune preuve que le domaine choisi influe d'une quelconque façon sur votre futur positionnement dans les pages de résultats des moteurs. En d'autres termes, que vous optiez pour un *.fr*, un *.info*, un *.biz* ou un *.com*, cela ne devrait pas jouer sur vos futurs positionnements, les moteurs n'en tenant pas compte, jusqu'à preuve du contraire.

Le choix du domaine sera donc plutôt issu d'une réflexion sur le site lui-même et sa cible. Par exemple :

- un site ayant une cible française pourra, de façon indifférente, être disponible sur un *.fr* ou un *.com* (ou autre) ;
- un site ayant une cible américaine optera plutôt pour un *.com* ;
- un site à vocation internationale pourra de façon habile être accessible selon plusieurs adresses : le *.fr* pour la version française, le *.com* pour la version en langue anglaise, etc. ;
- une association pourra sans soucis opter pour le *.org*.

À une époque, certains moteurs (Excite notamment) prenaient en compte, pour une recherche sur le Web francophone, uniquement les pages issues de sites en *.fr*. Cette époque est aujourd'hui révolue (heureusement !) et une recherche sur le Web francophone, par exemple sur Google, s'effectuera prioritairement sur la langue utilisée dans la page web et ne se basera pas sur le domaine du site.

Un bémol sera cependant appliqué à cette réflexion au paragraphe suivant.

L'hébergement est-il important ?

L'hébergement de votre site peut avoir son importance, et notamment la localisation géographique de l'hébergeur choisi. En effet, sur la plupart des moteurs (Google, Yahoo!, Bing), trois boutons de recherche sont disponibles (voir figure 4-12) : « Web », « Pages francophone » et « Pages : France » (appellation de Google, légèrement différente sur Yahoo! et Bing).

Figure 4-12

Page d'accueil de Google France

Si les deux premiers choix sont simples (« Web » = recherche sur la totalité de l'index, « Pages francophones » = pages écrites en langue française), le choix « Pages : France » restreint la recherche sur :

- soit les pages accessibles sur un site en *.fr* ;
- soit les pages accessibles sur un serveur hébergé sur le territoire français.

Cette restriction peut sembler peu importante sur la France (qui utilise réellement la fonction « Pages : France » dans l'Hexagone ?). Elle l'est cependant bien plus en Suisse ou en Belgique par exemple, où de nombreux internautes effectuent des recherches spécifiquement sur leur territoire.

Ainsi, si la cible suisse, par exemple, est importante pour vous et que votre site est hébergé, disons, sur le territoire français ou américain, vous n'aurez pas d'autre choix que d'utiliser le domaine .ch pour une version helvétique de votre site, si vous désirez apparaître dans l'option « Pages : Suisse » du site de Google (figure 4-13).

Figure 4-13

Page d'accueil de Google Suisse

Dans ce cas, c'est donc l'hébergeur et sa localisation géographique qui induiront le choix le plus judicieux du nom de domaine de votre site.

En revanche, le choix de l'hébergeur ne semble pas devoir causer de soucis de référencement, en dehors de la problématique de localisation géographique et pour ce qui est du strict point de vue du nom de domaine.

Vous pouvez avantageusement utiliser des outils comme Whois.sc (*http://www.whois.sc/*) qui vous indiqueront dans quel pays est situé votre hébergeur sur la base de l'adresse de son site ou du numéro IP de l'un de ses serveurs.

Il faut également tenir du compte du fait qu'il semblerait que Google prenne en considération le fait que certains sites, notamment s'ils échangent des liens, soient hébergés sur une adresse IP proche (même classe C), ce qui réduirait le « poids » de ces liens. En clair, Google penserait dans ce cas que les sites sont sur le même serveur, ou chez le même hébergeur, et qu'il y a des chances qu'ils appartiennent à la même entité. Aucune preuve formelle n'est venue confirmer ou infirmer ce point pour l'instant. Difficile de toute façon, si on gère une dizaine de sites web, de les transférer chacun chez un hébergeur, voire sur un serveur, différent...

L'ancienneté du domaine est-elle importante ?

L'ancienneté du nom de domaine semble clairement importante et notamment pour Google. On peut penser que si ce moteur a mis en place la démarche de s'enregistrer en tant que gestionnaire de noms de domaine (*registrar* en anglais) (*http://actu.abondance.com/2005-05/google-registrar.php/*), c'est certainement pour avoir un accès plus direct à un certain nombre d'informations disponibles dans les bases Whois des DNS (*Domain Name Systems* ou *Domain Name Servers*).

Si vous disposez de plusieurs noms de domaine, privilégiez, toutes choses étant égales par ailleurs, le nom de domaine que vous avez déposé à la date la plus ancienne. Il semblerait que Google accorde plus de confiance à votre site si le nom de domaine de ce dernier est ancien, d'où la notion de *TrustRank* (dont nous reparlerons au chapitre suivant) pour désigner certains critères pris en compte par le moteur de recherche. Certains disent même que les dates de renouvellement jouent également un rôle : un nom de domaine renouvelé par exemple tous les 5 ans serait une preuve de confiance plus grande qu'un domaine renouvelé tous les ans, suscitant une méfiance du moteur portant sur des opérations à courte échéance.

Noms composés : avec ou sans tirets ?

Question fréquemment posée : si votre société s'appelle « Matelas Bon Sommeil », faut-il acheter « matelasbonsommeil.com » ou « matelas-bon-sommeil.com » ?

Ici, la question est simple en théorie : le nom de domaine contenant les mots séparés par un tiret est à privilégier : matelas-bon-sommeil.com, puisque dans ce cas, les tirets séparant les différents mots, le site sera plus réactif pour le moteur sur des requêtes comme « matelas », « matelas sommeil », « bon sommeil » ou « matelas bon sommeil ». Dans le premier cas, les termes n'étant pas séparés, le moteur ne comprendra que le mot-clé « matelasbonsommeil ».

Cependant, une autre question doit se poser : sur quel nom de domaine faut-il communiquer lorsqu'on parle de son site, en « offline » (cartes de visites, publicité papier, papier à en-tête, etc.) ou « online » (référencement, liens, etc.). En tout état de cause, nous verrons bientôt qu'il est bon de ne jouer que sur un seul nom de domaine pour la communication de façon globale. Ce sera donc à vous de le choisir en fonction d'un certain nombre de critères.

- **La cible.** Une cible professionnelle, technophile, pourra ne pas être dérangée par la présence du tiret et cette version pourra être privilégiée. Une cible « grand public » sera peut-être gênée par le tiret et la version en un mot pourra éventuellement être privilégiée. Voici un exemple simple : si l'adresse du site doit être énoncée à la radio ou à la télévision, faut-il parler de « tiret », de « trait d'union » ?
- **La préférence de la promotion.** Si la visibilité sur les moteurs de recherche est essentielle dans votre stratégie, préférez la version avec tiret qui sépare bien les termes et permet au moteur de les prendre en compte.

En tout état de cause, la version avec tirets est préférable pour les moteurs de recherche, c'est tout à fait clair. Mais si vous préférez la version sans tiret, sachez que cela n'est pas rédhibitoire. Vous pourrez tout à fait compenser ce problème en jouant, par exemple, sur des URL « bien conçues » pour y insérer des mots-clés. Exemple :

http://www.matelasbonsommeil.com/matelas/bon-sommeil/gamme/prix-reduits/promotions.html/

Ce type d'URL, très optimisée pour les moteurs de recherche, peut tout à fait compenser l'absence des tirets séparant les mots dans le nom de domaine. Il est faux de penser que le fait d'utiliser des noms de domaine en un mot est fortement pénalisant pour le référencement. Utiliser les tirets est un plus mais l'utilisation d'URL et de pages optimisées peut tout à fait compenser ce fait, du moins en grande partie. La tendance, en 2009, semblait d'ailleurs clairement aller vers les noms de domaines sans tiret...

Faut-il utiliser le nom de la société ou un nom contenant des mots-clés plus précis comme nom de domaine ?

Si votre société s'appelle « Tartempion » et qu'elle vend des matelas en mousse, que devez-vous acheter comme nom de domaine : « tartempion.com » ou « matelas-mousse.com » ? Là encore, il n'existe pas de réponse gravée dans le marbre.

On serait quand même tenté de conseiller d'acheter « tartempion.com » car, stratégiquement parlant, il est plus logique d'acheter son nom de marque comme nom de domaine et donc de communiquer sur celui-ci. Le domaine « matelas-mousse.com » est certainement plus optimisé pour les moteurs de recherche (il contient les deux mots les plus importants pour votre activité) mais il n'est optimisé *que* pour les moteurs de recherche. C'est un peu juste *a priori* pour étendre cette vocation à toute votre communication.

Là encore, on revient au cas précédent : vous pouvez toujours communiquer sur votre nom d'entreprise en jouant sur des URL optimisées comme :

http://tartempion.com/matelas-mousse/gamme/promotions.html/

Cette solution devrait être efficace et présente l'avantage de refléter une certaine logique dans le cadre d'une stratégie de communication globale sur le Web.

Faut-il baser une stratégie de référencement sur plusieurs noms de domaine pointant vers un même site ?

Cette question découle des deux précédentes. On peut penser qu'il est nécessaire aujourd'hui d'être assez catégorique sur ce point. **Il est important de ne communiquer que sur un seul nom de domaine** pour :

- la communication online : liens vers votre site (amélioration du PageRank), citations dans les articles, liens internes de votre site, etc ;
- la communication offline : papier à en-tête, cartes de visites, posters, publicité papier, PLV, etc.

En tout état de cause, il semble clair qu'un seul nom de domaine est à privilégier pour ne pas brouiller les esprits de vos clients et prospects, futurs visiteurs de votre site. De plus, la présence étrange de plusieurs noms de domaine pointant tous vers votre site est à déconseiller, les moteurs de recherche pouvant détecter des tentatives de spam.

Entendons-nous bien : rien ne vous interdit d'acheter par exemple les versions *.com*, *.fr*, *.net* et *.info* de votre nom de domaine pour éviter que quelqu'un ne les réserve à votre place (le site Abondance, par exemple, dispose des noms de domaine en *.com*, *.net* et *.fr* qui redirigent tous vers le .com au niveau du DNS). En revanche, nous déconseillons fortement de mettre en place une stratégie de référencement basée sur un nombre important de noms de domaine pour une même source d'information, pointant donc sur une même page d'accueil. Ce type de tactique de référencement peut fonctionner à court terme mais est extrêmement dangereuse à moyen et long terme, la détection de spam par les moteurs étant quasi certaine dans les mois qui viennent sur ce type de méthodes.

En tout état de cause, il existe des milliers de sites web très bien référencés sur des mots-clés très concurrentiels tout en ne proposant pas ces termes dans leur nom de domaine. Ce champ est certes important, mais il n'est pas primordial aujourd'hui dans les algorithmes de pertinence des moteurs. Si votre stratégie globale de communication rejoint les contraintes des moteurs, tant mieux. Mais si ce n'est pas le cas, ce n'est pas d'une gravité rédhibitoire. En résumé, choisissez le plus logiquement possible votre nom de domaine et privilégiez le contenu textuel de vos pages : votre référencement ne s'en portera que mieux.

Nous irons même plus loin : si vous avez, par exemple au démarrage d'un projet, le loisir de choisir votre nom de domaine en partant de zéro, essayez de l'optimiser au mieux : nom de votre entreprise ou termes décrivant votre activité. Si ce n'est pas le cas, si votre nom de domaine est déjà connu sur le Web, si des liens ont déjà été créés vers lui par d'autres sites, laissez tomber et gardez la situation actuelle. Il y a bien d'autres points à optimiser pour obtenir une bonne visibilité et vous risqueriez de « casser » pas mal de choses en tentant une stratégie basée sur les noms de domaine.

Des mini-sites valent mieux qu'un grand portail

La solution certainement la plus efficace mais pas la plus simple à mettre en œuvre si votre site existe déjà, est de créer des « mini-sites » plutôt qu'un grand portail. C'est l'option que nous avons prise en créant le « Réseau Abondance » qui regroupe plus d'une vingtaine de sites, chacun ayant son contenu propre, identifié (l'actualité pour abondance.com, l'audit de site avec outiref.com, les forums avec forums-abondance.com, les jeux avec googlefight.com, etc.). Plusieurs avantages à cela.

- Cela évite de créer une « usine à gaz » présentant parfois trop d'informations par rapport à ce que recherche l'internaute. Ce dernier peut aller directement sur le site qui lui convient et y trouver l'information en quelques clics.

- Chaque site peut avoir sa propre cible, son contenu adéquat, sa propre charte graphique, son propre modèle économique… bref, sa propre vie, indépendamment l'un de l'autre.
- Cela multiplie la visibilité du « réseau » dans les résultats des moteurs de recherche.
- Cela renforce le PageRank (popularité) de chacun des sites du réseau en multipliant les liens de l'un vers l'autre et donc l'interconnexion des pages, toujours importante pour les moteurs de recherche.
- Cela permet d'accélérer la prise en compte des nouveaux sites du réseau en jouant sur les liens croisés (voir chapitre 8).

Bien sûr, si votre site est déjà créé, cela peut poser problème car il vous faudrait, pour créer un réseau, refaire bon nombre de choses et rebâtir votre stratégie. Mais cela est peut-être envisageable lors d'un remaniement du site ou de la mise en place d'une nouvelle version ?

Les sous-domaines

Autre solution pour augmenter votre visibilité : créer des sous-domaines pour certaines zones de votre site web. Tapez le mot-clé « abondance » sur Google. Vous verrez apparaître les sites suivants dans les deux premières pages de résultats sur la figure 4-14 :

- *www.abondance.com/*
- *outils.abondance.com/*
- *actu.abondance.com/*
- *blog.abondance.com/*
- *methodologies.abondance.com/*
- *docs.abondance.com/*

Figure 4-14

Les sous-domaines (ici outils.abondance.com, blog.abondance.com, docs.abondance.com et actu.abondance.com), s'ils sont utilisés avec parcimonie et bon sens, permettent de démultiplier la visibilité d'un site web dans les résultats des moteurs.

Quelles conclusions pouvons-nous en tirer ?

- La plupart des moteurs de recherche majeurs actuels pratiquent le *clustering* : pour un site donné, ils affichent au maximum deux pages web pertinentes. Pour les autres, un lien « Pages similaires » est, parfois, proposé comme le montre la figure 4-15.

Figure 4-15

Clustering par Google : 2 pages au maximum d'un même site sont présentées dans les résultats, la deuxième étant décalée vers la droite.

- Pour les moteurs de recherche, les adresses *actu.site.com*, *site.com*, *www.site.com*, *info.site.com* représentent 4 sources d'informations, soit 4 sites différents.

> **La création des sous-domaines**
>
> Notez bien ici qu'il existe de nombreuses façons de créer, techniquement parlant, des sous-domaines. Certains hébergeurs proposent même une interface spécifique dans leurs outils d'administration à cette fin. Le mieux est de contacter votre hébergeur pour en savoir plus.

Dans ce cadre, vous pouvez tout à fait créer des « sous-domaines » comme, *actu.votre-site.com*, qui pointe par exemple vers l'adresse *www.votresite.com/actu/*.

Cela présente, là aussi, plusieurs avantages.

- L'adresse est plus rapide à taper pour l'internaute et plus mnémotechnique (c'est l'un des avantages essentiels des sous-domaines). Il vaut mieux saisir *http://referencement.abondance.com/* plutôt que *http://www.abondance.com/docs/ref/index.php/*.
- Votre visibilité est multipliée sur les outils de recherche.
- Votre visibilité est accrue dans les pages de résultats des moteurs.
- Les liens entre vos différents sous-domaines sont considérés comme externes par Google (de site à site), ce qui renforce leur efficacité (voir chapitre 5).
- Cela n'est pas considéré comme du spam par les moteurs à partir du moment où :
  - les différents sous-domaines créés pointent vers des pages différentes ;
  - vous gardez raison en ne créant pas des centaines, voire plus, de sous-domaines différents. Le mieux est de créer un sous-domaine par rubrique de votre site, dans les limites que nous allons évoquer ci-après.

À partir du moment où cette stratégie est mise en place avec bon sens, elle peut réellement considérablement augmenter votre visibilité sur les moteurs de recherche. Mais retenez bien que cette possibilité n'est acceptable que si elle est utilisée avec parcimonie

et loyauté. La première motivation pour créer un sous-domaine doit toujours être de fournir un raccourci à l'internaute pour aller plus rapidement à l'information présente sur une page interne d'un site. D'ailleurs, Google lui-même les utilise (*labs.google.com*, *news.google.com*, *images.google.com*, etc.). Mais si cela devient une stratégie de référencement plus ou moins « tordue », cela devient moins évident tout en se rapprochant du spam. Avec un marteau, on peut taper sur un clou ou sur la tête de son voisin… mais il ne faut pas faire, dans ce dernier cas, le procès du marteau. Évitez tout abus en privilégiant, là encore, le bon sens lors de la création de vos sous-domaines. Rappelons également que la règle valable pour les domaines l'est également pour les sous-domaines : créer de nombreux sous-domaines différents pointant vers une même page web est considéré comme du spam.

Attention cependant, chaque médaille a son revers. En effet, il semblerait que le nombre de pages d'un site web soit également un critère que prend en compte Google dans sa recherche de la meilleure pertinence d'un document. Si une page web appartient à un « gros » site proposant de nombreuses pages, Google aurait plus confiance en elle, et classerait donc mieux cette page. Ce qui signifie que si vous avez un site qui contient, par exemple, 1 000 pages et que vous l'éclatez en 10 sous-domaines comportant chacun 100 pages, cette manœuvre provoquerait une « perte de confiance » du côté des moteurs de recherche puisque vous portez à leur connaissance 10 petits sites à la place d'un gros... Là encore, aucune preuve réelle de ce phénomène n'a été portée à notre connaissance, mais le référencement n'est pas une science exacte... Ceci dit, on pourra retenir comme règle de ne pas créer de sous-domaines générant de trop « petits » sites web, de quelques dizaines de pages. Ne créez un sous-domaine que si son contenu représente un volume assez conséquent pour être pris en compte comme une source d'informations à part entière par les moteurs de recherche.

Les intitulés d'URL

Les noms de domaine et sous-domaines ne sont pas les seules zones importantes dans l'adresse de vos pages web. Tout l'intitulé peut avoir une importance. Utilisez donc des termes clairs et précis plutôt que des abréviations, des chiffres ou des signes cabalistiques que vous seriez seul à comprendre.

Une adresse telle que :

http://www.stela.com/produits/stylos/recharges/acheter.html/

propose cinq mots-clés intéressants : « stela », « produits », « stylos », « recharges » et « acheter ». C'est loin d'être négligeable. Et c'est toujours mieux que :

http://www.societestela.com/prods/sty-fr/PK470012/pricing.html/

Insérer, par exemple, la référence catalogue d'un produit dans l'URL peut être intéressant pour la maintenance des pages, mais moins pour l'internaute. Or, n'est-ce pas pour lui que vous avez créé votre site ? Et les moteurs n'y trouveront pas non plus de grain à moudre.

Utilisez également le tiret (-) pour séparer les mots plutôt que l'underscore (_) car ce caractère ne représente pas un séparateur pour les moteurs en général et Google en particulier (bien que cette donnée ait évolué notamment chez Google qui est devenu beaucoup plus souple sur ce point, mais il n'a toutefois jamais communiqué officiellement – tout du moins à la mi-2009 – sur le fait que ce caractère ne lui posait plus de problèmes).

Ainsi, l'adresse ci-dessous est valide et optimisée pour les moteurs de recherche :

http://www.stela.com/produits-papeterie/stylos-encre/recharges/encre-noire.html/

mais pas celle-ci :

http://www.stela.com/produits_papeterie/stylos_encre/recharges/encre_noire.html/

En effet, dans le second cas, le moteur risque de comprendre ainsi les mots composés : « produitspapeterie », « stylosencre » et « encrenoire ». Alors que le tiret sera, quant à lui, remplacé par un espace. Google semblait avoir réparé à la mi-2007 ce bug (*http://actu.abondance.com/2007/07/google-prend-enfin-lunderscore-comme.html/*) avant de revenir sur cette déclaration sans être tout à fait clair à ce sujet. Prudence donc...

En règle générale, utilisez également des lettres et des chiffres simples, et bannissez les caractères tels que « * », « + » ou « ? » de vos URL : certains moteurs ne les acceptent pas (nous en reparlerons au chapitre suivant).

Vous pouvez éventuellement, pour faire plus simple, saisir toutes vos URL en minuscules, cela ne pénalisera pas vraiment votre référencement (notez bien que cela ne l'améliorera pas non plus) mais ce sera plus lisible pour l'internaute. Rappelons en effet que la casse des lettres, si elle n'a pas d'importance dans le nom de domaine, est discriminante pour un navigateur dans le reste de l'intitulé de l'adresse : `INDEX.HTML` est différent de `Index.html` et d'`index.html`. En revanche, elle n'a pas d'importance pour les moteurs de recherche.

Bien sûr, il n'est pas toujours possible d'insérer des mots-clés importants dans les URL des pages. Selon les systèmes d'édition utilisés (pages dynamiques, gestion de contenu, etc.), les URL peuvent être plus ou moins absconses, complexes, sans que vous ayez la main sur leur intitulé. Peut-être sera-t-il intéressant, dans ce cas, de passer par des systèmes d'*URL rewriting* (voir chapitre 7).

Idéalement, recopiez également le contenu de votre titre éditorial (en `<h1>` dans votre code) à la fin de votre URL. Prenons un exemple avec l'article *Google Suggest condamné* paru en juillet 2009 sur le site Abondance.

1. Son titre éditorial est contenu dans une balise `<h1>` :

```
<h1>Google Suggest condamné</h1>
```

2. L'URL reprend ce titre dans sa dernière partie (le « é » de « condamné » étant remplacé par un « e ») :

   http://actu.abondance.com/2009/07/google-suggest-condamne.html

3. Et la balise `<title>` commence par ce même titre :

```
<title>Google Suggest condamné - Abondance : R&eacute;f&eacute;rencement et moteurs
➥de recherche</title>
```

Résultat : l'article s'est retrouvé en première page de Google en quelques heures pour la requête « Google Suggest » qui génère pourtant 91 200 000 résultats (voir figure 4-16)... Même si ce positionnement n'est pas dû qu'à ces trois critères uniquement, ils ont fortement joué sur le résultat obtenu.

Figure 4-16

Un positionnement obtenu en quelques heures...

L'« œil du référenceur » se posera d'ailleurs souvent sur ces trois critères au minimum pour voir si l'optimisation d'une page web est bâtie sur de bonnes fondations :

- Le titre éditorial de la page est-il descriptif et inséré dans une balise `<h1>` ?
- Ce titre termine-t-il l'URL de la page ?
- Ce titre débute-t-il l'intitulé de la balise `<title>` de la page ?

Si la réponse est « oui » à chacune de ces trois questions, la page part sur de bonnes bases. Pas toujours suffisantes, certes, mais hautement nécessaires pour obtenir de bons résultats en référencement...

Autres points à prendre en compte :

- Pas de lettres accentuées ou de caractères diacritiques dans les URL. Ainsi, un « é » sera remplacé par un « e », un « ç » par un « c », etc.

- Idem pour la ponctuation : remplacer les apostrophes, les guillemets, etc., par un tiret.
- Éviter également d'insérer trop de caractères « / » dans vos URL.
- Notez que la terminaison de l'adresse (.html, .asp, .php, etc.) n'a aucune importance pour votre référencement.
- Enfin, sachez que, pour voir vos contenus indexés dans Google Actualités (Google News – *http://news.google.fr/*), vos URL devront également contenir au moins trois chiffres. Par exemple, *http://www.stela.com/actualites-papeterie/clairefontaine-sort-un-nouveau-stylo-123456.html/*.

  Alors que cette URL serait refusée : *http://www.stela.com/actualites-papeterie/clairefontaine-sort-un-nouveau-stylo.html/*.

  Cette restriction, un peu ridicule, est clairement demandée par Google pour voir un article indexé dans son agrégateur d'actualité. Cependant, à la mi-2009, un bruit venu de Grande-Bretagne semblait signaler, qu'à moyen terme, Google pourrait supprimer cette contrainte. Mais rien d'officiel au moment où ces lignes sont écrites.

Voici donc quelques exemples d'URL parfaitement optimisées pour les moteurs de recherche (et Google News) :

- *http://www.stela.com/actualites-papeterie-123456-clairefontaine-cree-un-nouveau-stylo.html/*
- *http://www.matelas-mousse.fr/12-09-2009-notre-societe-a-recu-le-prix-de-l-ingeniosite-francaise-en-2009/*
- *http://tempsreel.nouvelobs.com/actualites-societe-20090716-l-unefaculte-denonce-des-frais-d-inscription-illegaux.html*

Pour résumer

Voici quelques conseils pour bien optimiser les URL de vos pages :

– Achetez un nom de domaine (*.com*, *.fr*, *.net*…) propre, sans système de redirection.

– Insérez, si possible, un ou deux mots importants pour votre activité dans le nom de domaine : votre nom, votre activité, etc.

– Ne fonctionnez que sur un seul nom de domaine pour vos promotions online et offline.

– Essayez de mettre en place un réseau de « petits sites » plutôt qu'un gros portail.

– Insérez des mots-clés importants et intelligibles dans l'intitulé complet de vos URL.

– Reprenez le titre `h1` de votre page à la fin de votre URL.

– Séparez les mots importants par des tirets dans les énoncés de vos noms de domaine.

– Remplacez les caractères accentués et diacritiques de vos URL par des équivalents non accentués.

– Insérez au moins trois chiffres (par exemple, la date de parution) dans l URL si vous voulez que vos contenus soient indexés par Google News.

– Créez des sous-domaines (*motclé.votresite.com*) pour faciliter la tâche des internautes et augmenter votre visibilité dans les résultats des moteurs.

– Le plus important : agissez toujours avec loyauté et évitez tout spam, jamais payant à court, moyen ou long terme, sur les moteurs !

Zone chaude 4 : balises meta

En HTML, les balises meta permettent de fournir aux moteurs de recherche un certain nombre d'informations sur le contenu d'une page web. « meta » est une abréviation de « metadata », ces balises signalent donc « de l'information sur l'information ».

Moins d'importance aujourd'hui

Il est important de noter que les balises meta ont aujourd'hui moins d'importance pour les moteurs de recherche qu'il y a quelques années en termes de critère de pertinence et de positionnement. Trop de spam a été réalisé au travers de cette zone et, petit à petit, les moteurs se sont lassés de prendre en considération des informations qui auraient dû être très pertinentes et qui tournaient, *in fine*, au « réservoir à spam »… Ceci dit, la présence de balises meta ne pénalise pas obligatoirement vos pages, mais on peut aisément dire aujourd'hui que l'aide apportée au positionnement est minime sur certains moteurs, voire nulle sur Google… Ce n'est pas une raison pour ne pas en parler, mais certainement d'une façon beaucoup plus rapide que si cet ouvrage avait été écrit il y a quelques années de cela.

Les balises meta permettent d'ajouter une description de la page affichée, ainsi que des mots-clés spécifiques, de façon transparente, à l'attention des moteurs. Elles ne garantissent pas cependant que les pages qui les contiennent obtiendront obligatoirement un meilleur classement que d'autres. La meilleure garantie d'un bon classement reste aujourd'hui la présence de mots-clés pertinents dans le titre de la page, dans le texte visible, dans l'URL, etc.

Les deux balises meta prises en compte par les moteurs (`name="description"` et `name="keywords"`) doivent être placées après la balise `<title>…</title>` et avant la balise de fin d'en-tête (`</head>`) comme ici :

```
<html>
  <head>
       <title>Titre de la page</title>
       <meta name="description" content="contenu de la balise description" />
       <meta name="keywords" content="contenu de la balise keywords" />
```

Si votre éditeur HTML les place avant la balise `<title>`, remettez cette dernière en première position, comme nous l'avons vu au début de ce chapitre.

Imaginons que vous réalisiez un site relatif à votre société, nommée Stela, et dont l'activité consiste à vendre des chaussures de sport. Sur la page d'accueil, vous indiquerez, par exemple, dans le code HTML les lignes suivantes :

```
<meta name="description" content="Stela, sp&eacute;cialiste de la vente de chaussures
➥de sport, bas&eacute; &agrave; Paris, France" />
<meta name="keywords" content="stela, chaussures de sport, tennis, running, footing,
➥stretching, chaussure, terre battue, dur, herbe, wimbledon, flushing meadow,
➥roland garros, flinders park, grand chelem" />
```

Notez ici l'accentuation en HTML des caractères accentués : `&eacute;` est le code de « é », `&agrave;` signifie « à ».

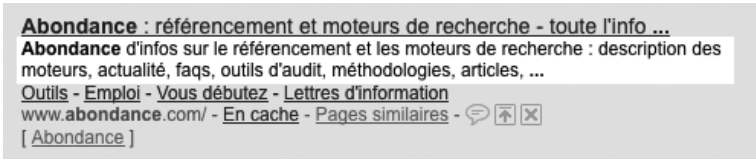
Nous allons maintenant étudier plus en détail ces deux balises.

Balise meta description : à ne pas négliger pour mieux présenter vos pages !

La balise meta `description` permet d'indiquer au moteur de recherche une phrase de résumé du contenu de la page. Cette description sera affichée par certains moteurs dans leur page de résultats, sous le titre. Exemple sur Google (mot-clé « abondance ») en figure 4-17.

Figure 4-17

Google reprend le contenu de la balise meta description pour présenter la page dans ses résultats.



Si la page ne contient pas de balise meta `description`, ou si elle est trop courte, ou si le moteur décide de ne pas l'afficher, un snippet (extrait textuel de la page contenant le terme demandé) sera indiqué comme le montre la figure 4-18.

Figure 4-18

Google peut également afficher un extrait textuel de la page contenant la requête.

Dans cet exemple, le moteur de recherche n'a pas trouvé de balise meta `description` dans le code source de la page. Il a donc créé un snippet, extrait textuel de la page contenant le mot demandé.

Un autre exemple est visible en figure 4-19.

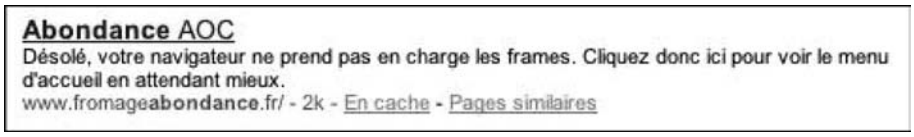


Figure 4-19

Pas facile de savoir de quoi parle cette page…

Ici, le moteur n'a trouvé ni balise meta `description` assez descriptive (elle existe pourtant), ni texte visible (mauvaise optimisation des frames). Le résultat est peu parlant.

L'algorithme d'affichage du résumé textuel de chaque résultat proposé par Google est le suivant :

1. Tout d'abord, il utilise trois sources différentes et possibles pour ce texte : le contenu de la balise meta `description`, un extrait textuel de la page ou la description de l'annuaire Open Directory (*http://www.dmoz.org/*) si le site est inscrit sur cet outil.
2. Tout d'abord, le moteur va privilégier la balise meta `description` et va chercher ce contenu dans le code de la page. Si cette balise n'existe pas ou si elle est vide, il va passer à l'étape 6.
3. Si le contenu de la balise meta `description` existe mais qu'il est trop court (moins de 100 caractères environ), il va passer à l'étape 6.
4. Si le contenu de la balise meta `description` n'est pas cohérent avec le contenu de la page (par exemple, même contenu sur toutes les pages du site), il va passer à l'étape 6.
5. Si le contenu de la balise meta `description` est assez long (supérieur à 100/150 caractères, les moteurs actuels n'en affichent que rarement plus) et qu'il est cohérent par rapport au contenu de la page, c'est ce texte que Google utilisera comme résumé textuel (snippet).
6. Si le contenu de la balise meta `description` est trop court ou pas assez cohérent, Google va tenter de chercher dans le contenu de la page un texte pour la décrire. Dans ce cas, vous ne maîtrisez plus ce que Google va indiquer comme résumé pour votre page... Il fait sa propre « cuisine » sur la base du texte qu'il lit dans votre page.
7. Si votre site est inscrit dans l'annuaire Open Directory et que la page proposée dans les résultats de Google est votre page d'accueil, Google pourra afficher le résumé indiqué dans cet annuaire comme snippet. Cette tendance semblait à la baisse chez Google en 2009 et vous pouvez l'interdire par l'utilisation de la balise meta `robots` (voir chapitre 9).

Yahoo!, de son côté, utilise également la balise meta `description` de façon assez similaire à Google, en complémentarité avec la définition issue de son annuaire ou des snippets (voir figure 4-20).

Figure 4-20

La description fournie par Yahoo! contient, au début, la balise meta description et est complétée par un snippet issu du texte de la page.

Cette balise permet donc de mieux maîtriser la présentation de la page proposée à l'internaute.

N'oubliez pas, comme nous l'avons vu précédemment, qu'il semblerait que la balise meta `description` soit affichée par les moteurs s'ils détectent une bonne homogénéité entre le titre de la page, le contenu de la balise meta en question et le texte du document.

Veillez donc bien à ce que le contenu de la balise meta `description` soit :

- un développement du titre de la page ;
- un résumé du contenu textuel de la page.

Ces deux conditions devraient faire en sorte que cette balise soit affichée dans les résultats des moteurs.

Si c'est possible, générez vos balises meta automatiquement : si vous utilisez un CMS (*Content Management System*), vous devriez avoir la possibilité de générer ces balises automatiquement en « piochant » des informations dans la page. Cela ne posera aucun problème aux moteurs de recherche, bien au contraire, ils encouragent même cette voie. Par exemple, vous pouvez y intégrer le chapô d'un article ou les 20 premiers mots d'un contenu éditorial qui résument souvent le contenu d'un texte, etc.

Plusieurs pistes peuvent ainsi être explorées pour améliorer vos balises meta `description` :

1. Proposez dans la balise meta `description` un contenu textuel différent de celui de la balise `<title>`. La balise meta doit compléter le titre sans – si possible – reprendre de façon littérale son contenu. Google donne, sur son blog (*http://googlewebmastercentral.blogspot.com/2007/09/improve-snippets-with-meta-description.html*), deux exemples de ce qu'il faut et ne faut pas faire.

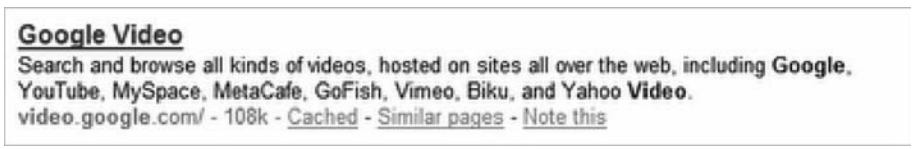


Figure 4-21

Balise meta « de qualité » selon Google

Voir ci-après plus d'explications sur cet exemple et un autre...

2. N'indiquez pas des listes de mots-clés séparés par une virgule dans cette balise. Cette forme de données est réservée aux balises meta `keywords` et les moteurs de recherche ne les apprécieront pas, ce qui induira leur non-affichage. Faites des « vraies » phrases contenant des mots descriptifs du contenu de la page et tout se passera au mieux.
3. Intégrez des données structurées. Pour un site d'actualité ou un blog, indiquez l'auteur, la date de parution ou toute autre information intéressante de ce type (on peut noter ici que, bizarrement, la plate-forme Blogger, qui appartient à... Google, n'offre pas de telles possibilités par défaut, mais il est possible de « ruser » pour y arriver). Bref, toute information qui ne sera pas affichée dans le titre mais qui peut le compléter est la bienvenue dans la balise meta...

Prenons l'exemple d'une balise meta jugée comme « non optimisée » par Google.

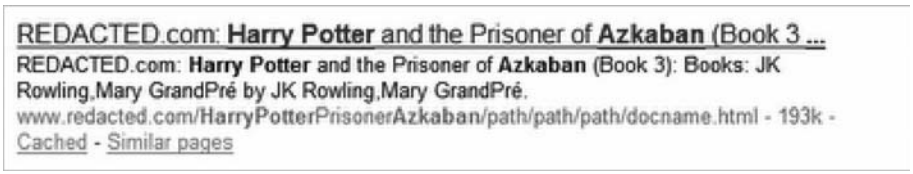


Figure 4-22

Balise meta « à revoir », toujours selon Google

Le contenu de la balise sera dans ce cas :

```
<meta name="description" content="[domain name redacted]: Harry Potter and the Deathly
➥Hallows (Book 7): Books: J. K. Rowling, Mary GrandPré by J. K. Rowling,Mary GrandPré">
```

Google explique, sur son blog, pourquoi ce type de balise meta n'est pas « recevable » selon lui :

- le titre du livre est repris dès le début et mot pour mot de la balise `<title>`, provoquant un doublon d'informations ;
- les noms de l'auteur (`J. K. Rowling`) et de l'illustratrice (`Mary GrandPré`) sont dupliqués à l'intérieur même de la balise ;
- certaines informations ne sont pas claires : qui est Mary GrandPré ? Il n'est pas indiqué qu'il s'agit de l'illustratrice du livre ;
- les espaces manquants et l'usage trop important des « : » rendent le descriptif complexe à lire.

Il s'agirait donc ici typiquement d'une balise meta `description` qui, malgré le fait qu'elle soit présente dans la page, pourrait ne pas être affichée par Google et qui de toute façon, si c'était le cas, ne rendrait pas service au site en question car elle n'inciterait pas au clic. Google, pour cet exemple, propose plutôt ce contenu :

```
<meta name="description" content="Author: J. K. Rowling, Illustrator: Mary GrandPré,
➥Category: Books, Price: $17.99, Length: 784 pages">
```

Ainsi, plus vous proposerez dans cette balise d'informations connexes permettant au moteur de mieux « comprendre » de quoi parle la page, meilleure sera la façon dont vous « rendrez compte » de son contenu auprès des internautes... et des moteurs.

Le travail sur le contenu des balises meta `description` peut s'avérer long et complexe, surtout si vous n'avez pas la possibilité de l'automatiser. Mais il sera certainement payant à moyen terme, non pas au niveau de vos positionnements – vous ne devriez *a priori* pas voir de grand changement de ce côté-là – mais plutôt sur la façon dont les internautes percevront et comprendront vos pages, bref, sur votre « retour sur investissement » et la satisfaction que vous apporterez à vos futurs visiteurs. Sur ce point, la balise meta `description` doit donc plutôt être appréhendée comme une zone marketing qui doit

donner envie aux internautes de cliquer pour venir sur votre site, beaucoup plus qu'une zone de pur positionnement algorithmique de pertinence...

Meta description : environ 200 caractères

Limitez le contenu de la balise meta `description` à 150, voire 200 caractères, espaces compris. La plupart des moteurs limitent l'espace alloué aux résumés. Si votre descriptif est plus long, faites en sorte que, réduite aux 150 premiers caractères, la phrase ait quand même un sens. Dans votre calcul, prenez en considération une lettre par caractère accentué - bien que la représentation de ceux-ci en langage HTML soit plus longue (8 caractères la plupart du temps, comme `&eacute;` pour le « e accentué »).

Bien entendu, pour être totalement efficace, chaque page de votre site doit contenir une balise meta `description` différente, décrivant exactement le contenu de la dite page ! Si ce n'est pas le cas, n'en proposez pas !

Notez enfin que Google développe de plus en plus une tendance à rallonger les snippets qu'il propose dans ses pages de résultats. Cela a commencé en octobre 2008 avec quelques tests (*http://actu.abondance.com/2008/10/des-rsums-plus-longs-en-test-sur-google.html*), puis l'affichage de résumés plus longs sur les requêtes contenant plus de trois mots-clés (*http://actu.abondance.com/2009/03/google-continue-semantiser-ses.html*). Enfin, en mai 2009, Google a proposé des options (*http://actu.abondance.com/2009/05/nouvelles-fonctionnalites-sur-la.html*) permettant d'allonger ou non le texte de ces snippets, comme indiqué sur la figure suivante.

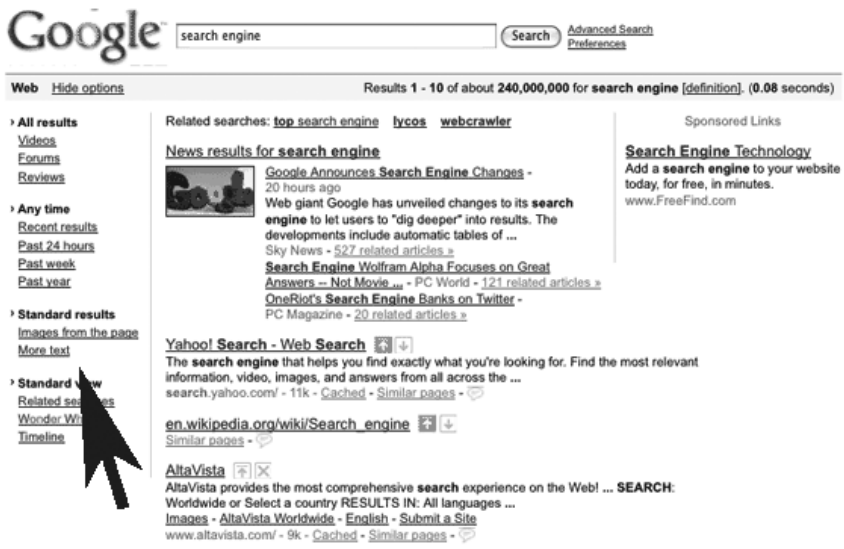


Figure 4-23

L'option « more text », sur la version américaine des pages de résultats de Google, permettant d'afficher des snippets plus longs

La morale de tout cela est qu'il va certainement falloir s'habituer, à l'avenir, à créer des balises meta `description` plus longues, tout en ne dévoilant pas trop d'informations. En effet, si toute l'information est dévoilée dans le snippet, l'internaute risque de ne plus cliquer et de ne pas aller sur votre site. C'est peut-être ce que désire Google (que l'internaute reste chez lui), mais ce n'est pas obligatoirement votre souhait...

Keywords : n'y passez pas trop de temps !

La balise meta `keywords` sert à fournir des mots-clés supplémentaires aux moteurs de recherche qui les prennent en compte, c'est-à-dire de moins en moins. Parmi les moteurs majeurs, seul Yahoo! semblait encore lire cette zone (il en est de même pour la balise meta `description`) en 2009, mais l'abandon de cette technologie au profit de celle de Microsoft pourrait bien condamner cette balise. Mais ce moteur octroie, dans son algorithme, un poids très faible à ce(s) champ(s). Autant être prévenu dès le départ...

Ces mots-clés servent à indiquer certains termes importants qui ne seraient pas présents dans le document. La balise meta `keywords` sert également à proposer diverses orthographes de vos mots importants aux moteurs de recherche. Ils sont séparés – au choix – par une virgule, un espace ou encore une virgule suivie d'un espace.

Dans le cas de notre société Stela, fabricant de chaussures de sport, une balise meta `keywords` pourrait être :

```
<meta name="keywords" content="stela, chaussure de sport, fabricant, tennis, running,
➥footing, stretching, athl&eacute;tisme, terre battue, dur, herbe, wimbledon,
➥flushing meadow, chaussures, Roland-Garros, flinders park, grand chelem" />
```

Il a longtemps été d'usage que la balise meta `keywords` contienne jusqu'à 100 mots-clés, ou 1 000 caractères. Au-delà, vous pouviez être considéré comme étant un spammeur, et votre page pouvait être pénalisée.

Le nombre de mots-clés présents dans vos balises meta `keywords` peut, en fait, être bien moins important, au vu de l'importance moindre aujourd'hui de cette zone d'information pour les moteurs. En règle générale, on pourra penser qu'avec une vingtaine, voire une trentaine, de mots-clés, vos balises meta `keywords` seront bien optimisées. N'y passez pas des heures, désormais le jeu n'en vaut plus la chandelle.

En règle générale, proposez dans cette balise les occurrences suivantes :

- noms communs : une occurrence au singulier et éventuellement une autre au pluriel et au féminin (singulier et pluriel). Par exemple : chien, chiens, chienne, chiennes ;
- lettres accentuées : indiquez une version non accentuée et une version accentuée en HTML. Par exemple : athletisme, athl`&eacute;`tisme. Il en est de même pour les caractères diacritiques, notamment le « ç » : francais, fran`&ccedil;`ais

Si possible, privilégiez :

- les noms communs : l'occurrence au singulier, puis éventuellement les occurrences au pluriel et au féminin. Privilégiez dans ce cas les mots les plus logiques qui sont susceptibles d'être saisis par un internaute ;

- les lettres accentuées (si prises en compte) : la version accentuée en HTML d'abord, puis la version non accentuée.

N'oubliez pas de proposer des expressions, des mots composés (chaussure de sport, Roland-Garros) en plus des mots isolés. Attention cependant à ne pas répéter un mot à l'intérieur de ces expressions.

Par exemple, la balise suivante :

```
<meta name="keywords" content="chaussure de sport, chaussure de tennis, chaussure de
➥footing, chaussure de training, chaussure de basket" />
```

risque d'être interprétée par les moteurs comme étant une tentative de fraude, car le mot « chaussure » est trop souvent répété. La bonne syntaxe serait plutôt :

```
<meta name="keywords" content="chaussure de sport, tennis, footing, training, basket" />
```

Vous choisirez dans la première expression (ici « chaussure de sport ») le mot le plus important pour votre activité afin de l'indiquer dans la première occurrence proposée. Les autres termes (dans l'exemple : tennis, footing, training, basket) viendront ensuite.

Par ailleurs, si vous reproduisez certains mots-clés, ne les mettez pas les uns à côté des autres (chien, chiens, chienne, chiennes, élevage, élevages), mais préférez la répétition d'une séquence de mots-clés : chien, élevage, bébé, chiot, puis chiens, élevages, bébés, chiots, puis chienne, élevage, bébés… En règle générale, comptez environ 10 mots entre deux occurrences proches d'un même terme. En clair, il faudra ici que 10 mots séparent chien et chiens, élevage et élevages, chien et chienne, etc. Belle gymnastique en perspective…

Attention ! Il ne s'agit pas de faire du remplissage dans la balise meta `keywords` avec des mots qui n'ont aucune chance d'être saisis par un internaute sur un moteur. Soyez réaliste, et ne retenez que des occurrences logiques et répandues de vos mots-clés.

Important : prêtez également attention aux éventuelles fautes de saisie qui pourraient être faites par les internautes si votre nom est complexe. Exemple : si votre société s'appelle Schmidt, insérez également les mots-clés schmit, chmidt, schmid, chmit, etc.

Terminons avec une mise en garde : si vous aviez l'idée d'inclure dans les balises meta `keywords` de vos pages tous les noms de vos concurrents – afin que vos pages soient trouvées même si l'internaute s'intéresse en priorité à vos rivaux économiques –, sachez que plusieurs entreprises américaines et françaises ont déjà attaqué en justice d'autres sociétés coupables de ce type de fraudes… et qu'elles ont gagné leur procès. Un webmaster averti en vaut donc deux.

Indiquez la langue

N'oubliez pas d'indiquer les mots-clés dans plusieurs langues, selon vos cibles et l'orientation linguistique de vos pages. Dans ce cas, l'option `lang` peut être indiquée :

```
<meta name="keywords" lang="fr" content="contenu de la balise" />
```

Si vous utilisez plusieurs langues dans votre page, employez plusieurs balises :

```
<meta name="keywords" lang="fr" content="contenu de la balise en français" />
<meta name="keywords" lang="en" content="contenu de la balise en anglais" />
```

Évitez cependant le plus possible les pages bilingues ou trilingues. Elles risquent de poser des problèmes lors de la recherche linguistique effectuée par les moteurs (voir précédemment).

Point important également : évitez d'insérer des retours chariots et autres renvois à la ligne en plein milieu d'une balise meta, quelle qu'elle soit. Tout passage à la ligne en plein milieu d'une balise pourrait poser des problèmes aux moteurs de recherche, certains d'entre eux n'arrivant pas à lire la balise de façon correcte.

Un rappel pour finir : très peu de moteurs de recherche prennent en compte aujourd'hui le contenu des balises meta dans leurs algorithmes de pertinence. N'y passez donc pas des heures, cela n'aurait qu'une utilité très limitée.

Utilité de la balise keywords dans certains cas

Si la présence de balises meta `keywords` est aujourd hui facultative dans vos pages, au vu du faible intérêt que leur portent les moteurs, elles peuvent avoir leur importance dans des pages contenant très peu de texte. Une page HTML lançant une animation Flash, par exemple et affichant un contenu textuel quasi inexistant, pourrait tirer parti d'une balise meta `keywords` bien remplie.

Notons également que le contenu de la balise meta `keywords` peut éventuellement être lu par votre technologie de recherche interne (intrasite). À vérifier…

Enfin, cet espace pourra éventuellement servir à créer une zone « mots-clés » si vous créez un Sitemap XML spécifique pour Google News (voir chapitre 8).

Ce sont à peu près là les seuls intérêts de la balise meta `keywords` à l'heure actuelle...

Seules comptent les balises meta description, keywords et robots

De très nombreuses autres balises meta sont disponibles et parfois visibles dans le code HTML des pages rencontrées au cours de vos pérégrinations sur le Web : `revisit-after`, `classification`, `distribution`, `rating`, `identifier-URL`, `copyright`, etc. Il faut savoir qu'elles ne sont clairement prises en compte par aucun moteur de recherche majeur. Leur présence est donc superflue dans vos pages, si ce n'est pour d'autres buts que le référencement.

Le mythe de la balise revisit-after

Lorsque vous trouverez une balise meta `revisit-after` dans le code source des pages web présentant l'offre d'une société de référencement, vous pouvez vous poser quelques questions sur ses compétences techniques. Et oui, cela arrive.

Une autre balise meta est cependant prise en compte : la balise `<meta name="robots">`, qui sera étudiée au chapitre 9 de cet ouvrage. Concernant les métadonnées Dublin Core, il est conseillé de ne les utiliser que si elles ont une utilité côté serveur et webmaster. Elles n'ont en revanche pas d'intérêt pour le référencement...

Pour résumer

Voici quelques conseils pour bien optimiser les balises meta de vos pages.

– Les balises meta `description` sont importantes pour mieux maîtriser la façon dont les moteurs affichent les résumés de vos pages. Leur importance en termes d'analyse de la pertinence d'une page par les moteurs est très faible.
– Une bonne balise meta `description` développe le titre de la page et en résume son contenu textuel. Sa taille moyenne est de 150 à 200 caractères.
– Idéalement, chaque page doit proposer une balise meta `description` qui lui est propre.
– Les balises meta `keywords` sont aujourd'hui moins prises en compte par les moteurs de recherche. Leur présence est facultative.
– Ces dernières balises permettent notamment d'indiquer plusieurs formes de mots importants (pluriel/singulier, masculin/féminin, et éventuellement l'accentuation).
– N'oubliez pas les éventuelles fautes de frappe et d'orthographe qu'il est toujours intéressant d'insérer dans les balises meta `keywords`.
– L'indication d'une trentaine de mots-clés est la plupart du temps suffisante.
– À part la balise meta `robots`, les autres balises meta (`revisit-after` et consors) n'ont aucune incidence dans le cadre d'un référencement.

Zone chaude 5 : attributs alt et title

De nombreux webmasters soignent les attributs `alt` et `title` (servant à l'accessibilité et à l'affichage de texte alternatif en l'absence d'image), dans les balises images et les liens du code HTML de leurs pages web, pensant que le contenu de ces options est pris en compte comme du texte par les moteurs de recherche. Exemple de code HTML de ce type :

```
<img src="http://www.votresite.com/images/logo.gif" alt="Mots-clés" title="Autres
➥mots-clés">
```

Les mots-clés insérés dans ces attributs sont-ils pris en compte par les moteurs ? Voici les conclusions tirées de tests effectués à la mi-2009 :

- l'attribut `alt` de la balise `<img>` (images) est pris en compte dans tous les cas par Google et jamais par Yahoo! et Bing ;
- l'attribut `title`, sur une image ou un lien textuel, n'est pris en compte par aucun moteur parmi les trois étudiés.

Notez bien deux points importants à ce sujet :

- la situation de Yahoo! et Bing, qui ne prenaient pas en compte ces champs à la mi-2009, peut évoluer… ;
- Google prend en compte les mots-clés insérés dans la balise `alt`, mais on peut parier que le poids qui leur est attribué dans l'algorithme de pertinence est faible.

Ne passez donc pas trop de temps sur ces champs qui n'ont qu'une utilité assez limitée (bien qu'elle ne soit pas nulle) pour l'optimisation de votre site.

Nous avons donc atteint la fin de ce chapitre sur les critères « in page ». Nul doute qu'il y a beaucoup de travail à faire dans ce domaine sur un site web. Mais ce n'est pas fini, car il nous faut maintenant parler des critères « off page », tout aussi importants... Vous êtes prêt ?

5

Optimisation des pages du site : les critères « off page »

Dans le chapitre précédent, nous avons essayé de traiter tous les aspects d'optimisation du code HTML d'une page web : balise `<title>`, prise en compte d'un texte visible bien structuré et dans lequel les mots importants sont mis en avant, balises meta, etc.

La plupart des premiers moteurs de recherche (AltaVista, par exemple) fonctionnaient sur ce mode, avec ce type de critère de pertinence. Google est ensuite arrivé et a changé la donne en introduisant des critères de pertinence basés sur l'analyse du contexte, de l'environnement de la page. La popularité (le célèbre PageRank) a été l'un des premiers, rapidement suivi par la réputation ou l'indice de confiance. Ces critères sont très importants sur Google et ses principaux concurrents – qui les ont adoptés dans la foulée – en 2009. Ils sont pourtant le plus souvent assez mal connus. Raison de plus pour les explorer en profondeur...

Liens et indice de réputation

De nombreux référenceurs vous le diront : la meilleure façon d'obtenir une meilleure visibilité sur les moteurs réside aujourd'hui dans une bonne gestion de vos liens ou backlinks. Nous allons voir pourquoi. Et, à tout seigneur tout honneur : nous allons commencer, de façon logique, avec les liens présents dans vos pages web, puisque, *a priori*, ce sont ceux que vous maîtrisez le mieux.

Il faut bien être conscient que les liens sont très importants pour les moteurs de recherche car ils permettent à leurs robots d'explorer votre site pour y « cueillir » d'autres documents. Les robots (*spiders*) suivent ainsi les liens présents dans vos pages et indexent de

nombreux documents sans que vous n'ayez à faire une quelconque soumission (voir chapitre 8). Il est donc très important que vos liens soient compatibles avec les *spiders* des moteurs, comme nous le verrons par la suite.

Un credo doit être le vôtre lorsque vous bâtissez vos pages afin que celles-ci soient réactives par rapport aux moteurs de recherche : créez des liens les plus simples possible !

Plus vos liens se rapprocheront de la forme « simple » HTML suivante, mieux cela vaudra :

```
<a href="http://www.votresite.com/page-de-destination.html">texte du lien</a>
```

Réputation d'une page distante

Attention : le texte du lien (qui apparaîtra donc dans vos pages comment étant cliquable) est primordial. Sur Google, par exemple, il va servir à donner un thème à la page de destination et représente pour elle un critère de pertinence crucial, d'où la notion de « réputation ». Explications…

Prenons un exemple. Vous gérez un site sur les assurances. Sur votre page d'accueil, vous proposez les liens suivants :

```
Notre offre en <a href="http://www.votresite.com/assurance-vie.html">assurance-vie</a>
Notre offre en <a href="http://www.votresite.com/assurance-auto.html">assurance auto</a>
Notre offre en <a href="http://www.votresite.com/assurance-moto.html">assurance moto</a>
```

On voit dans cet exemple que le texte du lien qui pointe vers ces pages contient les mêmes mots-clés. La page qui parle de l'assurance-vie est pointée par un texte qui s'intitule « assurance-vie ». Idem pour les deux autres pages.

Ce critère va alors faire en sorte que la page en question (page cible) sera certainement bien considérée par Google – et les autres moteurs majeurs – pour l'expression citée dans le texte du lien (« assurance-vie »). Elle a la « réputation » de parler d'assurance-vie...

Si cette page de destination contient de plus un bon titre et du texte optimisé, vous n'êtes plus très loin de la première page, voire de la première place, sur Google.

Le Google Bombing

Ce fait est bien illustré par les actes de *Google Bombing* entrevus ces derniers temps sur le Web. On se souvient que Georges Bush a été victime d'une action de ce type (*http://actu.abondance.com/2003-50/miserable-failure.html*). Lorsqu'on tapait la requête « miserable failure » sur Google, le premier résultat affiché était… la biographie officielle de George W. Bush, sur le site de la Maison Blanche. En février 2004, c'est le député Jean Dionis, partie prenante dans la nouvelle loi sur l'économie numérique, qui a fait les frais de ce type d'action : son site sortait premier sur Google pour la requête « député liberticide » (*http://actu.abondance.com/2004-07/jean-dionis.html*). Depuis, les actes de Google Bombing se sont multipliés et de nombreux hommes politiques français, notamment, en ont été victimes… Mais, en 2007, Google a travaillé sur le sujet devant la multitude de *bombings* perpétrés (*http://actu.abondance.com/2009/01/google-bombing-sur-barack-obama-google.html*) et il est devenu de plus en plus difficile de mettre en place ce type de châtiment numérique. La plupart d'entre eux ont aujourd hui disparu...

Lancer une opération de Google Bombing n'est pas très complexe en soi : il suffit de multiplier, sur le plus de sites possible, les liens pointant vers le site à « bomber », tout en indiquant la requête désirée dans le texte du lien. Exemple : vous désirez que le site Abondance soit premier sur la requête « meilleur site sur les moteurs de recherche » ? Nous n'en doutons pas un instant. Vous multipliez alors dans vos pages les liens de ce type : meilleur site sur les moteurs de recherche, pointant sur la page d'accueil du site Abondance, en répétant ce lien sur le plus de sites possible. Et le tour est joué ! Sur des requêtes générant peu de résultats (moins de 100 000) ou sur des mots-clés peu concurrentiels, il y a de fortes chances pour que le positionnement attendu soit au rendez-vous. Et pourtant, le site cible en question ne contient pas obligatoirement les termes de la requête !

Soignez les libellés de vos liens

Le texte du lien (texte cliquable) est donc extrêmement important pour le positionnement de vos pages. Ne le sous-estimez pas. Par exemple, évitez des phrases comme :

- « Pour consulter nos offres d'assurance-vie, cliquez ici ».
- « Notre offre d'assurance-vie est l'une des meilleures du marché. Elle vous propose un rapport qualité-prix incomparable. Lire la suite… »
- « En savoir plus… »

En effet, les expressions « cliquez ici », « Lire la suite » ou « En savoir plus » ne sont pas toujours très pertinentes pour qualifier les pages sur lesquelles l'internaute se rendra s'il clique sur le lien. Elles perdront donc inévitablement du poids, donc des positions, pour les moteurs de recherche.

Dans ce cas, préférez donc :

- « Consultez nos offres d'assurance-vie ».
- « Notre offre d'assurance-vie est l'une des meilleures du marché. Elle vous propose un rapport qualité-prix incomparable ».

Pour résumer, on peut dire que les liens hypertextes insérés dans les pages web de votre site sont importants :

- pour insérer des mots-clés donnant un poids plus fort à la page qui les contient (que l'on peut appeler « page origine »), comme nous l'avons vu précédemment dans le chapitre 4 ;
- pour insérer des mots-clés donnant un poids plus fort à la page vers laquelle il dirige (« page cible »).

Point important également pour la réputation d'une page : plus la page contenant le lien pointant vers ce document dispose d'un PageRank (voir plus loin) élevé, plus sa réputation sera forte. Plus clairement, si A pointe vers B, le fait que A ait un PageRank élevé (supérieur ou égal à 6) augmentera encore la notion de réputation de B. Mais nous y reviendrons très bientôt.

À éviter le plus possible : images, JavaScript et Flash

Nous l'avons vu, les liens textuels les plus simples sont les plus efficaces. Mais il existe d'autres façons de construire des liens. Par exemple, les liens images comme dans le code suivant :

```
<a href="http://www.votresite.com/page-de-destination.html">
<img src="image.gif"/>
</a>
```

Dans ce cas, le texte du lien est remplacé par une image. Si l'on clique sur celle-ci, on est redirigé vers la page de destination. Dans ce cas, le lien est « lisible » par les moteurs (leurs robots sauront « suivre » ce lien pour indexer la page cible). En revanche, l'absence de texte sera préjudiciable pour la réputation de la page cible. Et l'attribut `alt` de la balise `<img>` ne suffira pas à compenser ce manque...Voici ce qu'il faut éviter :

- **Le JavaScript.** Les moteurs, en règle générale, n'aiment pas le JavaScript et ne lisent pas les adresses qui y sont insérées. Nous y reviendrons au chapitre 7.
- **Le Flash.** Comme pour le JavaScript, les moteurs aiment peu le Flash. Si Google suit parfois les liens insérés dans certaines animations Flash, il semble que cela ne soit pas une règle établie et systématique. On préférera donc retenir l'adage selon lequel « tout ce qui est présent dans une animation Flash est ignoré par les moteurs ». Donc les liens le sont également. Là encore, reportez-vous au chapitre 7 pour plus d'explications.
- **Les formulaires.** Certains liens peuvent être proposés sous la forme de formulaires, et notamment de menus déroulants, comme dans l'exemple de la figure 5-1.

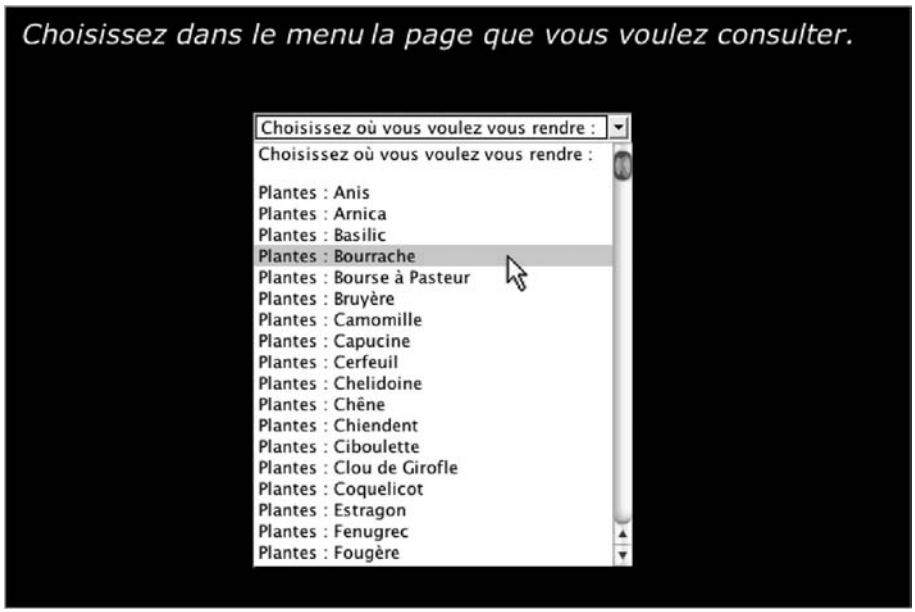


Figure 5-1

Exemple de formulaire web

Là encore, la situation risque de se complexifier pour les moteurs de recherche. Certains se débrouilleront avec ce type de lien, tandis que d'autres – la majorité – auront plus de mal. Ils liront le contenu des intitulés du menu déroulant comme du texte à part entière, mais les liens ne seront pas considérés comme tels. En tout état de cause, les formulaires, en guise d'outils de navigation, constituent un obstacle, plus ou moins bloquant, pour la plupart des moteurs. Ils sont donc à utiliser avec parcimonie. Rendez-vous au chapitre 7 pour plus d'informations.

Les liens sortants présents dans vos pages

Les liens sortants représentent les liens insérés dans vos pages et pointant vers l'extérieur ou vers des sites qui ne vous appartiennent pas.

A priori, ces liens ne jouent aucun rôle dans l'algorithme de classement de vos documents par les moteurs (même s'il semblerait que ce soit là une piste de réflexion dans certains laboratoires de recherche). Le nombre et la destination des liens de vos pages ne leur serviront donc pas à être mieux classées sur les moteurs de recherche, à part pour le calcul du PageRank (voir plus loin). Mettre en place dans vos pages un lien vers le site de Google, d'Amazon ou d'eBay ne jouera donc en rien sur votre positionnement dans les pages de résultats des moteurs.

De même, ce n'est pas parce que vous avez inséré un formulaire de recherche Google ou un lien vers le célèbre moteur dans vos pages que vous y serez mieux positionné pour vos mots-clés favoris. Il s'agit là d'une croyance que l'on rencontre parfois sur certains forums. Il n'en est rien !

Pour résumer

Voici quelques conseils pour bien optimiser les liens de vos pages :

- L'optimisation d'un lien est importante à la fois pour la page qui le contient (page origine) et pour la page vers laquelle il pointe (page cible).
- Soignez particulièrement le texte du lien (le texte cliquable) qui doit être pertinent par rapport à la « réputation » de la page cible.
- Vos liens doivent être le plus simple possible pour que les robots des moteurs puissent les suivre afin d'indexer les autres documents de votre site.
- Privilégiez les liens textuels et évitez le plus possible les liens images, JavaScript, formulaires ou Flash.

Liens, PageRank et indice de popularité

Ce n'est un secret pour personne, tous les moteurs de recherche majeurs actuels, de Google à Bing en passant par Yahoo! et Exalead, utilisent l'indice de popularité (*link popularity* ou *link analysis* en anglais) dans leurs critères de pertinence. On peut même dire que ce paramètre est aujourd'hui devenu une partie importante des algorithmes de pertinence et que l'indice de popularité (IP, à ne pas confondre avec l'*Internet Protocol*

du « numéro IP »…) figure parmi les cinq critères majeurs sur tous les moteurs avec le titre des pages, le texte visible, l'URL et la réputation.

Comment l'indice de popularité est-il calculé ?

Au départ, il y a quelques années de cela, cet indice de popularité n'était calculé que selon un mode quantitatif : plus une page avait, dans l'index du moteur, de liens qui pointaient vers elle, plus son indice de popularité était élevé. Il n'en est rien aujourd'hui et tous les moteurs de recherche ont mis en place des modes de calcul bien plus élaborés pour quantifier ce critère, en tenant notamment compte de la qualité des liens trouvés vers la page cible. Il n'est donc pas réellement nécessaire d'avoir énormément de liens pointant vers vous pour obtenir une bonne popularité sur Google ou Yahoo!, mais il vaut mieux, et de plus en plus, avoir des liens « à forte valeur ajoutée ». Pas obligatoirement plus que vos concurrents, mais de meilleure qualité, donc émanant de pages elles-mêmes populaires. Google a fait de son système d'analyse de la popularité des pages, appelé « PageRank » du nom de son concepteur Larry Page (cocréateur de Google avec Sergey Brin), l'un des fleurons de son algorithme de pertinence et de sa communication.

Aujourd'hui donc, les moteurs de recherche utilisent plusieurs familles de données et de critères pour calculer ce paramètre (rappelons que le calcul est effectué sur la base des pages présentes dans l'index du moteur et seulement celles-ci. Il ne sert à rien d'avoir des liens forts vers son site, encore faut-il que les pages qui les contiennent soient bien dans l'index du moteur en question pour être prises en compte). Voici quelques données qui devraient vous être utiles pour améliorer votre situation à ce niveau.

- Les aspects quantitatif et qualitatif sont le plus souvent pris en compte à deux niveaux. Le moteur calcule non seulement l'IP d'une page, mais également celui des pages pointant vers lui (voir ci-après). Donc, un lien depuis une page à forte popularité sera plus important qu'un lien émanant d'une page perso *lambda*. Il peut suffire d'avoir peu de liens mais provenant de pages très populaires, plutôt qu'une multitude de liens émanant de pages peu connues et isolées. Le quantitatif a vécu, place au qualitatif. Ceci dit, si vous disposez d'une multitude de liens émanant de pages très populaires, c'est encore mieux.
- Le nombre de liens présents dans les pages pointant vers vous est également de plus en plus important (voir la formule de calcul de Google ci-après). Plus la page qui pointe vers vous contiendra de liens divers et variés, plus son importance diminuera, plus elle sera diluée parmi tous les liens proposés. Ceci peut défavoriser les longues pages de liens, de type FFA ou *links farms* (voir plus loin), qui sont rarement lues et n'ont finalement que peu d'intérêt, autre que celui de faire croire qu'elles vont augmenter votre IP, ce qui est faux en grande partie.
- Le fait que les liens vers une page soient internes ou externes peut être important. Certains moteurs peuvent compter les liens internes de votre site dans leurs calculs, soit les exclure (rarement), soit leur donner un poids plus faible (le plus souvent) pour prendre davantage en considération les liens provenant d'autres sites que le vôtre, ce qui est assez logique.

- L'indice de popularité est calculé par rapport à une page précise, et non pour un site de façon globale. La page d'accueil de votre site aura donc, le plus souvent, le plus important indice de popularité parmi toutes vos pages, car il y a fort à parier – sauf exception – que la plupart des liens du Web renvoient vers elle. Attention, notamment, à la façon dont vos pages sont adressées. Par exemple, la page d'accueil du site Abondance est accessible *via* les adresses suivantes : *abondance.com, www.abondance.com, www.abondance.net, www.abondance.fr, www.abondance.com/index.html*, etc. Sur certains moteurs (dont Google), chaque URL sera considérée comme un site différent. Sur d'autres, peu importe l'adresse, toutes les formulations seront identiques.
- Seuls les liens pointant vers vous (liens entrants ou backlinks) sont pris en compte. Les liens émanant de vos pages pour aller vers d'autres sites (liens sortants) ne sont, pour le moment, pas pris en compte dans le calcul de l'indice de popularité de cette page.

Mode de calcul du PageRank

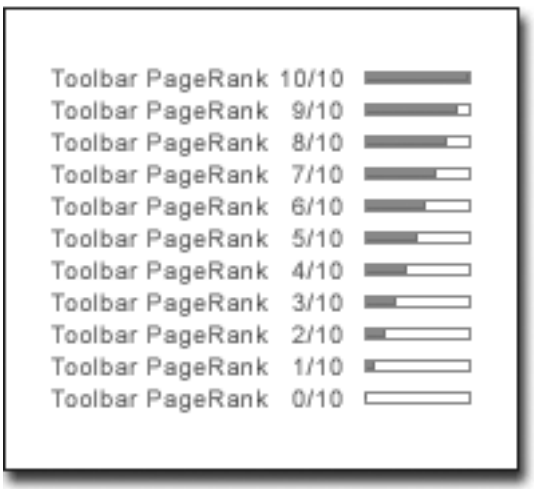
Le moteur de recherche Google utilise fortement l'indice de popularité baptisé « PageRank » dans son algorithme. Comme vous pouvez le voir sur les figures 5-2 et 5-3, il est affiché dans la barre d'outils de Google (*http://toolbar.google.fr/*) sous la forme d'une « note » allant de 0 à 10.

Figure 5-2

Affichage du PageRank dans la barre d'outils de Google

Figure 5-3

La valeur affichée du PageRank varie de 0 à 10.



Dans le document intitulé *The Anatomy of a Large-Scale Hypertextual Web Search Engine* (*http://infolab.stanford.edu/~backrub/google.html*), les deux créateurs de Google (Sergey Brin et Larry Page) fournissent la formule itérative de calcul de cet indice :

$$PR(A) = (1-d) + d(PR(T1)/C(T1) + PR(T2)/C(T2) + \ldots + PR(Tn)/C(Tn))$$

où :

- PR(A) est égal au PageRank de la page A ;
- Tn (pages sources) représente les pages pointant (ayant mis en place un lien) vers la page A (page cible) ;
- C(Tn) représente le nombre de liens présents dans la page Tn ;
- d est un facteur multiplicatif, égal au lancement de Google à 0,85.

Google justifie ainsi sa formule : elle peut être imaginée comme représentative du comportement d'un internaute qui effectuerait une séance de surf sur le Web et partirait d'une page web, au hasard, puis cliquerait sur tous les liens qu'elle présente, et continuerait ainsi à cliquer sur tous les liens qu'il rencontre. Éventuellement, cet internaute « cliqueur fou » pourrait se lasser et repartir, à un moment ou à un autre, d'une nouvelle page de départ. Dans cette métaphore, la probabilité qu'une page soit visitée par l'internaute est représentée par son PageRank. Et le facteur « d » représente le fait que l'internaute fou change, à un moment ou à un autre, de page de départ pour repartir sur un nouveau surf. Nous vous laissons méditer quelques minutes sur ce paragraphe…

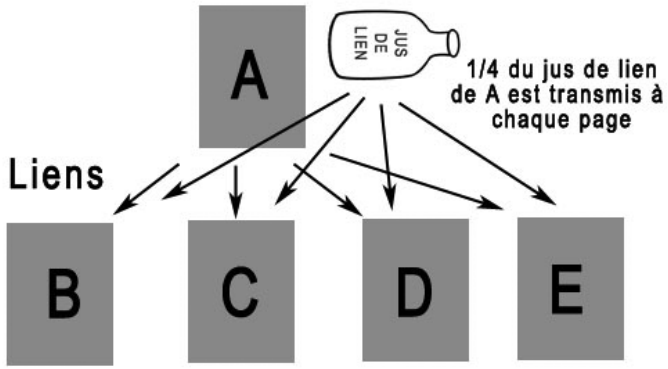
Bien sûr, la formule originelle du calcul du PageRank a certainement été améliorée, les centaines d'ingénieurs de haut vol embauchés depuis par Google travaillant dessus au quotidien. Mais il y a également de fortes chances pour que le « noyau » du calcul soit resté fidèle au modèle d'origine qui montre plusieurs points importants qui restent ainsi certainement d'actualité.

- La valeur de l'IP sur Google est proportionnelle au nombre de liens pointant vers une page et à leur qualité (l'IP de chaque page source pointant vers le document cible).
- L'IP d'une page cible est inversement proportionnel au nombre de liens présents à l'intérieur des pages sources. Plus la page qui pointe vers vous contient de liens, plus son influence sur votre IP faiblit. La transmission de la popularité d'une page au travers des liens est souvent appelée « jus de lien » ou *Link Juice*. Une page détient une certaine popularité (son PageRank) et la transmet aux autres pages au travers de ses liens sortants. Plus il y a de liens dans la page et moins chaque page pointée par un lien reçoit de jus de lien. On peut tout à fait symboliser ce concept sous la forme d'une bouteille de jus de lien détenue par une page donnée. Si cette page met en place dix liens, chacune de ces pages reçoit un dixième de la bouteille de jus de lien. Si cent liens

sont créés, chacune des pages distantes reçoit un centième du jus de lien de la page originale, etc.

Figure 5-4

Un lien transmet du « jus de lien » vers les pages distantes. La bouteille de jus de lien de A est partagée équitablement entre chaque page pointée par un lien. Plus il y a de liens dans A, plus la part de jus de lien reçue par chaque page est faible.



- Les pages internes d'un site sont prises en compte dans le calcul du PageRank d'un document donné. Tous les liens de sources interne et externe ont donc leur importance.

Le processus de calcul de l'IP est très long car il faut manipuler d'énormes matrices de plusieurs centaines de millions de liens (et de milliards pour les plus gros index). Il s'agit en fait d'un processus itératif de type « point fixe » : le calcul étant rétroactif, il faut itérer un certain nombre de fois la formule pour que l'algorithme converge. La théorie indique qu'après N itérations, l'algorithme doit converger vers une solution stable (un point fixe de type $PR = M \times PR$, où PR est l'équivalent du PageRank de la page en question et M une matrice de liens).

Notons, pour être complet, un ensemble de travaux portant plus spécifiquement sur le paramètre des liens sortants. Les travaux du projet Clever d'IBM (*http://www.almaden.ibm.com/projects/clever.shtml*) et les algorithmes dits « HITS » et « HITS improved » (*http://www.math.cornell.edu/~mec/Winter2009/RalucaRemus/Lecture4/lecture4.html*) travaillent sur les liens sortants d'un site. Ils permettent de trouver les sites de communautés (c'est-à-dire, pour un sujet donné, les sites contenant beaucoup de liens sur ledit sujet) – contrairement aux algorithmes de type PageRank qui trouvent les sites de référence (donc les sites les plus cités pour un sujet donné). Un moteur comme Ask (*http://www.ask.com/*) est, par exemple, conçu sur ce principe issu des algorithmes HITS (technologie Teoma).

Ces algorithmes basés sur les communautés sont intéressants car ils permettent de trouver des points de départ très pertinents pour certaines recherches génériques (ainsi que des périmètres, c'est-à-dire des ensembles de pages traitant d'un même sujet).

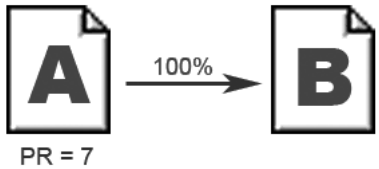
Le PageRank en images

Pour avoir l'esprit plus clair sur le PageRank (PR), nous pouvons l'expliciter au travers d'images et d'exemples.

- **Exemple 1.** A (de PR 7) pointe vers B

Figure 5-5

A (PR 7) vote pour B.

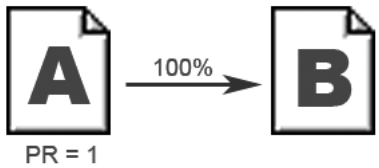


La page A, qui bénéficie d'un PageRank (PR) de 7 va fortement influer sur le PR de B et le faire augmenter en proposant un lien vers cette page. De plus, comme le seul lien sortant de la page A va vers B, cette dernière page profite de 100 % de la capacité de vote (un lien étant considéré comme un vote par les moteurs de recherche), du « jus de lien » de A.

- **Exemple 2.** A (de PR 1) pointe vers B

Figure 5-6

A (PR 1) vote pour B.

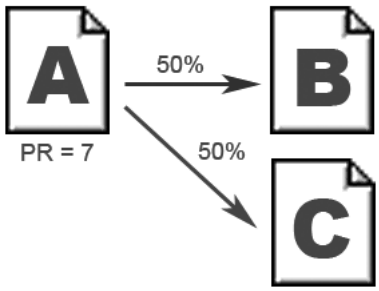


Dans ce cas, B profite toujours des 100 % de capacité de vote de A, mais cette dernière page étant très peu populaire (PR = 1), ce lien ne fera que faiblement augmenter le PR de B. Notons également qu'il n'influencera pas de façon négative le PR de B.

- **Exemple 3.** A (PR 7) pointe vers B et C

Figure 5-7

A (PR 7) vote pour B et C.



Dans ce cas, le PR de A est fort (7) et les liens vers B et C vont augmenter l'IP de B et de C. En revanche, du fait qu'il existe maintenant deux liens sortant de A (un vers B et un vers C), chacune des deux pages de destination va donc se partager pour moitié le jus de lien de A. L'impact de A sur le PR de B et C sera donc moins fort que dans notre premier exemple…

Spamdexing ou non ?

Le système d'indice de popularité a été mis en place, au départ, car il était très difficile à contourner par les webmasters, comparé notamment à des critères dans la page et contenus dans le code HTML comme les balises meta, aujourd'hui presque complètement abandonnées. Cependant, certaines pratiques, parfois utilisées aujourd'hui, sont assez contestables, même si leur impact reste très faible.

- Les FFA (*Free For All Links* ou *links farms*), qui sont d'immenses pages de liens sur lesquelles les webmasters peuvent inscrire gratuitement et sans contraintes un lien vers leur site. Le faible IP de ces pages ainsi que le grand nombre de liens qu'elles proposent (ajoutés au fait que les moteurs les chassent pour les enlever de leurs index) en limitent grandement la portée.
- Les offres de création de liens factices. Actuellement, il existe de nombreuses offres qui vous proposent d'augmenter de façon artificielle l'IP de votre site web en vous inscrivant, parfois de façon payante, dans des systèmes d'échange de liens « bidons ». Le principe en est simple : plusieurs milliers de pages mettent en place un lien vers la page d'accueil de votre site pour faire grandir votre indice de popularité dans l'algorithme de pertinence des moteurs de recherche et donc votre positionnement. Sur le papier, la théorie semble donc tenir la route. Mais en pratique, qu'en est-il ? En fait, le système se heurte à quelques obstacles non négligeables.
  - Tout d'abord, avoir des liens vers votre site web est une chose, mais cela ne peut améliorer votre popularité que si les pages contenant lesdits liens se trouvent intégrées dans l'index des moteurs de recherche majeurs. Et rien ne prouve que cela soit le cas. Là aussi, les moteurs deviennent très pointus dans la chasse au *link spamming* et toutes ces pages sont de plus en plus traquées, ôtées de l'index et le site les proposant est parfois mis en *blacklist*, c'est-à-dire en liste noire.
  - Le système fonctionne uniquement si le moteur prend en compte un indice de popularité à un niveau, de façon quantitative, dans son algorithme. Or, ce n'est pas le cas et ces pages de liens sont elles-mêmes très rarement populaires. Leur impact est donc très limité.
  - Enfin, et c'est peut-être le plus important à nos yeux, ces offres proposent purement et simplement de tromper l'algorithme de classement des moteurs de façon artificielle. Il s'agit de pratiques qui peuvent être assimilées à du spam.

À vous de voir, donc, au vu de ces quelques éclaircissements, si vous désirez profiter de ces offres (qui semblent d'ailleurs en perte de vitesse) selon votre propre vision des ambitions de votre site sur le Web. Mais attention : comme nous l'avons dit, la plupart

des moteurs ont mis en place des algorithmes de détection des tentatives de *link spamming*. Ne vous amusez pas trop à ce petit jeu…

Le PageRank seul ne suffit pas

Il est important de ne pas oublier que l'indice de popularité n'est qu'un critère de pertinence, parmi d'autres, utilisé par les moteurs de recherche. Si votre site ne propose pas (ou que peu) de texte visible, que vos pages n'ont pas été modifiées depuis deux ans (la fraîcheur des documents est également un point important) et que les titres de vos pages sont bâclés, le moteur n'y trouvera pas les mots-clés importants pour votre activité et ne prendra tout simplement pas en compte vos documents, quel que soit leur popularité. Le PageRank seul ne suffit pas à voir vos pages bien classées dans les pages de résultats des moteurs, ne l'oubliez pas !

On en revient donc toujours à la même conclusion : faites des pages avec du vrai contenu, intéressant, original, en texte visible. Les webmasters créeront ensuite tout naturellement des liens vers votre site et vos pages bénéficieront d'un bon PageRank. Elles seront mieux classées sur les moteurs de recherche et tout sera plus facile…

Mise à jour du PageRank

Un autre point à ne pas oublier également : il est clair aujourd'hui que le PageRank affiché par la barre d'outils de Google n'est pas le même que celui utilisé par Google dans son algorithme de pertinence.

Un officiel de Google a indiqué en 2005 que l'affichage du PageRank dans la barre d'outils du moteur était là uniquement « pour le fun » (*http://www.searchenginejournal.com/google-pagerank-is-for-entertainment-purposes-only/1105/*). Voici ce qu'il a dit à ce propos : « The PageRank that is displayed in the Google Toolbar is for entertainment purposes only. Due to repeated attempts by hackers to access this data, Google updates the PageRank data very infrequently because is it not secure. On average, the PR that is displayed in the Google Toolbar is several months old. If the toolbar is showing a PR of zero, this is because the user is visiting a new URL that hasn't been updated in the last update. The PR that is displayed by the Google Toolbar is not the same PR that is used to rank the webpage results so there is no need to be concerned if your PR is displayed as zero. If a site is showing up in the search results, it doesn't not have a real PR of zero, the Toolbar is just out of date. »

Au moins, cela a le mérite d'être clair : les informations de la barre d'outils de Google semblent donc obsolètes et uniquement présentées à titre d'information. Il semblerait en fait qu'il existe plusieurs PageRank : un qui est affiché dans la barre d'outils (parfois appelé TBPR pour *ToolBar PageRank*), un autre qui est affiché dans l'annuaire de Google (différent de celui du moteur) et le « vrai », utilisé par Google dans son algorithme de pertinence qui, lui, n'est pas public.

Sachez également que la valeur du PR d'une page affichée dans la barre d'outils de Google est mise à jour environ tous les deux à trois mois. C'est cette période de remise à jour du PageRank que l'on appelle aujourd'hui la *Google Dance*. Une page web peut donc afficher un PR nul dans la Google Toolbar et bénéficier d'un PR réel plus élevé qui ne sera révélé qu'à la prochaine Google Dance. Mais pour ce dernier, il est impossible de le connaître, à moins d'être embauché par Google.

Le netlinking ou comment améliorer son indice de popularité

L'échange de liens (ou *netlinking*) est une stratégie de promotion de sites web très efficace depuis que le Web existe et permet d'obtenir de nouveaux liens afin d'améliorer un indice de popularité. En effet, il cumule deux avantages principaux et indéniables :

- Les liens créent en eux-mêmes du trafic puisque les internautes qui naviguent sur la Toile vont aller sur votre site s'ils découvrent sur d'autres pages des voies d'accès vers vos documents.
- Le nombre et la qualité des liens pointant vers votre site et vos pages sont des caractères essentiels et très importants de calcul de l'indice de popularité (selon la notion du PageRank pour Google) de votre source d'information en ligne, comme nous venons de le voir. Plus vous aurez de liens émanant de sites de qualité vers vos pages, meilleur sera votre positionnement sur les moteurs. Une bonne raison pour soigner cet aspect de la promotion de votre site.

Mais est-il important de faire de l'échange de liens tel qu'on le faisait il y a encore quelques années, un pur échange « lien pour lien » (je pointe sur toi, tu pointes vers moi) sans autre considération ? Pas si sûr… Pour en savoir plus, nous avons essayé de rassembler dans les paragraphes suivants plusieurs conseils pour vous permettre d'optimiser vos échanges avec des sites distants. À vous de voir lesquels sont les plus intéressants pour vous.

Conseils d'ordre général

- Lors de vos campagnes d'échange de liens, ciblez des sites à forte popularité plutôt que des sites peu connus. Si un site vous intéresse, regardez le PageRank de sa page d'accueil et/ou faites une requête de type « link:www.sitepartenaire.com » sur Google pour avoir une idée du nombre de liens pointant vers lui. Préférez éventuellement une requête de type « linkdomain:votresite.com » sur Yahoo!, les résultats sont le plus souvent bien plus fiables et exhaustifs que sur Google !

  Si plusieurs sites vous intéressent, faites la même chose pour chacun d'eux et contactez ceux disposant du plus grand nombre de liens entrants (ou backlinks). Même si on a vu que l'aspect quantitatif n'était pas suffisant, il donne quand même une bonne idée du potentiel d'un site.
- De plus, ne tentez des échanges de liens qu'avec des sites parlant de sujets les plus proches, les plus connexes possibles des vôtres. Vous renforcerez ainsi votre pertinence

sur les mots-clés les plus importants, car les liens qui pointeront vers vous seront issus de sites proposant du contenu très proche du vôtre.

- Référencez votre site dans les annuaires majeurs, dans les meilleures catégories possibles. Le fait d'être dans Yahoo!, l'Open Directory ou tout site à forte popularité, améliorera votre popularité (voir chapitre 3).

  Attention cependant aux systèmes de redirection, comme sur Yahoo.com. En effet, allez, par exemple, dans la catégorie *http://dir.yahoo.com/Business_and_Economy/Employment_and_Work/News_and_Media/Magazines/*.

  Si vous regardez la liste des sites proposés, le premier est « Equal Opportunity Publications ». Le site est présent à l'adresse *http://www.eop.com/* mais l'adresse proposée sur la page de Yahoo.com est en fait *http://rds.yahoo.com/S=70868:D1/CS=70868/SS=59913/SIG=10njuuqgj/*http%3A//www.eop.com/*. Le lien n'est donc plus vers le site EOP mais plutôt vers le site RDS de Yahoo!, qui est un système de traçage des clics effectués sur l'annuaire. Cela signifie que toute inscription et toute apparition dans l'annuaire de Yahoo! ne sera pas prise en compte dans le calcul de l'indice de popularité des moteurs.

  Lors d'une inscription dans un annuaire, vérifiez donc bien que le lien se fait bien « en dur » (lien direct vers la page) et donc sans système de redirection. Il en est de même avec les bannières publicitaires (qui pointent vers une régie) ou de nombreux systèmes d'affiliation…

- Autre enseignement du paragraphe précédent : le nombre de liens sur la page source étant important dans un annuaire (qui propose le plus souvent des listes de sites), il vaut mieux essayer de demander, sur les annuaires majeurs, des catégories proposant peu de sites, plutôt que des rubriques déjà bien remplies et affichant plus de 50 sources d'information, diluant ainsi la capacité de vote de la page où est référencé votre site.

Les X commandements du « bon lien »

Un « bon lien » doit présenter un maximum de particularités parmi la liste ci-dessous.

– Il doit émaner d'une page populaire (PageRank supérieur ou égal à 4 ou 5).
– Il doit émaner d'une page issue d'un site de la même thématique que le vôtre.
– Si c'est un site de référence du domaine, c'est encore mieux.
– Il doit émaner d'une page contenant le moins possible de liens sortants.
– Le texte du lien (*anchor text*) doit décrire ce que l'internaute trouvera dans la page (éviter les « cliquer ici » ou les « pour en savoir plus »).
– Un lien aura une meilleure efficacité au niveau de la pertinence s'il est placé au cœur de la page, intégré dans un contenu et sur plusieurs pages du site au lieu d'une seule (voir ce billet sur le blog d'Abondance : *http://blog.abondance.com/2007/04/entre-lien-structurel-et-contextuel.html*).

Évitez le simple « échange de liens »

Si vous êtes webmaster d'un site, vous avez certainement reçu, un jour ou l'autre, un e-mail de ce type :

> *Bonjour*
>
> *J'ai particulièrement apprécié le contenu de votre site. Je vous propose d'échanger un lien avec le nôtre, disponible à l'adresse :*
>
> *http://www.tartempion.fr/*
>
> *Merci et bien cordialement*
>
> *Le webmaster du site Tartempion.fr*

Honnêtement, parmi tous les e-mails de ce genre que vous avez reçus, combien ont reçu une issue positive ? Certainement très peu. Ces messages, qui sentent bon (*sic*) le e-mailing massif effectué sans distinction sur des centaines voire des milliers de sites, ne sont pas très efficaces : aucune personnalisation, aucune information sur ce que propose le site demandeur et son adéquation à votre propre source d'information… En règle générale, tout cela part à la corbeille en moins de temps qu'il n'en faut pour l'écrire.

Soyons clair : ce type d'échange n'est en rien un vrai partenariat, mais plutôt un système de troc ponctuel qui ne transforme pas les deux sites éventuellement liés par un lien en réels partenaires.

Il ne nous semble pas que ce type de pratique par e-mailing doive être mis en place pour tenter d'échanger des liens avec d'autres sites web. Il n'est certainement pas nécessaire de contacter des centaines de sites pour obtenir des liens efficaces. Quelques sites, voire quelques dizaines de sites, peuvent suffire. Mais il faut bien les traiter, aller voir leur contenu et essayer de leur faire une vraie proposition, qui leur profite autant qu'à vous. Oubliez les e-mailings stériles et impersonnels, fixez-vous des objectifs réalisables sur un nombre de sites limités et faites du travail très personnalisé. Vos résultats n'en seront que meilleurs.

Visez la qualité plutôt que la quantité

On l'a dit, les moteurs sont aujourd'hui plus sensibles à la qualité des liens pointant vers vos pages plutôt qu'à leur quantité. Bien entendu, rien ne vous empêche de tenter de coupler les deux notions.

Toujours est-il qu'actuellement, il vaut mieux avoir 10 sites populaires (disposant d'un bon indice de popularité) pointant vers vos pages que 100 pages perso (le terme n'a rien de péjoratif ici) très peu populaires et disposant elles-mêmes de peu de liens populaires pointant vers elles.

Répétons-le car cette notion est très importante : ce n'est pas la quantité de liens vers votre site qui est importante mais bien leur qualité.

Encore une fois, fixez-vous des objectifs raisonnables mais efficaces : peu de sites web contactés mais avec de vrais arguments, un contenu intéressant et original et une stratégie d'approche affinée et personnalisée. Laissez la triche à ceux qui n'ont pas d'imagination et de contenu…

Prenez en compte le PageRank des sites contactés

Comment trouver les sites web les plus intéressants pour leur proposer un échange de liens ? Tout simplement en contactant les plus populaires. Et comment trouver ceux qui sont les plus populaires ? Il s'agit certainement de ceux qui ont le meilleur PageRank. Vous pouvez en trouver la liste notamment sur l'annuaire de Google (*http://directory.google.com/* ou *http://directory.google.fr/*), sagement classés par ordre de PageRank décroissant de leur page d'accueil.

Vous recherchez, par exemple, les sites les plus populaires dans le domaine des moteurs de recherche ? Vous les trouverez ici : *http://directory.google.com/Top/World/Fran%C3%A7ais/Informatique/Internet/Recherche/Moteurs_de_recherche/*.

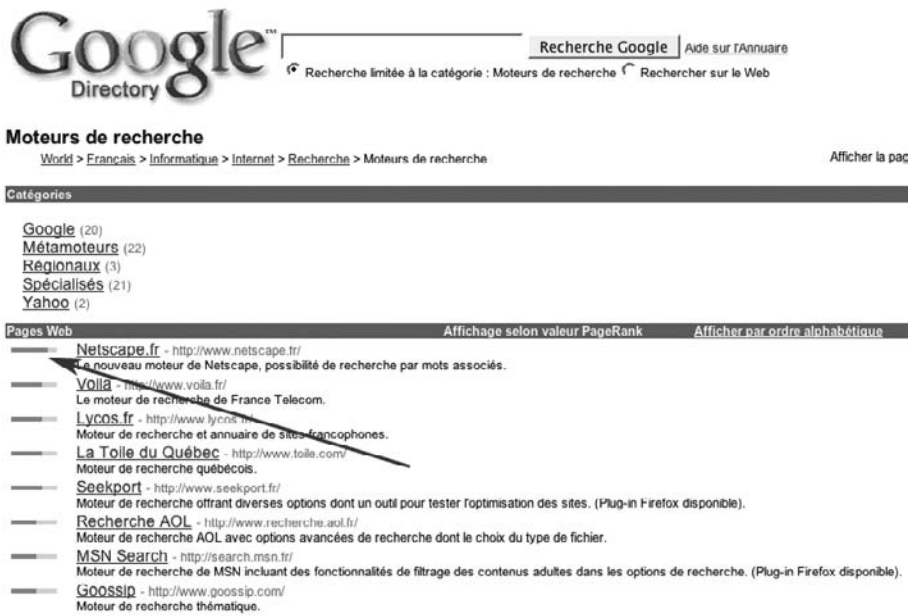


Figure 5-8

Affichage du PageRank dans l'annuaire de Google

La barre verte, à gauche, précise le PageRank du site en question. Mieux il est classé, meilleure est sa popularité. Ce sont donc avec les sites web présents en tête de liste qu'il vous faudra discuter pour échanger des liens (veillez bien à ce que le classement s'effectue

« Pages affichées selon le classement Google PageRank », comme indiqué dans la barre verte).

Quoi qu'il en soit, limitez-vous à un certain nombre de sites (une vingtaine peut paraître un nombre intéressant, mais tout dépend bien entendu de votre ambition sur le Web et de votre domaine d'activité) et faites-en une liste détaillée avant de commencer vos démarches.

Par ailleurs, sachez qu'on n'est jamais si bien servi que par soi-même. Créez, si vous en avez la possibilité, des liens vers votre nouveau site depuis des sites existants qui vous appartiennent.

Exemple : nous avons créé, sur toutes les pages d'accueil des sites du réseau Abondance, des liens vers les différents sites du réseau. Exemple en figure 5-9 sur la page d'accueil du site Abondance.com, en bas de page.

Figure 5-9

Footer des pages du site Abondance.com

Par l'intermédiaire de cette astuce, qui sert également à orienter les internautes vers les autres sites du réseau, dès qu'un nom de domaine est créé dans le réseau Abondance, nous ajoutons un lien sur chacun des sites du réseau et nous constatons que le site correspondant se retrouve très rapidement dans l'index de Google en profitant du PageRank des autres sites (voir chapitre 8). Comme il est normal de faire de la publicité offline pour votre site web (mention sur vos cartes de visite, vos papiers à en-tête, les étiquettes de vos produits), il est également logique de signaler vos autres sites sur chacune de vos sources d'information. Et si celles-ci disposent d'un bon PageRank, c'est encore mieux.

Ne spammez jamais !

Ne vous avisez pas de franchir la frontière entre signalement et spam (lien caché, lien sur des pages bidons, dans des *layers*, etc.), vous pourriez vous en mordre les doigts assez rapidement…

Vous pouvez également utiliser des logiciels comme PR Weaver (*http://www.prweaver.fr/*) qui affichent le PageRank des résultats de Google. Tentez donc des accords d'échanges de liens avec ces sites, sauf si ce sont des concurrents bien sûr. Mieux : faites également une requête de type « link: » sur ces sites pour voir qui pointent vers eux et tentez également des échanges de liens avec ces nouveaux sites identifiés.

Utilisez la fonction « sites similaires »

Autre façon de trouver des sites intéressants : utiliser les fonctionnalités permettant de rechercher des sites similaires. Sur Google, cela se fait au travers du lien « Pages similaires » dans les résultats du moteur ou de la syntaxe « related: ». Exemple : *related:www.abondance.com.*

Attention cependant : ces fonctions ne sont pas forcément très efficaces sur des domaines très – ou trop – pointus ou sur des adresses de sites ayant un PageRank assez faible. Ces outils fonctionnent sur la base de l'analyse des liens entre les sites sur le Web. Un site peu populaire (car peu connu même s'il est très pertinent) sera peu pointé et donc l'analyse par ce biais en sera assez peu efficace. Ne soyez donc pas trop précis dans vos recherches sur les sites similaires et prenez en compte comme site de départ une source d'information la plus populaire possible.

Prenez en compte la valeur du PageRank du site distant

Autre critère à prendre en compte : tentez le plus possible de demander des échanges de liens à PageRank (PR) égal ou proche. Plus le PR de la page sur laquelle vous désirez obtenir un lien sera élevé, plus populaire sera la page en question. Un PR de 6 ou 7 peut déjà être considéré comme excellent.

En revanche, si votre page n'a qu'un PR de 2, il vous sera difficile de demander un échange de liens avec un site dont la page a un PR de 7. Une disproportion trop forte des PR peut faire capoter votre demande. En effet, si votre PR est bien plus faible que celui du site à qui vous demandez un échange, il est facile de constater qu'il va vous apporter beaucoup plus que ce que vous pouvez lui proposer en termes d'échange de popularité. Un échange de ce type n'est donc pas logique et bien proportionné : vous êtes en position de faiblesse. À vous de voir ce que vous pouvez apporter pour compenser la différence de PR. Du contenu ? Des prestations publicitaires ? De la visibilité ? Un partenariat ? Nous vous laissons imaginer tout cela. Il semble donc assez difficile de proposer un simple échange de liens si les deux sites en question sont par trop disparates au niveau du PR. Un bon échange de liens se fera toujours en « gagnant-gagnant » et tout écart de PR devrait être compensé d'une façon ou d'une autre.

Au contraire, si vos PR sont assez proches, faites-en un argument et indiquez au webmaster contacté que vous avez un bon PR et que cela peut être bénéfique d'échanger un lien avec son site qui aura tout à y gagner. Tout comme vous !

Une stratégie qui fonctionne très bien, lorsque vous avez un site qui contient un contenu de qualité (et c'est sans aucun doute le cas du vôtre), est de proposer de l'échange de contenus avec le site distant.

Tous les sites web et portails recherchent du contenu de qualité pour leurs visiteurs. N'hésitez donc pas à proposer le vôtre, directement issu de vos pages ou créé de façon spécifique pour eux, aux sites avec qui vous désirez échanger des liens. Vous voyez ainsi que l'échange de liens constitue non seulement un art, mais aussi un vrai travail…

Un exemple de partenariat efficace et pérenne

C'est ce que nous avions fait avec le site Abondance quelques mois après son lancement. Des accords avaient notamment été passés avec les sites Voila.fr et Journaldunet.com pour leur écrire des articles sur le référencement et la recherche d'information. En échange, chaque page de ces sites contenant un article affichait un lien vers le site Abondance.com. L'échange, ici, était logique, même si à cette époque, on ne parlait encore que très peu de PageRank : celui d'Abondance étant plus faible que celui de Voila et du Journal du Net, une compensation a été effectuée en livraison de contenus, équilibrant l'accord. Les liens sur ces deux sites ont ensuite fait énormément pour la popularité du site Abondance, que ce soit de façon directe (les gens venaient en lisant les articles et en cliquant sur le lien proposé) ou indirecte (au travers de l'indice de popularité sur les moteurs).

Il est également possible de proposer de nombreux partenariats avec le site distant : affiliation, information « cobrandée » (sous l'égide des deux sites), concours, etc. La seule limite est l'imagination et la complémentarité entre les deux sites. Mais nous avons la faiblesse de penser que seuls les vrais partenariats durent et sont profitables pour tous.

Paid linking : bonne ou mauvaise idée ?

Comment trouver des sites partenaires influents dans leur secteur d'activité et acceptant de placer des liens vers les pages d'un autre site ? L'idée émise par certains professionnels serait d'acheter ces liens.

L'état des lieux

Le phénomène est de plus en plus important aux États-Unis. Des plates-formes comme ReviewMe (*http://www.reviewme.com/*) ou Text Link Ads (*http://www.text-link-ads.com/*) proposent de mettre en relation acheteurs et vendeurs. Et quelques recherches autour des mots-clés « pagerank », « for sale » et « links » permettent en quelques minutes d'obtenir une liste de nombreux sites qui monnayent leurs liens au *prorata* du PageRank de leurs pages web, de façon ouverte...

En Europe, le commerce de liens semble beaucoup plus discret. Certains sites mettent parfois en vente une partie de leur espace sans que cela soit clairement identifié comme tel. L'emplacement peut être sous forme d'un encart publicitaire (du type Google AdSense) en lien direct ou sous forme de pages dédiées exclusivement au(x) partenaire(s).

Quels sont les prix pratiqués ?

Il apparaît extrêmement difficile de donner un prix exact car cela dépend de nombreux facteurs. En tout état de cause, la prestation doit être prise dans la durée (minimum 6 ou 12 mois) pour que le lien puisse avoir un réel impact.

Le prix dépend également de la pertinence du site, du PageRank de la page qui contiendra le lien ou de l'emplacement de ce lien (sur tout le site ou sur seulement une page).

Le prix de départ peut commencer à une petite dizaine d'euros mensuels pour finir à plusieurs milliers d'euros mensuels.

Certains outils ont développé des interfaces pour connaître le montant d'un lien comme Text Link Ads (*http://www.text-link-ads.com/link_calculator.php*, non opérationnel au moment où ces lignes sont écrites). On peut s'apercevoir que le prix varie considérablement en fonction de l'emplacement et du type du partenaire.

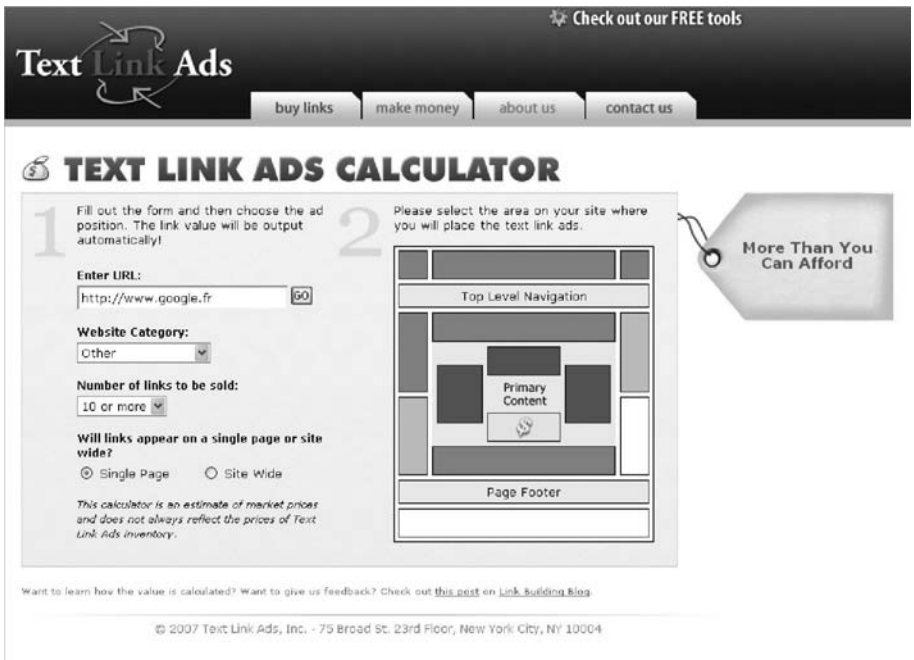


Figure 5-10

Obtenir un lien sur Google semble être hors de prix (« More Than You Can Afford ») pour le logiciel !

Quels sont les risques encourus lors de l'achat d'un lien ?

Matt Cutts, ingénieur technique responsable de la qualité des résultats de Google, détaille sur son blog personnel (*http://www.mattcutts.com/blog/hidden-links/* et *http://www.mattcutts.com/blog/how-to-report-paid-links/*), la position (non officielle) de Google concernant la stratégie de rémunération des liens (*paid links*). En voici quelques extraits détaillés :

« Use the unauthenticated spam report form and make sure to include the word "paidlink" (all one word) in the text area of the spam report. »

Matt Cutts demande à tous les webmasters qui constatent des abus concernant les liens payants de prendre contact directement avec Google *via* le formulaire de spam pour dénoncer la supercherie. Google propose en effet un formulaire spécialisé pour le *Paid*

Linking Report afin de dénoncer toute tentative de ce type. Autant dire que la chasse à l'achat de liens est bien ouverte chez Google (même s'il est très difficile pour le moteur de recherche de détecter ce qui est vendu et ce qui ne l'est pas lorsque ce n'est pas explicitement indiqué sur le site web en question).

Matt Cutts explique également sur son blog que l'utilisation du formulaire *Spams Report* doit servir à tester de nouvelles techniques chez Google.

« If you want to sell a link, you should at least provide machine-readable disclosure for paid links by making your link in a way that doesn't affect search engines. There's a ton of ways to do that. For example, you could make a paid link go through a redirect where the redirect URL is robot'ed out using robots.txt. You could also use the rel=nofollow attribute. »

Suite des explications : si un lien est acheté, il faut le rendre non indexable pour les moteurs de recherche afin de ne pas nuire aux résultats naturels. Pour cela, plusieurs moyens sont possibles, par exemple paramétrer le fichier `robots.txt` en refusant l'indexation par les moteurs de recherche de la page vers laquelle le lien pointe. On peut ajouter la notion `rel="nofollow"` au lien (voir plus loin, la notion de sculpture de PageRank).

« Google is going to be looking at paid links more closely in the future. »

À l'avenir, Google va regarder les liens payants de plus près. Tout un programme !

Même si les arguments lancés par Matt Cutts n'ont rien d'officiel, l'avertissement ne doit pas être pris à la légère. Reste à savoir ce que va faire le moteur lorsqu'il va détecter un lien non pertinent : les rumeurs concernant Google ne sont pas prêtes de s'éteindre… Alors le risque en vaut-il la chandelle ? À vous de voir ! Mais si vous décidez malgré tout de partir sur cette voie, choisissez scrupuleusement vos partenaires pour que les résultats soient à la hauteur des risques pris !

Quelques informations supplémentaires sur le Paid Linking

Voir également ce post récent à ce sujet sur le blog du site Search Engine Watch :

http://blog.searchenginewatch.com/blog/070514-153234

Attention aux pages des sites distants et de votre site

N'oubliez pas que la notion de PageRank s'applique à chaque page d'un site, pas au site web de façon globale. Une page interne de votre site aura, la plupart du temps, un PR plus faible que celui de la page d'accueil, même si ce n'est pas absolument obligatoire.

Dans vos tractations, prenez donc en compte à la fois le PR de la page de votre site sur laquelle vous désirez créer un lien vers le site distant, mais également le PR de la page du site distant sur laquelle vous voudriez voir apparaître un lien vers votre page. Leurs PR respectifs devraient être très différents de ceux des pages d'accueil. Attention aux discours du type « Mon site a un PageRank de 6 » qui doit être traduit ainsi : « Mon site a une page d'accueil de PR 6 ». Mais est-ce bien sur cette page que le lien sera mis en place ?

N'oubliez pas également que le PR de votre page sera fonction du PR de la page pointant vers vous, mais également du nombre de liens que cette page contient. N'hésitez donc pas à choisir, sur le site distant, une page ne proposant pas trop de liens sortants. Bien choisir les pages qui vont échanger des liens, c'est également tout un art…

Créez « une charte de liens »

Si vous avez lu les pages précédentes, vous savez certainement que le texte des liens pointant vers vos pages est important pour votre « réputation ». Aussi, un texte ainsi libellé : « Découvrez un site web sur les moteurs de recherche » amènera un poids très fort à la page distante (pointée par le lien) pour l'expression « moteurs de recherche » (contenu textuel du lien).

Si vous le pouvez, n'hésitez donc pas à créer une « charte des liens » sur votre site, en indiquant au site distant une liste de mots-clés importants pour votre activité qu'il peut, s'il en a la possibilité bien sûr, intégrer dans le texte de ses liens ou mieux, lui fournir directement le code HTML du lien (voir figure 5-11). Bien sûr, ne l'obligez pas à le faire, mais il y a de fortes chances pour qu'il apprenne quelque chose à cette occasion et qu'il vous en soit reconnaissant.

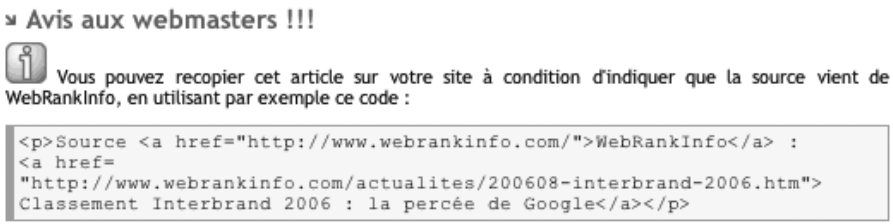


Figure 5-11

Le site WebRankInfo (http://www.webrankinfo.com/) fournit aux webmasters le code HTML à insérer dans les pages pour réaliser un lien.

Bien entendu, proposez de faire la même chose avec le site distant en lui demandant ses mots-clés les plus importants afin de les inclure dans vos liens.

Suivez vos liens

Vous avez obtenu de nombreux liens vers vos pages web ? Bravo ! Mais ne vous endormez pas sur vos lauriers. Aucune situation n'est établie. Aussi :

- Vérifiez bien, à intervalles réguliers, que les liens pointant sur votre site, notamment depuis des sites populaires, existent encore et qu'ils n'ont pas été effacés ou modifiés à l'occasion d'un remaniement des sites distants.

- Attention aux adresses de vos pages : si elles changent et que des liens pointaient vers elles, n'oubliez pas de réagir en conséquence (redirections, messages, etc.).
- Vérifiez bien, grâce à la syntaxe « link: » de Google ou « linkdomain: » de Yahoo!, la liste des pages web ayant mis en place un lien vers vous.
- Vérifiez également, dans vos statistiques, les sites web qui drainent le plus de trafic vers vos pages, notamment dans la rubrique « urls referrers » ou « référents » que votre hébergeur ou votre logiciel de statistiques doit normalement vous proposer. Dans cette liste doivent apparaître les sites avec lesquels vous avez noué un échange de liens. Si certains génèrent beaucoup de trafic, fidélisez-les pour que le partenariat continue, se renforce et soit profitable pour les deux parties. Si le trafic généré par d'autres est faible, essayez de comprendre pourquoi et réagissez en conséquence.
- Faites une veille sur les nouveaux sites apparaissant dans votre domaine d'activité et sur votre métier. Remettez-vous ensuite à l'ouvrage pour obtenir de nouveaux liens…

Pour résumer, suivez bien vos liens, suscitez de nouveaux échanges avec des sites populaires et vous devriez vivre des jours de webmasters heureux.

Des liens triangulaires plutôt que réciproques

Souvent, l'échange de liens s'effectue entre deux pages A et B sous la forme « A pointe vers B qui pointe vers A ». Ce type d'échange est détecté par les moteurs de recherche et n'est pas obligatoirement optimisé. Nous préconiserons plutôt une autre forme d'échange, certes plus complexe à mettre en œuvre, mais bien plus efficace en faisant intervenir une troisième page : l'échange en triangle du type « A pointe vers B qui pointe vers C qui pointe vers A ».

Figure 5-12

Échange de liens « classique » entre deux pages

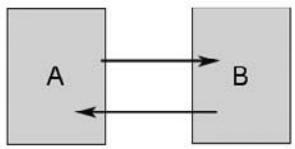
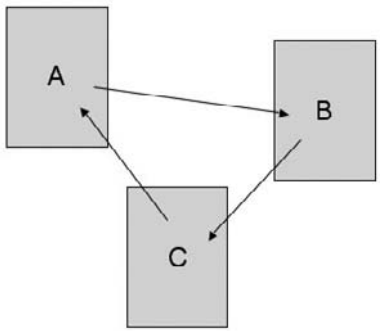


Figure 5-13

Échange de liens « en triangle » entre trois pages



La page C peut correspondre (idéalement) à un troisième site, à une page interne de A ou à une page interne de B, à votre convenance. Cette « triangularisation » des liens fait perdre l'aspect de « réciprocité » de l'échange et améliore son efficacité.

Privilégiez le lien naturel en soignant la qualité de votre site

Voici un moyen quasi infaillible de générer des liens de qualité vers un site web : créer un site original au contenu de haute qualité. Dans ce cas, les liens vers vos pages se créeront de façon naturelle, quasi instantanée, le bouche à oreille jouant son rôle (et Dieu sait si le bouche à oreille est important sur le Web !). Votre site deviendra alors populaire en quelques mois et votre positionnement sur les moteurs de recherche en sera le reflet… Oui, nous nous répétons, mais « Le contenu est votre capital » est une devise qui doit tout le temps rester à votre esprit.

N'oubliez pas cependant d'effectuer des échanges de liens loyaux et honnêtes (oubliez les *links farms*, systèmes d'échanges de liens massifs ou les liens payants) et, pour ce qui est des liens, privilégiez la qualité à la quantité. Votre référencement, votre positionnement et leur pérennité ne s'en porteront que mieux !

Le linkbaiting ou comment attirer les liens grâce à votre contenu

Connaissez-vous le linkbaiting (ou *link baiting* ou encore *link-baiting* comme on le voit parfois écrit) ? En fait, il s'agit d'un concept dont le nom a bien sûr été inventé par les Américains mais qui est en lui-même vieux comme le Web. Il s'agit d'appâter (*bait* signifie « appât », « amorce ») les webmasters par du contenu provoquant une envie naturelle de créer un lien vers lui. Il s'agit donc ici de partir à la « pêche aux liens » en créant du contenu, sur votre site web, qui va générer du lien vers lui, et lui donner une popularité importante !

L'avantage du linkbaiting est double :

- les liens créés de façon quasi instantanée vont générer du trafic direct vers votre site ;
- ces liens vont améliorer votre popularité (PageRank), et donc, à un moment ou à un autre, favoriser vos classements dans les pages de résultats des moteurs de recherche.

L'idée principale du linkbaiting est de créer du contenu qui attirera les liens de façon naturelle, plutôt que de se lancer dans des campagnes d'échanges de liens, qui sont souvent très aléatoires et fastidieuses, ou dans de longues inscriptions dans des annuaires de seconde zone afin de générer du lien parfois bien peu efficace. Le principal inconvénient des campagnes classiques d'échanges de liens est qu'il est difficile de contacter un site souvent inconnu pour ne lui proposer rien d'autre que ce type de citation textuelle d'une page vers une autre et réciproquement. Le taux de perte est parfois monstrueux. De plus, chaque webmaster d'un site un tant soit peu populaire est aujourd'hui assailli par ce type de demandes (parfois bien folkloriques d'ailleurs) auxquelles il ne prête plus attention. Quant au message du type « Échangeons des liens et nous pourrons ainsi détourner les algorithmes des moteurs (et plus si affinités) », nous vous laissons seul juge de leur

teneur éthique comme vu précédemment. Il est en effet temps maintenant de passer à une approche plus professionnelle et surtout plus efficace de la notion de netlinking.

De nombreuses façons de faire du linkbaiting

Le concept du linkbaiting n'est donc pas nouveau, même si le terme peut le faire croire. Pourtant, l'idée est intéressante car il existe plusieurs pistes de réflexion que l'on peut prendre en considération pour créer du contenu attractif. Pour cela, il vous faut mettre en place des articles hameçons qui vont servir à ferrer le lien – pour reprendre l'analogie avec la pêche, suscitée par le terme *bait*. Par exemple :

- Créer un concours de sites web, de blogs, de personnalités, etc., qui va faire parler de lui. Ou bien créer un sondage de type « Oscars » ou « Awards » pour élire un site ou une personne. Ou alors une enquête en ligne sur un sujet « dont on parle »… De même, un jeu peut avoir le même impact s'il est original (mais attention, il existe de très nombreux jeux sur le Web).
- Une interview exclusive d'une personnalité peut également susciter du lien si elle contient des informations intéressantes, nouvelles et pertinentes.
- Publier sur son site un article « coup de poing », ou « coup de gueule », bien argumenté cependant, qui nourrira le « buzz » (ou bouche à oreille numérique), notamment dans la blogosphère. Les contenus assez personnels du type « Je ne suis pas d'accord avec… », s'ils sont bien écrits et bien argumentés, peuvent créer un fort afflux de liens… Attention cependant : si vous écrivez ce type d'article, vous devez bénéficier d'un minimum de crédibilité sur Internet pour faire valoir vos arguments. N'oubliez pas, également, de rester courtois et constructif dans vos arguments sans diffamer quiconque. L'attaque gratuite ne vous amènera pas obligatoirement des retours très positifs. Sachez également que la polémique peut rapidement amener des réactions plus ou moins violentes d'autres internautes. On récolte parfois ce que l'on sème… Mais une opinion tranchée, intelligente, sur un sujet porteur est toujours la bienvenue si, là encore, elle est bien argumentée.
- Il est toujours possible de reprendre en français des rumeurs lues sur des sites anglais (ou entendues dans des salons parisiens). Mais attention, ce ne seront que des rumeurs donc il faudra peut-être les présenter comme telles…
- La traduction en français d'articles anglais intéressants peut également être une piste valable. N'oubliez pas, bien entendu, de citer votre source et d'indiquer clairement que votre travail s'est limité à de la traduction… Une demande d'autorisation à l'auteur de l'article initial peut également être nécessaire si ce n'est polie…
- Les articles comparatifs de plusieurs produits, sites web ou autres sont toujours assez prisés et repris sur le Web. De même, les articles ayant des titres chiffrés du type « 10 façons de… » ou « 12 choses à ne pas faire pour… » ont une bonne côte et « sonnent bien ». Le post de Danny Sullivan intitulé *25 Things I Hate About Google* (*http://blog.searchenginewatch.com/060313-161500*) en est un excellent exemple. À noter que

son pendant, *25 Things I Love About Google* (*http://blog.searchenginewatch.com/060313-161501*), ne l'est pas moins…

- Publier une étude gratuite sur un sujet « chaud » est une excellente façon de faire parler de soi. Mais encore faut-il que l'étude en question soit fouillée et pertinente… L'article *Search Engine 2009 - Ranking Factors* (*http://www.seomoz.org/article/search-ranking-factors/*) ou l'étude du référencement des sites web du secteur du champagne (*http://blog.ranking-metrics.fr/analyse-champagne/*) par Ranking Metrics ont généré de nombreux liens dans le microcosme du référencement. À vous de vous en inspirer !

- Une liste de ressources intéressantes peut également provoquer un bon *linkbait*. Exemple : une page reprenant la liste des blogs « officiels » de Google voire affichant, pour chacun d'entre eux, les trois derniers billets publiés, ou des fiches descriptives. La rubrique « Ressources » du site Abondance (*http://ressources.abondance.com/*) a ainsi généré près de 8 000 liens vers elle en proposant des listes de sites dans le domaine des moteurs de recherche et du référencement. Il peut s'agir également d'une compilation d'articles ou de chiffres intéressants sur un thème donné.

- Un article humoristique sur un sujet donné peut également vous servir… ou vous desservir, car n'oubliez jamais que la notion d'humour n'est pas toujours partagée de la même façon par les internautes. Une requête qui donne un résultat amusant sur un moteur, des images dénichées ici ou là, un site amusant et original que vous chroniquez, etc. Tout cela est bon pour faire parler de vous !

- L'idéal reste d'obtenir ou de trouver un scoop sur un sujet précis et porteur, mais ce n'est bien entendu pas si facile que cela. Le scoop peut être textuel, photographique ou sous la forme d'un podcast, d'une vidéo ou autre… L'explosion de sites comme YouTube ou Dailymotion peut être un tremplin pour faire connaître rapidement une vidéo humoristique par exemple… ou une « manœuvre politique », comme on l'a vu récemment en France avec certains candidats à l'élection présidentielle qui ont vu des extraits vidéo de certains de leurs meetings se répandre rapidement sur la Toile. La vidéo du rappeur Kamini (*http://www.kamini.fr/*) reste également un incroyable exemple de linkbaiting : une simple vidéo postée en ligne, il y a quelques mois de cela, a valu au site du rappeur français plus de 15 000 liens ! C'est ce que l'on appelle le gros lot.

- Certaines dates sont plus propices au linkbait telles que le 1er avril. Un poisson d'avril bien senti peut rapidement générer de nombreux liens… Le site Googland (*http://www.googland.com/*) que nous avions créé pour le 1er avril 2004 a attiré en quelques semaines plus de 10 000 liens vers lui, ce qui n'est pas négligeable…

- Bien entendu, un site original, voire amusant, sera rapidement l'objet d'un linkbait qui peut s'avérer fulgurant. Exemple avec le site GoogleFight (*http://www.googlefight.com/*) qui a vu ses backlinks se développer très rapidement (près de 140 000 liens vers le site) et susciter des articles dans des revues comme *USA Today* ou des passages sur Canal Plus.

- La création d'un *widget* – concept très en vogue actuellement – original peut également vous valoir une bonne couverture de « buzz » (ou bouche à oreille électronique)…

Ces gadgets sont très faciles à concevoir, aussi n'hésitez pas à exercer votre créativité dans ce domaine.

Bref, soyez inventif, créatif, amusant, informatif, voire agressif (dans le bon sens du terme) et vous devriez voir les liens venir vers vous de façon quasi automatique. L'avantage de ce type de promotion est qu'elle ne coûte rien, hormis le temps passé à la créer, et peut-être à initier sa connaissance *via* le Web. L'inconvénient majeur est qu'il faut être inventif, créatif, amusant, informatif, voire agressif…

On pourrait parfois penser que ces techniques sont un peu sensibles et que les moteurs de recherche les voient d'un mauvais œil et pourraient les considérer comme du *spamdexing*. Pourtant, Matt Cutts, le spécialiste du référencement chez Google, en parle sur son blog (voir encadré page suivante) et semble les voir plutôt d'un bon œil. Encore faut-il, bien sûr, que la stratégie mise en place soit de bonne qualité et que les ficelles ne soient pas trop grosses. Comme toujours en référencement, tout est le plus souvent une question de bon sens.

L'une des idées principales du linkbaiting est de suivre une procédure qui fera en sorte de créer le plus de backlinks possible dans un minimum de temps.

1. Imaginez, créez le contenu et mettez-le en ligne.
2. Contactez éventuellement des « meneurs d'opinion » dans le domaine traité pour leur signaler ce contenu et les engager à en parler, notamment sur leur blog. Jouez fin : ne demandez pas à ce qu'ils créent un lien vers votre article ni même qu'ils en parlent expressément, mais indiquez que vous leur écrivez uniquement pour leur signaler cette information. À eux de juger de ce qu'ils veulent en faire… Mais proposez-leur éventuellement d'intervenir dans les commentaires de votre blog (si cette possibilité existe).
3. Vérifiez que votre contenu est bien référencé sur les différents moteurs comme Google, voire Google News et Google Blog Search, Technorati, etc. Laissez faire ensuite le système… Si votre linkbait est bon, tout devrait s'enchaîner rapidement. Si la pompe est difficile à amorcer, travaillez à améliorer votre contenu ou à en créer un autre.

Il est clair que le linkbaiting n'est pas un concept neuf mais plutôt une « stratégie d'attraction de liens » au travers d'un contenu adapté et de qualité. Et s'il peut rappeler aux éditeurs de sites web que cette qualité du contenu est le capital, cela ne peut qu'être bénéfique ! Cela ne peut tirer le Web que vers le haut. Et il y a fort à parier qu'à l'instar de ce bon M. Jourdain, nombreux sont ceux, parmi vous, qui font du linkbaiting depuis de nombreuses années, sans le savoir !

Cette notion révèle également une chose importante : aujourd'hui, éditer un site web « plaquette » n'est plus suffisant. Il faut véritablement créer du contenu pour être visible sur le réseau. S'asseoir et attendre les visiteurs est une stratégie totalement dépassée. Là aussi, si le linkbaiting nous le rappelle, c'est également une bonne chose !

Mais c'est aussi une stratégie à prendre avec des pincettes car vouloir générer du lien à tout prix peut vite se retourner contre son auteur. Le linkbaiting ne peut être profitable

que s'il est appliqué avec parcimonie et, surtout, avec intelligence et qualité ! Il ne faut surtout pas l'oublier, sous peine de s'en mordre les doigts très rapidement…

Quelques liens sur le linkbaiting

– L'un des premiers articles en français traitant du linkbaiting : *Linkbait & linkbaiting : une tentative de traduction* de Jean-Marie Le Ray

http://adscriptum.blogspot.com/2006/02/linkbait-linkbaiting-une-tentative-de.html

Il existe également de très nombreux articles en anglais abordant ce sujet. Voici une liste de ceux qui nous ont semblé les meilleurs. Comme quoi, écrire des articles sur le linkbaiting est en soi du linkbaiting :

– Matt Cutts, porte-parole SEO de Google, parle du linkbaiting :
http://www.mattcutts.com/blog/seo-advice-linkbait-and-linkbaiting/

– *Link Baiting & Effective Link Building* de Rob Sullivan
http://www.searchenginejournal.com/index.php?p=2797

– *Linkbaiting for Fun & Profit* de Rand Fishkin
http://www.searchenginejournal.com/index.php?p=2541

– *Use Link Bait to Catch Better Rankings* de Bill Hartzer
http://www.searchengineguide.com/hartzer/006598.html

– *Linkbaiting with Attack* de Darren Rowse
http://www.problogger.net/archives/2005/11/10/linkbaiting-with-attack/

– *Linkbaiting, How Hard Is It?* de Joe Balestrino
http://www.isedb.com/db/articles/1515/

– *Learning About Linkbaiting* de Jennifer Laycock
http://www.searchengineguide.com/searchbrief/senews/009242.html

– *2007 Guide to Linkbaiting: The Year of Widgetbait?* de Nick Wilson
http://searchengineland.com/070118-074231.php

– *Linkbait Articles & Is It Linkbait Or Link Bait?* de Danny Sullivan
http://searchengineland.com/070117-145217.php

– *Are You Linkbaiting The Right Audience?* de Eric Ward
http://searchengineland.com/070129-130949.php

– *The Links That Can't Be Baited* de Eric Ward
http://searchengineland.com/070205-093613.php

– *Andy Hagans' Ultimate Guide to Linkbaiting and Social Media Marketing* de Andy Hagan
http://tropicalseo.com/2007/andy-hagans-ultimate-guide-to-link-baiting-and-social-media-marketing/

– The Enormous Linkbait List
http://www.cornwallseo.com/search/index.php/2007/02/09/the-enormous-linkbait-list/

Si après avoir lu tous les articles de cette liste, vous ne devenez pas un spécialiste mondial du linkbaiting, c'est à désespérer de tout.

Link Ninja : de la recherche de liens classique

Le Link Ninja est une forme dérivée de linkbaiting où l'auteur du contenu n'attend pas que les liens se créent d'eux-mêmes. Dans cette stratégie, le webmaster va aller négocier un lien auprès d'autres sites web ou créer des liens par exemple dans des commentaires de blogs, dans des forums, etc. Bref, faire du « Link Ninja » n'est rien moins que de rechercher des liens vers son propre contenu. Loin d'être un concept révolutionnaire... On aime bien inventer des mots compliqués pour décrire des choses simples sur Internet...

Pour résumer

Voici quelques conseils pour bien optimiser l'indice de popularité de vos pages :

- Choisissez bien les pages qui vont pointer vers votre site : fort PageRank, faible nombre de liens sortants.
- Attention aux redirections et programmes d'affiliations ou autres bannières publicitaires dont les liens ne sont pas pris en compte dans la plupart des cas. Vérifiez bien que les liens créés vers vous sont « en dur », donc sans redirection.
- Faites des échanges de liens à popularité égale et compensez d'une façon ou d'une autre (en évitant le côté financier) en cas de déséquilibre.
- Attention également aux liens JavaScript, Flash ou autres, mal pris en compte par les moteurs.
- Ne vous faites pas berner par des offres d'augmentation artificielle de la popularité. Visez de vrais liens issus de partenariats forts.
- Créez une charte de liens en donnant des indications aux webmasters distants sur la meilleure façon de faire un lien vers votre site.
- Privilégiez la qualité à la quantité.
- Nouez de vrais partenariats, prévus pour durer.

La sculpture de PageRank

On entend souvent dire, dans *le Landerneau* du référencement, que les liens sortants n'ont pas d'importance pour l'optimisation d'une page en vue d'une meilleure visibilité sur les moteurs de recherche.

Cela n'est pourtant pas totalement vrai. En effet :

- On peut penser que l'analyse des liens sortants d'une page et d'un site joue un rôle non négligeable dans la notion de « TrustRank » (voir plus loin dans ce chapitre). En quelque sorte, l'étude des liens sortants émanant de sites de référence permettrait de calculer une « note de confiance » qui réduirait le *spamdexing* dans les résultats des moteurs de recherche. Par ailleurs, des algorithmes comme ceux de Ask.com, basés sur la définition de communautés web, utilisent également la notion de lien sortant. On pourrait encore trouver de nombreux exemples de l'importance de ces liens sortants en termes de référencement d'une manière macroscopique. Leur influence est donc loin d'être nulle dans ce domaine.

- Les liens sortants sont clairement utilisés par Google pour choisir les Sitelinks qu'il affiche dans ses pages de résultats.
- Enfin, la notion de popularité (au sens du PageRank de Google) tient grandement compte des liens sortants (qui deviennent automatiquement des liens entrants pour la page dont il faut calculer le PageRank) pour calculer ce critère de pertinence.

On l'a vu, pour calculer la popularité d'une page B, Google tient compte de la « capacité de vote » des pages qui pointent vers elle pour attribuer un « jus de lien » partagé entre toutes les pages pointées. Ainsi, la notion de lien sortant est très importante, non pas directement pour la page que vous désirez utiliser, mais pour celle qui « pointe » vers elle...

De la bonne utilisation des liens sortants dans une stratégie de référencement

Ainsi, il sera très important de tenir compte de ce système de calcul, notamment sur votre page d'accueil. En effet, de par le fait que votre *homepage* reçoit la plupart des backlinks (liens entrants) du Web, c'est très souvent elle qui bénéficie de la meilleure popularité, de façon assez logique.

Aussi, cette « note de popularité » doit être transmise avec parcimonie aux autres pages vers laquelle elle pointe, sous peine de dilution trop forte de cette « capacité de vote » ou « Link Juice ». Ainsi, si la page d'accueil de votre site contient 100 liens sortants, chaque page distante ne recevra qu'un centième de la popularité de cette homepage. Des miettes, en quelque sorte... Pour optimiser votre transfert de popularité, vous devrez donc « faire la chasse » aux liens sortants et optimiser au mieux leur nombre afin de le réduire au maximum pour empêcher le plus possible la « fuite de popularité ». Voici quelques astuces pour y parvenir...

Éviter les liens multiples

On voit souvent des présentations d'articles comportant un lien sur :

- le titre ;
- le contenu ;
- une vignette image.

Si trois liens sont proposés, c'est excellent pour l'internaute (il peut cliquer où il le désire pour aller lire l'article en question). En revanche, l'optimisation pour les moteurs n'est pas excellente : vous proposez trois liens au lieu d'un seul vers une même page, ce qui dilue d'autant la popularité fournie par la homepage...

D'autant plus que, si l'on en croit le site SEOmoz (*http://www.seomoz.org/blog/results-of-google-experimentation-only-the-first-anchor-text-counts*), seul le premier lien rencontré dans le code HTML vers une page donnée est pris en compte par le moteur de recherche. S'il s'agit du lien mis en place sur l'image, vous perdez ainsi toute notion de réputation (donnée par le contenu textuel du lien et qui n'est pas totalement compensée par l'attribut `alt` de l'image).

Figure 5-14

Site de l'Équipe (http://www.lequipe.fr/) : plusieurs liens pointant vers la même page sont proposés (un sur le titre, un sur le chapô et un autre sur l'image).

Deux solutions s'offrent alors à vous pour vous permettre de choisir quel lien sera suivi par le moteur : soit vous n'indiquez qu'un lien unique qui prend en compte les trois zones, si cela est techniquement possible sur votre site et par rapport à votre charte graphique, soit vous mettez deux liens en `rel="nofollow"`, qui est un microformat (*http://microformats.org/wiki/rel-nofollow*) initié par Google en 2005 (et suivi depuis par Yahoo! et Bing) et qui signifie que le lien en question ne sera pas pris en compte par le moteur de recherche. On appelle cela de la « sculpture de PageRank » ou *PageRank Sculpting* en anglais.

Ce lien sera ainsi suivi et analysé par les moteurs de recherche :

```
<a href="http://www.votresite.com/">Texte du lien</a>
```

Alors que celui-là sera ignoré (avec un bémol cependant, comme on le verra plus tard) :

```
<a href="http://www.votresite.com/" rel="nofollow">Texte du lien</a>
```

Dans ce cas, marquez (en `rel="nofollow"`) les liens sur l'image mais aussi sur le titre (s'il n'est pas descriptif), ou sur le texte/chapô (s'il n'est pas trop long), à votre convenance. Dans tous les cas, il faudra choisir le lien à donner « en pâture » au moteur de recherche (donc celui qui n'est pas en `rel="nofollow"`) comme celui qui propose le texte le plus explicite par rapport au contenu de la page distante (notion de « réputation », encore une fois).

De la même façon, si plusieurs liens sont proposés dans votre page vers un même document, il vous faudra mettre en place la même procédure. Un seul lien suffit pour les moteurs de recherche, il n'est donc pas nécessaire d'en fournir plus...

Autre solution si vous ne désirez pas utiliser le `rel="nofollow"` : certains utilisent des liens JavaScript, non suivis par les moteurs. Une autre façon de rendre ce type de lien invisible (mais attention aux internautes qui ont désactivé JavaScript sur leur navigateur...). D'autres méthodes (comme l'obfuscation qui permet de « cacher » du code

source dans un programme) peuvent également être employées pour cacher le code en question...

Attention cependant, Google a modifié en 2009 sa façon de prendre en compte le paramètre `rel="nofollow"` des liens (*http://www.mattcutts.com/blog/pagerank-sculpting/*).

- *AVANT* : si la page A comportait 10 liens sortants, dont 5 en `dofollow` (le `dofollow` est l'inverse du `nofollow` : dans ce cas, on n'indique pas d'attribut `nofollow`, c'est donc un lien « classique ») et 5 en `nofollow` : sa bouteille de jus de lien était partagée en 5 parts égales distribuées auprès des 5 pages pointées par des liens en `dofollow`. Ça va, vous suivez ?
- *AUJOURDHUI* : si la page A contient 10 liens sortants, 5 en `dofollow`, 5 en `nofollow` : sa bouteille de jus de lien est maintenant partagée en 10 auprès des 5 pages pointées par des liens en `dofollow`. Beaucoup moins intéressant…

Attention donc aux liens émanant de pages web à fort PageRank mais contenant un nombre très important de liens sortants (en `nofollow` ou pas)... La popularité distribuée peut être semblable à des miettes...

Éviter les destinations non pertinentes

Certaines pages de votre site sont certainement très peu pertinentes pour les moteurs de recherche : conditions générales de vente, informations sur l'entreprise, mot du PDG, crédits, etc. Ces données sont souvent présentes sous forme de liens dans le *footer* de vos pages. Là encore, on peut penser que ces liens « diluent » votre popularité et n'apportent pas grand-chose en termes de référencement ou à l'internaute qui les trouverait dans les pages de résultats d'un moteur. Aussi, il peut être envisagé de marquer ces liens en `rel="nofollow"` pour qu'ils ne soient pas suivis par les moteurs de recherche.

Cela ne signifie pas pour autant qu'il faut les désindexer, mais plutôt qu'il ne faut leur fournir que la popularité « qu'elles méritent ». Ces pages secondaires pourront tout à fait être trouvées par les *spiders*, par exemple, au travers du plan du site ou d'autres pages. En revanche, sur vos pages les plus populaires, on peut estimer qu'elles « polluent » votre optimisation...

Il en sera de même avec des liens vers des sites externes avec qui vous n'avez pas de réelle affinité ou de contrat de partenariat. Certains liens peuvent ainsi devenir « transparents » pour les moteurs et cela aidera à diminuer la dilution du transfert de votre popularité... Ceci dit, les nouvelles règles édictées par Google en termes de sculpture de PageRank (voir précédemment) rendent ces pratiques moins efficaces...

Par ailleurs, n'oubliez pas que les moteurs de recherche savent aujourd'hui faire la part des choses entre un lien « structurel » (qui sert à la navigation) et un lien « contextuel » (dans le contenu éditorial de la page) (*http://blog.abondance.com/2007/04/entre-lien-structurel-et-contextuel.html*). Soignez donc en priorité les liens contextuels (au sein même de vos textes) car il y a de fortes chances pour que leur poids soit bien plus fort dans les algorithmes de compréhension de vos contenus par les moteurs de recherche... Et ceci convient pour les liens internes tout comme pour les liens externes...

Notez que ces techniques de *PageRank Sculpting* sont largement débattues actuellement, certains se posant la question d'un éventuel *spamdexing* lorsqu'elles sont utilisées. Selon nous, si leur utilisation est légère et non exagérée, elle sont tout à fait « recevables » et ne devraient pas poser de problème. Bien entendu, tout restera dans la nuance et le bon sens pour ne pas « dépasser des limites que l'on ne connaît pas », comme souvent dans le domaine du référencement... Ceci dit, on a vu précédemment que Google n'a pas tardé à revoir sa politique de gestion et d'analyse des liens dans ce domaine, on peut donc estimer que la sculpture de PageRank n'est pas obligatoirement une technique très efficace aujourd'hui pour peaufiner sa popularité.

Conclusion

Il va de soi que cette optimisation des liens sortants n'est intéressante que si vous avez déjà bien optimisé vos pages par ailleurs : balise `<title>`, titre éditorial en `<h1>`, code HTML optimisé, contenu riche et de qualité, etc. De plus, la structure et l'indexabilité de votre site sont aujourd'hui des valeurs essentielles (voir chapitre 7) à prendre en compte. La « chasse aux liens sortants » n'est donc utile que pour affiner une optimisation, il s'agit d'une « finition » qui ne peut être mise en place que si le « gros œuvre » est déjà terminé... Ne l'oubliez pas !

Cinq règles d'or pour vos liens sortants

Voici, pour terminer, cinq « règles d'or du lien sortant » pour votre référencement qui résument bien le contenu du paragraphe que vous venez de lire :

- Règle 1 : proposer le moins possible de liens sortants aux moteurs de recherche depuis une page populaire.
- Règle 2 : ces liens sortants doivent le plus possible pointer vers un contenu traitant du même domaine, de la même thématique que la page qui les contient.
- Règle 3 : les textes des liens sortants doivent être le plus explicites et descriptifs possible (notion de « réputation »).
- Règle 4 : les liens sortants les plus importants sont ceux qui sont contenus dans la partie éditoriale de vos pages (contrairement aux liens de navigation). Soignez-les du mieux possible !
- Règle 5 : n'exagérez pas la « chasse aux liens sortants ». Restez dans les limites du bon sens et tout devrait bien se passer...

Le TrustRank ou indice de confiance

Google s'est fait connaître sur le Web au travers de son célèbre PageRank, étudié en détail dans les pages précédentes. Mais, depuis quelques années, on parle de plus en plus d'un autre indice, baptisé TrustRank, et dont Google se servirait pour mesurer la confiance que l'on peut avoir dans un site web donné, sur la base de critères humains et automatiques. Qu'en est-il exactement de ce fameux TrustRank, dont le nom vient en fait d'ingénieurs de chez Yahoo! ?

Définition du TrustRank

On a commencé à entendre parler de ce nouveau critère de classement au travers d'un article rédigé en mars 2004 par deux chercheurs de l'université de Stanford et intitulé *Combatting Web Spam With TrustRank* (*http://www.vldb.org/conf/2004/RS15P3.PDF*). Les chercheurs (qui travaillaient pour Yahoo!) proposaient de créer une liste de sites de confiance (*trusted sites*) et d'accorder une attention particulière aux liens du Web provenant de ces sites, partant de l'hypothèse qu'un lien issu d'un site de confiance pointe généralement vers un autre site de confiance.

Le TrustRank, au cœur de ce nouveau système, désigne ainsi l'indice de confiance accordé à un site web, et ce signal se propage d'un site à l'autre de façon décroissante. Plus on est « loin » du site de confiance initial (au sens du nombre de clics nécessaire pour y arriver), plus le TrustRank diminue.

Pour la petite histoire, Google a déposé en 2005 le nom de marque TrustRank mais il n'aurait pas de rapport direct avec le projet développé par les chercheurs de Yahoo!. Dans une vidéo publiée sur YouTube en novembre 2007, on voit notamment Matt Cutts expliquer que le TrustRank de Google n'a rien à voir avec celui décrit dans l'article de Stanford (*http://fr.youtube.com/watch?v=p8mUXQzwEvs*).

Néanmoins, bien qu'il n'ait jamais été officiellement reconnu par Google, le TrustRank semble être au cœur de son algorithme : pour être bien classé dans Google, il est en effet important d'obtenir des liens depuis des sites « de confiance ».

Ceci est confirmé dans plusieurs explications données dans l'aide en ligne de Google (*http://www.google.fr/support/webmasters/bin/answer.py?answer=66356&topic=15263*) :

« Le classement de votre site dans les résultats de recherche Google est en partie basé sur l'analyse des sites qui comportent des liens vers vos pages. La quantité, la qualité et la pertinence de ces liens sont prises en compte pour l'évaluation de votre site. Les sites proposant des liens vers vos pages peuvent fournir des informations sur l'objet de votre site, et peuvent indiquer sa qualité et sa popularité. »

« Le PageRank tient également compte de l'importance de chaque page qui « vote » et attribue une valeur supérieure aux votes émanant de pages considérées comme importantes. Les pages importantes bénéficient d'un meilleur classement PageRank et apparaissent en haut des résultats de recherche. » (*http://www.google.com/corporate/tech.html*)

Quelques informations ont filtré ici et là sur la façon dont Google pourrait déterminer qu'un site est un site de confiance ou non. Citons :

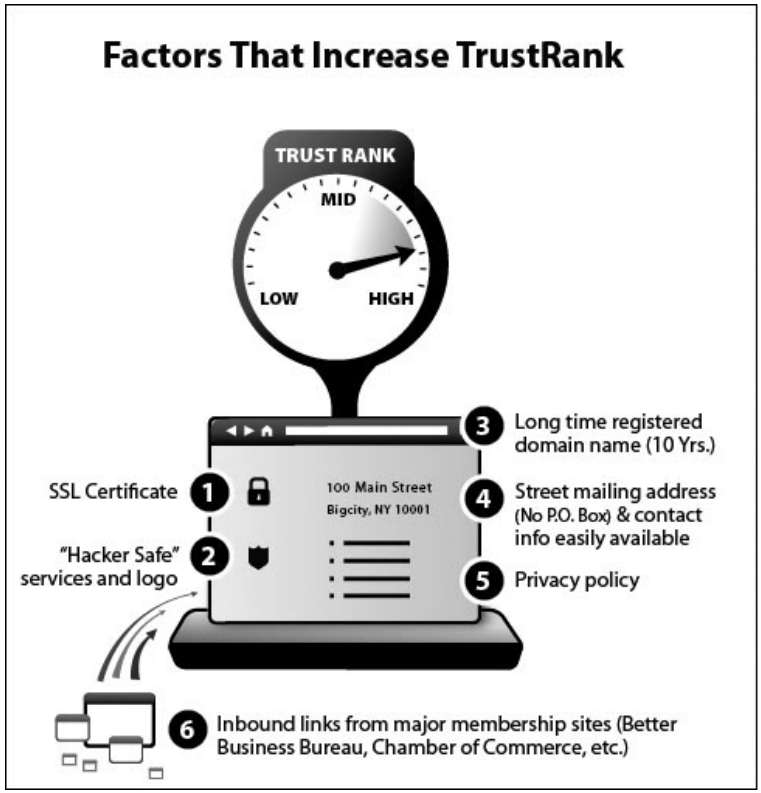
- les données Whois ;
- la durée d'enregistrement du nom de domaine ;
- la politique relative à la vie privée ;
- les informations de contact ;
- l'affichage de l'adresse postale de l'entreprise sur le site ;

- la taille du site (en nombre de pages) ;
- le trafic ;
- le niveau de sécurisation du site ;
- d'éventuelles certifications apportées par des organismes officiels ;
- l'ancienneté des liens acquis ;
- une note attribuée par un être humain, chez Google, chargé de recenser un certain nombre de « sites incontournables » dans certains domaines thématiques (ce qui expliquerait, par exemple, l'omniprésence de Wikipedia sur le moteur de recherche) ;
- d'autres paramètres ?

La liste peut être longue. Ceci peut être résumé dans le schéma de la figure 5-15, proposé par le site *http://searchengineoptimization.elliance.com*.

Figure 5-15

Quelques facteurs influençant le TrustRank



Un brevet, déposé par Google, parle également d'une notion de « TrustRank » pour son site Google News, indiquant ces critères pour l'établissement d'un « indice de confiance » dans une source d'informations sur l'actualité :

- le nombre d'articles produits par la source ;
- la longueur moyenne des articles ;
- la « couverture » de la source ;
- la réactivité de la source (*breaking score*) ;
- un indice d'utilisation (en nombre de clics sur cette source) ;
- une opinion humaine sur la source ;
- une statistique extérieure d'audience telle que Media Metrix ou Nielsen Netratings ;
- la taille de l'équipe ;
- le nombre de bureaux ou agences différents de la source ;
- le nombre d'entités nommées « originales » citées par la source (personnes, organisations, lieux) ;
- l'étendue (*breadth*) et le nombre de sujets couverts par la source ;
- la diversité internationale ;
- le style de rédaction, en termes d'orthographe, de grammaire, etc.

Sources : Brevet *Systems and Methods for Improving the Ranking of News Articles* déposé par Google, selon le blog Technologies du Langage : *http://aixtal.blogspot.com/*.

Le TrustRank sous toutes ses formes

Comme nous l'avons vu au début de ce chapitre, le terme de TrustRank a été originellement développé par Yahoo! et c'est un principe qui est en réalité utilisé par de nombreux moteurs. La notion de « site de confiance » est en effet à la base de nombreux algorithmes, et c'est une méthode de base pour identifier les sites les plus intéressants à présenter dans les résultats des moteurs de recherche.

Le moteur Ask.com utilise par exemple un « ExpertRank », qui mesure la popularité d'un site vis-à-vis de pages « expertes » (c'est-à-dire de sites de confiance) :

« L'algorithme ExpertRank du moteur Ask assure la pertinence des résultats de recherche en identifiant les sites les plus fiables et les plus respectés sur le Web. Avec la technologie de recherche Ask, il ne s'agit pas d'être le plus grand : il s'agit d'être le meilleur. Notre algorithme ExpertRank ne s'arrête pas à la popularité des liens (c'est-à-dire au classement des pages en fonction du nombre brut de liens pointant vers une page particulière) pour mesurer la popularité des pages dites expertes sur un sujet de recherche donné. À cet effet, on parle de popularité thématique. L'identification des sujets (également nommés « clusters »), des pages expertes sur ces sujets et de la popularité des millions de pages les plus fiables en la matière – à l'instant précis où vous lancez votre recherche – demande de nombreuses analyses supplémentaires non pratiquées par les autres moteurs de recherche. Résultat : une pertinence inégalée proposant souvent une

touche rédactionnelle unique par rapport aux autres moteurs de recherche. » (*http://sp.fr.ask.com/fr/docs/about/asksearch.shtml*)

Autre outil de classement, un « BrowseRank » a été développé en juillet 2008 par les chercheurs de Microsoft. Cette fois, la notion de popularité ne se base plus sur la qualité des liens entrants mais plutôt sur le comportement des visiteurs : temps passé sur le site, nombre de liens cliqués, nombre de visiteurs... (*http://actu.abondance.com/2008/07/browserank-larme-de-microsoft-pour.html*). Pour obtenir ces données stratégiques, Microsoft utilise la barre d'outils de son moteur Bing (et on imagine bien que Google fait de même avec sa propre barre d'outils).

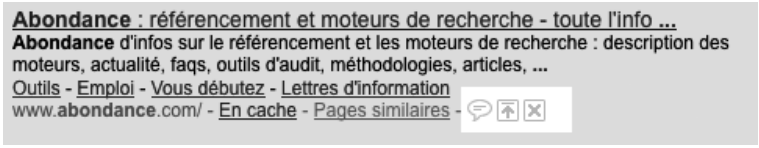
Le TrustRank en 2009/2010

Le TrustRank (s'il existe) est sans doute désormais une combinaison de nombreux facteurs. Cela fait longtemps que les moteurs de recherche ne se basent plus uniquement sur le nombre de liens pointant vers un site mais qu'ils prennent aussi en compte la qualité du contenu, l'origine des liens, le comportement des utilisateurs...

La fin d'année 2008 a été également marquée par l'arrivée d'un élément important pour le référencement : l'apparition des résultats personnalisés dans Google, avec le système baptisé « SearchWiki » (*http://googleblog.blogspot.com/2008/11/searchwiki-make-search-your-own.html*).

Figure 5-16

Les pictogrammes du service SearchWiki de Google



Concrètement, ce système, que l'on peut visualiser sur la figure 5-16, permet à chaque utilisateur possédant un compte Google de supprimer, enrichir ou commenter les résultats de recherche. Il est donc fortement probable que Google va tenir compte de l'avis des internautes pour le classement des résultats de recherche. En conséquence, la notation TrustRank pour Google pourrait bien s'enrichir de la directive suivante : « un site est un site de confiance s'il a reçu beaucoup de votes de la part des utilisateurs ».

Ceci compliquera certainement le travail des webmasters et des référenceurs : les techniques de « triche » seront de moins en moins fructueuses, pour faire de son site un site de confiance, il faudra désormais plaire aux internautes...

Dans les années qui viennent, un site de confiance ne sera pas forcément un site ayant été validé par d'autres sites de confiance. Il sera aussi un site ayant obtenu un grand nombre de visiteurs et de votes grâce un buzz ou un marketing viral efficace sur le Web.

L'autre évolution du Web qui se profile depuis plusieurs années, la recherche universelle, pourrait également de changer la donne. En effet, la notion de TrustRank risque de prendre en compte non seulement un site web, mais également ses « produits dérivés » tels

que les images, les vidéos, les documents... Nous en parlerons longuement au chapitre suivant.

Il est certain qu'assurer un positionnement sur différents médias permettra de profiter de la recherche universelle et il est possible qu'un site possédant, par exemple, beaucoup de vidéos indexées dans YouTube verra son TrustRank revu à la hausse. Une bonne stratégie sera alors de positionner des documents multimédias dans des sites de confiance, ceux-ci assurant ensuite un vote de qualité vers le site web principal.

Des portails comme Flickr, par exemple, permettront sans doute d'augmenter son TrustRank. En effet, il s'agit d'un site où les images peuvent être notées et commentées par les internautes (ce qui ajoute le critère de vote humain vu précédemment) et où le contenu est soumis à modération (voir à ce sujet les règles communautaires de Flick : *http://www.flickr.com/guidelines.gne*).

C'est donc l'aspect humain qu'il faudra peut-être privilégier à l'avenir : créer du contenu à destination des internautes, plutôt que penser son site pour les moteurs de recherche. Avec les nouvelles règles du TrustRank, un site en *full Flash* pourra, par exemple, avoir une chance plus importante de ressortir dans les moteurs...

Les autres critères…

Nous avons vu dans ce chapitre les principaux points que vous aurez à optimiser pour rendre vos pages compatibles et réactives par rapport aux critères de pertinence des moteurs. Mais un outil comme Google utilise plus d'une centaine de critères. Il en existe donc bien d'autres ayant un poids plus faible que ceux décrits dans ces pages. Il faut quand même les avoir en tête lors de l'élaboration de vos pages. Tous ne sont pas connus, mais en voici une liste non exhaustive avec quelques hypothèses s'y rapportant.

- **Validité W3C.** Pensez à passer vos pages au validateur du W3C (*http://validator.w3.org/*). Un code HTML propre peut grandement aider à sa compréhension par les robots des moteurs. Il est surtout important de vérifier, grâce à l'outil de validation du W3C, que rien ne va gêner l'analyse du code HTML par les moteurs de recherche (exemple : une balise ouverte mais non fermée). Ainsi, une page web peut tout à fait retourner des dizaines, voire plus, d'erreurs au niveau de ce validateur et être tout à fait *search engine friendly* si aucune erreur relevée ne permet l'analyse du code...
- **Date de création du document.** Plus un document est ancien, plus son poids pourrait être considéré comme fort.
- **Nombre de pages du site.** Plus un site contient de pages, plus il peut être considéré comme étant « de confiance », comme on l'a vu dans l'étude du TrustRank.
- **Fréquence de mise à jour des pages.** Plus une page est mise à jour, plus elle est considérée comme pertinente.

- **Historique du site.** Google pourrait analyser la vie d'un site et notamment le taux de création de nouvelles pages, de modification de documents dans le temps, etc. Rappelez-vous que Google horodate toutes les informations qu'il trouve...
- **L'ancienneté des liens acquis**. Plus un lien est créé depuis longtemps, plus il a de poids.
- etc.

Il ne s'agit ici que d'une sélection de critères complémentaires possibles. On en trouve bien d'autres dans des articles et forums sur le Web. Certains sont intéressants, d'autres complètement délirants… mais ceux évoqués dans ce chapitre (et le précédent) sont, et de loin, les plus importants. Et au moins, vous pouvez être sûr qu'ils fonctionnent…

6

Référencement multimédia, multisupport

Le concept de « recherche universelle » (affichage, dans les pages de résultats, de documents émanant de plusieurs bases de données différentes : Web, images, vidéos, actualités, cartographie, etc.), adopté par de nombreux moteurs, a rendu plus forte encore la nécessité de penser son référencement de façon globale et « multimédia ». Aussi, l'optimisation de fichiers autres que les pages web *stricto sensu* (en langage HTML) devient une stratégie essentielle pour obtenir une meilleure visibilité sur ces outils. Dans ce chapitre, nous allons voir comment optimiser une image, une vidéo ou un fichier PDF, entre autres, pour lui faire gagner des positions dans les résultats des moteurs et obtenir un bon référencement... Le référencement local, basé sur des outils comme Google Maps, ou dans l'actualité ne sera pas oublié. Bref, nous allons traiter ici de tout ce qui ne concerne pas le référencement Web proprement dit, mais qui prend pourtant une part de plus en plus prépondérante dans une stratégie de visibilité globale sur les moteurs de recherche. Commençons par les images...

Référencement des images

Le monde de la recherche d'images sur le Web est devenu un marché en constante progression et les outils de recherche y deviennent de plus en plus nombreux, de plus en plus pointus dans leurs investigations. Le référencement de fichiers images est donc également devenu un domaine sur lequel la plupart des webmasters doivent aujourd'hui se pencher. En effet, Google, par exemple, propose parfois des images – en première

position ou non – de son moteur de recherche web sur certaines requêtes. Exemple sur la figure 6-1 sur le mot-clé « désert ».

Figure 6-1

Recherche universelle : ajout d'images dans les résultats par Google

De plus, la recherche d'images est une préoccupation majeure des internautes actuels. Selon Hitwise (*http://actu.abondance.com/2006-20/trafic-google.php*), si la recherche web contribuait en 2006 à hauteur de 80 % au trafic de Google, son moteur Google Images était proche des 10 % et constituait son deuxième outil le plus utilisé... En 2007, comScore (*http://blog.abondance.com/2007/12/le-palmars-des-outils-de-google.html*) a fourni des chiffres plus récents sur la base d'une étude entre les mois de novembre 2006 et 2007 aux États-Unis. Cette étude confirme clairement le phénomène, Google Images ayant connu une croissance de 35 % sur cette période...

Il semble donc clair que, si l'indexation de vos images ne vous pose pas de problème stratégique (car il peut arriver que, pour des raisons de droits d'auteur notamment, on ne désire pas qu'une image soit indexée...), il vous faut vous pencher au plus vite sur leur optimisation afin de gagner du trafic sur votre site.

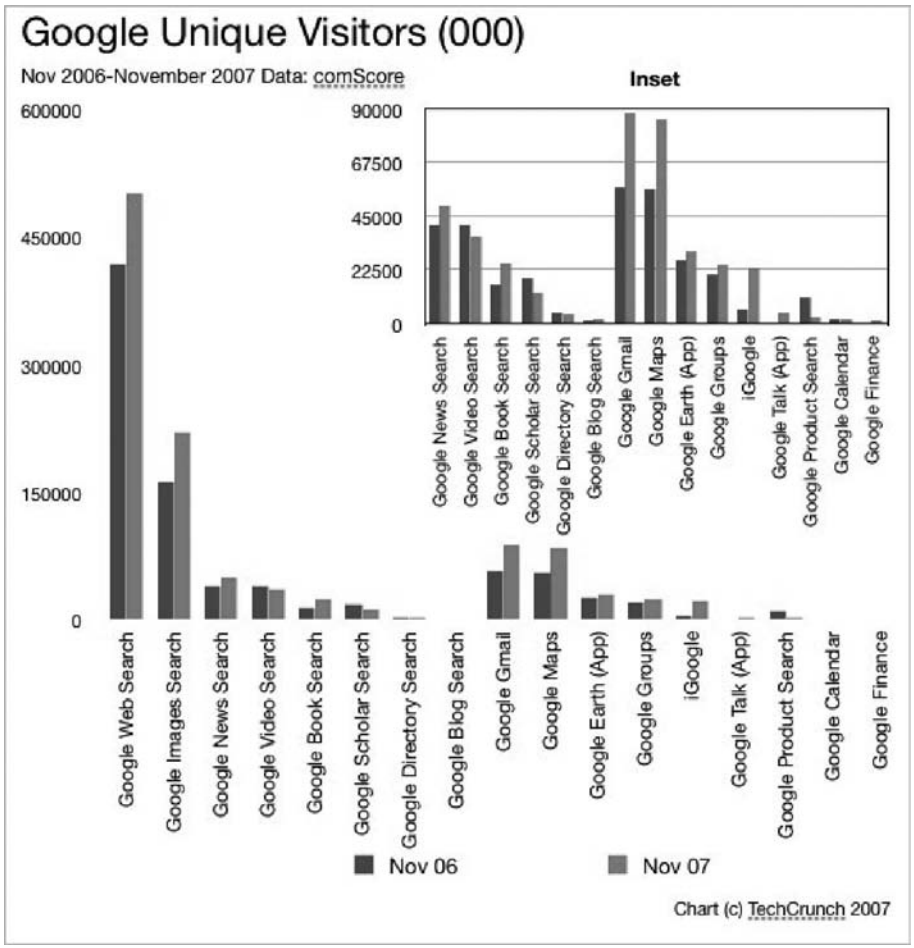


Figure 6-2

La recherche d'images, deuxième préoccupation des internautes utilisant les outils Google

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

```
<img src="http://www.votresite.com/images/nom-de-l-image.jpg" width="245" height="152"
➥alt="texte alternatif décrivant l'image">
```

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

Critère numéro 1 : le 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.

Pour séparer les mots, utilisez le tiret (-) ou l'underscore (_), en préférant le tiret si vous avez le choix (toujours ce vieux débat – pas totalement tranché aujourd'hui – sur le fait que Google prenne en considération ou non l'underscore comme séparateur dans les URL, voir chapitre 4...).

En revanche, évitez les mots « collés ». En d'autres termes, préférez `nicolas-sarkozy.jpg` à `nicolassarkozy.jpg`. L'utilisation d'un séparateur (tiret ou underscore) va « détacher » plusieurs mots dans une même expression et les rendre « réactifs » à une recherche.

Par ailleurs, il ne semble pas qu'il y ait actuellement une limite en termes de nombre de caractères pour le nom de l'image à partir du moment où ce nom reste dans les limites du « raisonnable »...

Critère numéro 2 : le format de l'image. Préférez les formats GIF (`.gif`) ou JPEG (`.jpg`). Certains moteurs comme Google peuvent indexer d'autres formats (PNG pour la plupart) mais le « tronc commun » pris en compte par tous les moteurs d'images reste le GIF et le JPEG. Un autre format risquerait d'exclure vos images de l'index (le PNG semble cependant de plus en plus « compatible » avec la plupart des moteurs).

Si votre image est de grande taille et représente une photo, préférez le format JPEG qui accepte mieux la réduction que le GIF. Et, pour être affichée dans les pages de résultats sous la forme d'une vignette, votre image sera presque obligatoirement réduite...

N'oubliez pas non plus d'indiquer la largeur (`width`) et la hauteur (`height`) de l'image, cela aidera le moteur de recherche à déterminer le format de l'image.

Critère numéro 3 : le texte alternatif. Ce texte, présent dans l'option `alt="..."`, est très important pour les moteurs de recherche. On a vu (chapitre 4) qu'il aidait au référencement web, mais sa réelle utilité concerne le référencement des images. 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. N'hésitez 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/nicolas-sarkozy-luxembourg.jpg" width="145"
➥height="152" alt="Discours de Nicolas Sarkozy au sommet européen du Luxembourg -
➥10 décembre 2009">
<img src="http://www.votresite.com/images/cathedrale-strasbourg.gif" width="100"
➥height="40" alt="Entrée ouest de la cathédrale de Strasbourg, Alsace, France">
```

Les textes ainsi insérés ne sont pas affichés dans 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. C'est également un excellent moyen (tout

comme dans le nom de l'image), de proposer une version non accentuée de certains mots-clés.

Ne dépassez pas pour cet attribut les 10 mots descriptifs de l'image (insérer dans cette zone des mots-clés n'ayant pas de rapport avec l'image ne sera pas très utile, faites en sorte qu'ils soient réellement en adéquation avec ce que propose l'image), cela suffira amplement... Évitez également de « profiter » de ces images pour les « truffer » de mots-clés, vous risquez d'être pénalisé par les moteurs pour *spamdexing*...

Si une image sert uniquement à la charte graphique et au design de votre page, indiquez un attribut `alt` vide (conformément au standard W3C) :

```
<img src="http://www.votresite.com/images/pixel-unique.gif" width="1" height="1" alt="">
```

Il existe également deux autres attributs, nommés `name` et `title` :

```
<img srcC="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 d'images, on dispose de moins d'informations sur ces deux derniers champs, mais, selon nos tests, il ne semble pas qu'ils soient pris actuellement en compte par les moteurs majeurs (mais l'attribut `title` sert, en revanche, à certains navigateurs pour afficher du texte lors du passage de la souris sur l'image, il n'est donc pas inutile par ailleurs...). N'y passez pas trop de temps, même si leur présence ne pénalisera pas vos images (mais cela peut éventuellement jouer sur le poids de votre page si elle contient beaucoup d'images) :

```
<img src="http://www.votresite.com/images/nom-de-l-image.jpg" width="145" height="152"
➥alt="Description 1 de l'image" name="image" title="Description 2 de l'image">
```

Dans tous les cas, privilégiez l'option `alt` pour décrire vos images.

Critère numéro 4 : le texte du lien. N'hésitez pas à indiquer, si l'image est affichée en cliquant sur un lien (notamment pour obtenir une version agrandie de l'image), des mots-clés de recherche importants dans le texte du lien pointant sur l'image. Exemple :

```
Visualiser <a href=http://www.votresite.com/images/nicolas-sarkozy-luxembourg.jpg"
➥target="_blank">une image de Nicolas Sarkozy au sommet européen du Luxembourg le
➥10 décembre 2007.</a>
```

Ce qui donnera comme résultat :

> Visualiser une image de Nicolas Sarkozy au sommet européen du Luxembourg le 10 décembre 2007.

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 :

```
Visualiser <strong><a href=http://www.votresite.com/images/nicolas-sarkozy-luxembourg.jpg"
➥target="_blank">une image de Nicolas Sarkozy au sommet européen du Luxembourg le
➥10 décembre 2007.</a></strong>
```

Ce qui donnera :

> Visualiser **une image de Nicolas Sarkozy au sommet européen du Luxembourg le 10 décembre 2007.**

Critère numéro 5 : le texte « autour de l'image ». Si vous en avez la possibilité, proposez, le plus proche possible de l'image dans le code HTML, du texte explicitant l'image. Exemple :

```
<img src="http://www.votresite.com/images/cathedrale-strasbourg.gif" width="100"
➥height="40" alt="Entrée ouest de la cathédrale de Strasbourg"> 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.
```

Les moteurs de recherche se servent, pour rechercher leurs images, non seulement du contenu de la balise `<img>` (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...

Par exemple, une bonne façon de proposer du texte sera d'afficher une légende au format textuel pour toutes vos images. Bien entendu, cette légende sera en rapport direct avec le contenu descriptif de l'image.

Évitez, en revanche, d'afficher l'image dans une fenêtre pop-up lancée grâce à un JavaScript non compatible avec les moteurs de recherche, ce qui aurait pour effet immédiat de rendre inaccessibles vos images par les *spiders*.

Critère numéro 6 : le texte de la page. Si l'indexation d'images est cruciale pour votre activité (photographe, artiste, etc.), nous vous conseillons de créer une page web par image et d'optimiser cette page de façon « classique » (balises `<title>`, `h1`, URL, réputation, etc.) par rapport au contenu de l'image en question. Il y a de très fortes chances qu'elle soit alors remarquablement indexée par les moteurs de recherche. Évidemment, cette stratégie n'est pas recommandée si les images de votre site sont avant tout des illustrations d'un contenu éditorial.

Utiliser l'outil Google Image Labeler

Google propose également son outil Google Image Labeler (*http://images.google.com/imagelabeler/*), qui lui permet de mieux « comprendre » vos images en les faisant tagguer par des internautes pris au hasard (*http://actu.abondance.com/2006-36/google-image-labeller.php*), comme on le voit sur la figure 6-3. Votre référencement ne s'en portera que mieux...

Figure 6-3

Google Image Labeler, un étonnant concept pour mieux tagguer vos images...

Désindexer ses images

Si l'on peut avoir envie de voir les images de son site indexées par Google et ses compères, 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. Cette procédure est décrite à la fin du chapitre 9.

L'avenir : reconnaissance de formes et de couleurs

Nous espérons que les quelques conseils qui se trouvent dans ce chapitre vous aideront à mieux optimiser vos images et à augmenter leur visibilité sur les moteurs de recherche. L'étape suivante, pour ces outils, sera certainement la reconnaissance de textes et de formes dans les images (Google a d'ailleurs déposé dernièrement un brevet à ce sujet : *http://actu.abondance.com/2008/01/un-brevet-sur-la-reconnaissance-de.html*). Un moteur comme Exalead, en France, a fait de nombreux progrès sur ces thématiques depuis quelques mois. Ainsi, dès maintenant, n'hésitez pas à soigner la qualité et la netteté de vos images afin que les formes, les textures, les couleurs et les textes puissent y être reconnus facilement. Comme pour la vidéo (voir ci-après), les futurs moteurs de recherche d'images passeront par ces critères pour effectuer leurs investigations. Facilitez-leur la tâche dès maintenant...

Par ailleurs, n'oubliez pas que les internautes utilisent également énormément des outils comme Flickr (*http://www.flickr.com/*) ou similaires (Webshots – *http://www.webshots.com/*, PBase – *http://www.pbase.com/* ou encore Fotki – *http://www.fotki.com/*) qui peuvent grandement servir à

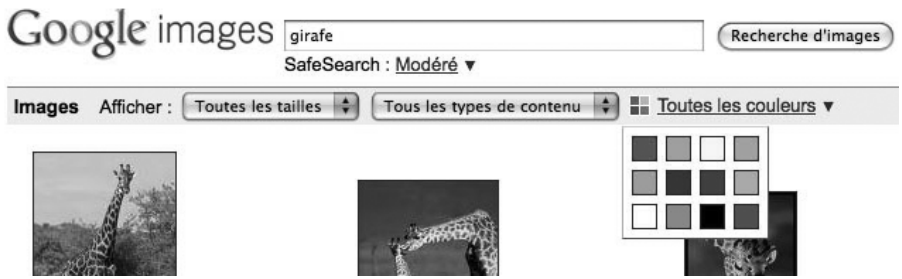


Figure 6-4

Google sait déjà rechercher par couleur, mais également des visages, etc. Sa recherche s'affine mois après mois.

la notoriété de votre site et de vos images sur le Web, notamment grâce à leur système de « tags » additionnels, tout comme les réseaux sociaux de type Facebook (voir plus loin dans ce chapitre). Ne les négligez pas...

Quelques liens sur le sujet...

Voici quelques liens complémentaires que nous vous conseillons de lire, car ils vous donneront quelques conseils supplémentaires sur la meilleure façon d'optimiser vos images pour les moteurs de recherche :

– *Comment optimiser le référencement des images : le guide complet* de Olivier Duffez

http://www.webrankinfo.com/actualites/200709-optimiser-le-referencement-des-images.htm

– *Conseils pour optimiser le référencement de vos images et augmenter votre trafic*

http://www.moteurs-news.com/blog/index.php/2007/01/15/67-conseils-pour-optimiser-le-referencement-de-vos-images-et-augmenter-votre-trafic

– *Optimisation des images pour Internet* (plus orienté « web » que « moteur » mais cela reste intéressant)

http://www.rankspirit.com/optimiser-images-web.php

– *Optimiser les images pour la recherche* de Adam McFarland

http://www.arkantos-consulting.com/articles-referencement/200610/optimisation-images-recherche.php

Et en anglais :

– *Optimizing Images for Search Engines* de Grant Crowell

http://searchenginewatch.com/showPage.html?page=3624327

– *Using Flickr for Search Engine Optimization*

http://www.naturalsearchblog.com/archives/2006/09/24/using-flickr-for-image-search-optimization/

– *Optimizing Images for Search Engines* de Manoj Jasra (contient de nombreux liens vers d'autres articles)

http://manojjasra.blogspot.com/2007/09/optimizing-images-for-search-engines.html

– *Feeling Sandboxed? How You Can Get 53% More Searches with One Twea*

http://www.pearsonified.com/2007/01/get_53_percent_more_searches_with_one_tweak.php

Référencement des vidéos

Le monde de la recherche de vidéos sur le Web est devenu un marché en constante progression et les outils de recherche deviennent de plus en plus nombreux et de plus en plus pointus dans leurs investigations. Le référencement de fichiers vidéo, tout comme pour les images, est donc également devenu un domaine sur lequel la plupart des webmasters doivent aujourd'hui se pencher.

On l'a vu au début de ce chapitre, de nombreux moteurs ont mis en place des systèmes de « recherche universelle » dans leurs pages de résultats.

Figure 6-5

La recherche universelle selon Google : des vidéos sont insérées dans les pages de recherche web (ici sur la requête « Matt Cutts »).

Il est donc clair que les vidéos peuvent aujourd'hui apporter un vrai « plus » en termes de visibilité dans les pages de résultats des moteurs. Raison de plus pour se pencher d'un peu plus près sur leur optimisation, qui est souvent assez proche de ce que l'on peut faire pour les images.

Des recherches incontournables sur les outils dédiés

La recherche de vidéos est aujourd'hui un phénomène qu'il est difficile de contourner, notamment grâce à l'avènement d'outils comme YouTube. Une étude publiée par le prestataire spécialisé Blinkx (*http://www.blinkx.com/seo.pdf*) indique que :

- 123 millions d'Américains ont vu une vidéo en ligne au moins une fois par mois en 2007 ;
- 26 % des internautes qui regardent des vidéos cherchent des vidéos d'actualité une fois par semaine ;
- 26 % des internautes recherchent des vidéos humoristiques ;
- 66 % de ces internautes ont regardé des publicités vidéos et elles ont influencé 44 % d'entre eux ;
- 76 % des internautes ont parlé à un ami d'une vidéo vue sur le réseau.

En France, le phénomène est également devenu important en quelques mois, notamment auprès des jeunes utilisateurs, grands utilisateurs d'outils comme Dailymotion, YouTube, Kewego ou encore Wideo.

Si l'on en croit comScore (*http://www.comscore.com/press/release.asp?press=1723*), 79 % des internautes français âgés d'au moins 15 ans et ayant accès à Internet depuis leur domicile ou leur travail ont visionné une vidéo en avril 2007. Les États-Unis, le Royaume-Uni et l'Allemagne enregistrent respectivement un résultat de 76 %, 80 % et 70 %. L'internaute français adepte du contenu vidéo sur Internet avait en moyenne regardé 64 vidéos en avril 2007 contre 80 au Royaume-Uni et 62 en Allemagne. Aux États-Unis en revanche, la moyenne atteignait 65 lectures vidéos en avril 2007.

En France, les adeptes du contenu vidéo passent 13 % de leur temps en ligne à visionner des vidéos sur Internet...

Figure 6-6

Temps de lecture de contenus vidéos par pays

Lecture de contenu vidéo sur Internet par pays
Avril 2007
Population totale des internautes - Domicile et travail, internautes de 15 ans et plus(1)
Source : comScore Video Metrix

	Part de la population internaute consultant des videos**	% du temps total de connecxion affecte \ à la lecture de video	Nombre lesctures de video par internaute
France	79%	13%	64
États-Unis	76%	6%	65
Allemagne	70%	9%	62
Royame-Unis	80%	10%	80

L'analyse cite également les sites de consultation de vidéos les plus populaires. Parmi les 1,28 milliards de lectures effectuées en France en avril 2007, 22 % ou 285,7 millions de personnes proviennent des sites Google (dont YouTube). Dailymotion suit de très près avec 249,2 millions de lectures en avril 2007. Parmi les autres sites populaires, on retrouve France Telecom avec 23,9 millions de lectures en transit, les sites Iliad/Free.fr avec 17,6 millions de consultations et les sites Microsoft avec 5,3 millions d'accès en avril 2007.

Quand on voit la croissance actuelle de ces utilisations de recherche en France et dans le monde, ces statistiques ne peuvent qu'augmenter fortement dans les mois qui viennent. Difficile, donc, d'ignorer un tel raz-de-marée...

Différents types de moteurs de recherche

Avant de parler des vidéos proprement dites, il est nécessaire de comprendre comment les moteurs de recherche fonctionnent... On peut classer les moteurs actuels et à venir en trois grandes familles :

- Les moteurs de première génération : ils basent leurs algorithmes le plus souvent sur l'analyse des métadonnées (titre, descriptif) fournies lors de la création de la vidéo ou de son téléchargement, ainsi que sur le nom du fichier et éventuellement d'autres données comme le texte dans la page qui lance la vidéo, etc. La majeure partie des moteurs actuels (Google, Truveo, Dailymotion, SingingFish, AltaVista, Yahoo!, etc.) fonctionnent de cette manière.
- Les moteurs de deuxième génération : ils ont mis en place des systèmes de reconnaissance vocale et permettent d'effectuer des recherches à l'intérieur de « ce qui est dit » dans la vidéo. Par exemple, une vidéo affichant le discours d'un homme politique pourra être trouvée car la personne en question énonce les termes de recherche lors de son discours... Blinkx, Google Audio Indexing ou Podzinger, entre autres, font partie de cette famille. Nous en reparlerons dans ce chapitre.
- Les moteurs de troisième génération fonctionneront selon le principe de la reconnaissance de forme. Dans ce cas, les systèmes seront capables de trouver des formes similaires, des couleurs, des textures, de reconnaître des visages, etc., dans les vidéos. Il s'agit certainement de moteurs qui apparaîtront dans les années qui viennent, ne serait-ce que parce que les potentialités de recherche et de publicité sont énormes... Mais de très nombreuses expérimentations sont d'ores et déjà en place dans les laboratoires de recherche...

Quelques moteurs de vidéos

À noter deux listes assez exhaustives de nombreux moteurs de recherche vidéos avec leurs caractéristiques ici :

– *http://www.lightreading.com/document.asp?doc_id=112147*
– *http://web2.econsultant.com/videos-hosting-sharing-searching-services.html*

Comment les moteurs trouvent-il les fichiers vidéo ?

Les moteurs spécialisés dans les vidéos ont, globalement, deux moyens à leur disposition pour trouver des fichiers et créer leur index :

- Comme un *spider* classique, ils suivent les liens trouvés dans les pages web et indexent ainsi les fichiers vidéo identifiés lors de leur navigation.
- Les internautes ont la possibilité, sur la plupart des outils, de charger (*uploader*) leur fichier directement. Il s'agit ici de la « soumission » du fichier comme on le faisait sur les moteurs de recherche web dans les années 1990. Un lien « Envoyer une vidéo » est alors proposé à cet effet, le plus souvent dès la page d'accueil...

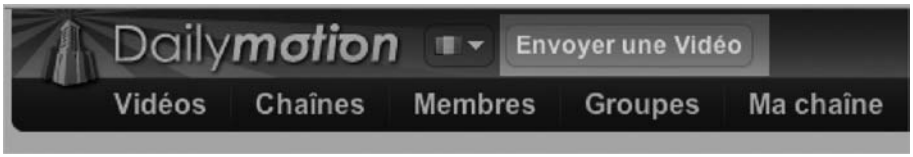


Figure 6-7

Lien permettant d'uploader une vidéo sur le site Dailymotion

L'optimisation des fichiers vidéo

Il est donc nécessaire, aujourd'hui, d'optimiser ses vidéos pour les moteurs de recherche avant de les « envoyer » sur le Web. Comment faire ? Voici une liste de critères à prendre en compte pour arriver à vos fins et obtenir la meilleure visibilité possible, tout en sachant que les techniques d'optimisation n'en sont encore qu'à leurs prémices (et que l'optimisation des vidéos est finalement assez proche de ce qui se fait pour les images) :

Critère numéro 1 : le nom du fichier. Indiquez un nom de fichier qui soit le plus possible en rapport avec le sujet de la vidéo. N'hésitez pas à y ajouter le mot « vidéo » (non accentué) car il semblerait que de nombreuses recherches sur le Web le contiennent. Chaque mot sera enfin séparé des autres par un tiret. Exemples :

- *video-discours-segolene-royal.wmv*
- *video-formation-maitrise-comptabilite.mpg*
- *mon-fichier-video.avi*

Critère numéro 2 : les métadonnées. La plupart des systèmes de création de fichier vidéo permettent d'ajouter des métadonnées, notamment un titre et un descriptif, comme pour un fichier Word ou PDF. N'hésitez pas à remplir ces champs et à être très descriptif en termes de mots-clés. Par ailleurs, si vous changez de format pour un fichier (par exemple, si vous le faites passer de MPEG à WMV), vérifiez que l'utilitaire de changement de

format que vous utilisez traite également les métadonnées, car ce n'est pas toujours le cas et vous risquez, avec certains outils, de les perdre lors de la conversion...

Critère numéro 3 : les caractéristiques techniques. La plupart des moteurs de recherche qui permettent d'*uploader* des fichiers vous donnent des indications et caractéristiques techniques à suivre – ou des préférences – pour vos fichiers (taille maximale, durée, taux de compression, etc.). Exemple pour YouTube (*http://www.google.com/support/youtube/bin/topic.py?hl=fr&topic=16560*) :

- Format : MPEG4 (Divx, Xvid) ;
- Résolution : 640 × 480 ;
- Format audio : MP3 ;
- 30 images par seconde ;
- Limites : durée = 10 minutes, taille = 100 Mo.

N'oubliez pas d'en tenir compte...

Pour les moteurs de recherche de deuxième génération, qui « comprennent » le texte contenu dans la vidéo, soignez particulièrement la qualité de la bande son afin que les systèmes de reconnaissance vocale utilisés par ces outils arrivent à bien décoder les phrases qui y sont énoncées.

Pensez d'ores et déjà aux moteurs de recherche de troisième génération, pour lesquels la netteté de l'image sera primordiale pour bien reconnaître les formes, les couleurs, les textures, etc. Les nouvelles façons d'indexer et de « comprendre » les fichiers arrivent très vite sur le Web. Pensez-y dès aujourd'hui...

Optimisation de l'environnement de la vidéo

Optimiser le fichier lui-même ne suffit pas. En effet, comme pour les images, vous devez faire attention à d'autres points cruciaux pour la bonne prise en compte de vos fichiers par les moteurs :

Critère numéro 4 : les tags. De nombreux moteurs et outils permettent d'insérer, lors de l'*upload*, des *tags* descriptifs de la vidéo et partagés par tous les internautes. N'hésitez donc pas à les utiliser pour décrire au mieux, encore une fois, vos fichiers en quelques termes « bien sentis »... Notamment un titre et un descriptif qui complèteront sans soucis les informations que vous avez éventuellement déjà indiquées dans les métadonnées du fichier lui-même lors de sa création (les mêmes conseils peuvent être édictés). Ils sont très importants pour tous les moteurs actuels. Le travail d'optimisation commencera donc par là... N'hésitez pas non plus à « occuper l'espace » fourni et à indiquer de nombreux mots-clés (notamment la catégorie/rubrique dans laquelle s'inscrit la vidéo) pour décrire vos fichiers. Des mots-clés précis mais aussi d'autres plus génériques sont toujours intéressants.

Figure 6-8

Ajout de tags sur le site de YouTube

Critère numéro 5 : la réputation. Soignez le texte des liens qui vont éventuellement lancer la vidéo lorsqu'on cliquera dessus. Exemple : Vidéo sur le discours de M. Georges Simenon, maire de Trifouillis-les-oies le 22 décembre 2009 à Paris. Ce critère est important pour tous les moteurs actuels, il faut absolument le prendre en compte.

Par ailleurs, il sera intéressant de présenter vos fichiers sous la forme d'une page HTML de présentation par vidéo. Évitez les pages qui présentent plusieurs vidéos les unes à la suite des autres... Un fichier = une page de présentation, c'est la règle pour une bonne optimisation, comme pour les images. Bien évidemment, toutes les méthodes d'optimisation de page web « classiques » devront être appliquées à cette page descriptive... Il s'agit d'ailleurs ici d'une stratégie de référencement qui est plébiscitée par de nombreux acteurs, qui ne désirent pas obligatoirement voir leurs vidéos référencées directement (en *upload*) sur YouTube ou Dailymotion, mais plutôt voir la page, sur leur site, qui présente la vidéo, référencée par Google ou Yahoo!. Car cette page, en plus de la vidéo, présente des bannières publicitaires et d'autres informations qui font venir l'internaute sur le site et en augmente le trafic. Nuance importante qui nous fait revenir à des optimisations HTML plus classiques...

Critère numéro 6 : le texte « autour » de la vidéo. Pensez à proposer une page de description de la vidéo et non pas le fichier seul (exemple : *http://ghq.lejdd.fr/2007/12/11/8-video-kadhafi-invite-une-journaliste-dans-sa-chambre*).

Si vous en avez la possibilité, accompagnez vos vidéos, sur une page web, d'un descriptif précis du contenu du fichier, voire un transcript de ce qui y est dit si vous l'avez à votre disposition. Certains moteurs, comme YouTube, acceptent également des sous-titres et

savent les lire (*http://www.google.com/support/youtube/bin/answer.py?hl=fr&answer=100076*). Excellent pour le référencement ! Autant d'informations textuelles qui vont permettre au moteur de bien comprendre « de quoi parle » la vidéo sous la forme d'une « fiche descriptive précise » qui lui est propre.

Critère numéro 7 : indexabilité. Un plan du site (au format HTML), sorte d'annuaire spécialisé présentant les vidéos présentes sur votre site peut également être une bonne idée pour donner aux *spiders* des moteurs un point d'entrée unique pour parcourir vos fichiers. Là encore, le texte des liens va être crucial (voir ci-dessus) donc soignez bien la « réputation » de vos vidéos.

N'hésitez pas non plus à intégrer vos vidéos dans vos flux RSS, c'est encore une autre voie pour faire connaître vos fichiers aux moteurs de recherche qui les prennent en compte.

Enfin, n'hésitez pas à soumettre ou *uploader* vos vidéos sur un maximum d'outils de recherche « YouTube-Like ». C'est encore le meilleur moyen de leur faire connaître vos « œuvres »... Certains moteurs de recherche vidéos vous proposent également de soumettre vos fichiers sous la forme d'un document au format RSS ou MRSS (*http://en.wikipedia.org/wiki/Media_RSS*), sorte de « Sitemap vidéo ». Utilisez également cette voie ! La soumission est importante car elle permet d'ajouter un descriptif et un titre, champs très importants pour les moteurs actuels.

Bien entendu, un Sitemap (voir chapitre 8) spécialisé pour les vidéos devra être présent sur votre site. Exemple :

```
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9"
        xmlns:video="http://www.google.com/schemas/sitemap-video/1.1">
<url>
  <loc>http://www.example.com/videos/some_video_landing_page.html</loc>
    <video:video>
      <video:content_loc>http://www.site.com/video123.flv<video:content_loc>
      <video:player_loc allow_embed="yes">http://www.site.com/videoplayer.swf?video
      ➥=123</video:player_loc>
      <video:thumbnail_loc>http://www.example.com/miniatures/123.jpg<video:thumbnail_loc>
      <video:title>Barbecue en été<video:title>
      <video:description>Pour des grillades réussies<video:description>
      <video:rating>4.2</Video:rating>
      <video:view_count>12345</video:view_count>
      <video:publication_date>2007-11-05T19:20:30+08:00.</video:publication_date>
      <video:expiration_date>2009-11-05T19:20:30+08:00.</video:expiration_date>
      <video:tag>steak</video:tag>
      <video:tag>viande</video:tag>
      <video:tag>été</video:tag>
      <video:category>barbecue</video:category>
      <video:family_friendly>yes</video:family_friendly>
      <video:expiration_date>2009-11-05T19:20:30+08:00<video:expiration_date>
      <video:duration>600</video:duration>
    </video:video>
</url>
```

Vous trouverez plus d'informations à ce sujet sur le site de Google dédié aux Sitemaps : *http://www.google.com/support/webmasters/bin/answer.py?answer=80472*.

Certains outils comme YouTube permettent également à leurs visiteurs de laisser des commentaires ou de noter les vidéos. Autant de façons de leur donner une meilleure visibilité...

Pour conclure, n'hésitez pas à toujours rechercher, tester, « fouiner » les résultats de recherche des moteurs en essayant de comprendre comment ils fonctionnent. Mettez en place une alerte « Google News » sur des mots-clés comme « référencement vidéo », « video SEO », « video optimization », etc. Le domaine de la recherche de vidéos – et donc de leur optimisation – n'en est encore qu'à ses balbutiements et nous apprendrons encore beaucoup de choses dans les années qui viennent... Bref, dans ce domaine peut-être encore plus qu'ailleurs, une veille est absolument indispensable...

Quelques liens à consulter sur le sujet pour en savoir plus

- *Make Your Videos Rank on the Search Engines* de Terri Wells
  http://www.seochat.com/c/a/Search-Engine-Optimization-Help/Make-Your-Videos-Rank-on-the-Search-Engines/
- *5 conseils pour optimiser le référencement de vidéo*
  http://www.pckult.net/articles/conception-web/993-5-conseils-pour-optimiser-le-rrencement-de-vid
- *Video Search Optimization*
  http://www.seroundtable.com/archives/006858.html
- *Search Illustrated: Video Optimization* de Elliance
  http://searchengineland.com/071120-133911.php
- *7 Ways to Optimize Your YouTube Tags* de Jonathan Mendez
  http://www.optimizeandprophesize.com/jonathan_mendezs_blog/2007/02/optimize_your_y.html
- *Balancing Video Quality and Search Optimization* de Grant Crowell
  http://searchenginewatch.com/showPage.html?page=3626098
- *Optimizing Video for Search Engines* de Amy Edelstein
  http://searchenginewatch.com/3624257

Le référencement de fichiers PDF et Word

Les moteurs de recherche actuels référencent, et classent même parfois très bien, les fichiers aux formats PDF (`.pdf`) et Word (`.doc`). Aussi, il nous a semblé important, dans cet ouvrage, de vous donner quelques informations sur la meilleure façon d'optimiser ces fichiers afin de les voir bien positionnés dans les pages de résultats des moteurs. Voici quelques trucs et astuces qui devraient vous y aider...

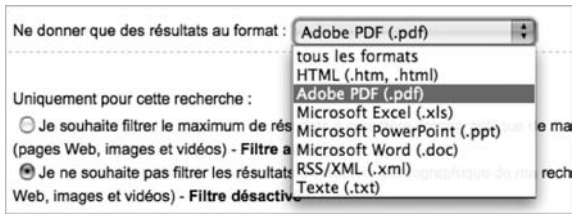
Prise en compte de ces fichiers par les moteurs

Tout d'abord, il est important de bien comprendre, comme nous le disions précédemment, que les moteurs de recherche indexent sans problème les fichiers PDF et Word. Pour les visualiser, vous pouvez :

- Sur Google et Bing, utiliser la syntaxe « filetype: ». Exemple : « abondance filetype:pdf » ou « abondance filetype:doc » pour indiquer le filtre adéquat au moteur.
- Sur Yahoo!, utiliser la recherche avancée du moteur (*http://fr.search.yahoo.com/web/advanced?ei=UTF-8&p=*) et opter pour le choix « Format de fichiers>Ne donner que des résultats au format : » qui propose notamment ces deux possibilités, entre autres, comme indiqué sur la figure 6-9.

Figure 6-9

Filtre de recherche sur le format de fichier. Recherche avancée de Yahoo!



Chacun de ces deux moteurs indique, dans ses résultats, le format des fichiers trouvés devant leur titre par la mention [PDF] sur les deux outils, [DOC] sur Google et [MICROSOFT WORD] sur Yahoo!. Bing, en revanche, indique « Fichier PDF » ou « Fichier DOC » à droite de l'URL si un tel fichier est proposé.

Figure 6-10

Fichier PDF dans les résultats de Google

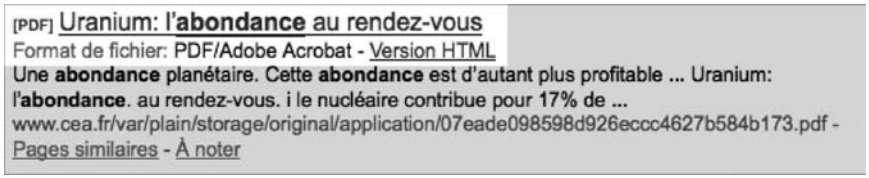


Figure 6-11

Fichier Word dans les résultats de Bing

Côté indexation, pas de problème donc... Un simple lien (*spider friendly*, voir chapitres 5 et 7) vers ces fichiers dans vos pages sera suivi par les robots des moteurs et générera automatiquement leur indexation. Exemple :

```
<a href="http://www.votresite.com/fichiers/votrefichier.pdf">Lien qui permettra aux
➥robots d'indexer votre fichier</a>
```

Sur ce point, pas de problème. Reste à envisager l'optimisation de ces fichiers : quelles informations les moteurs de recherche lisent-ils en leur sein ?

Zones reconnues par les moteurs de recherche

Voici les différentes zones que l'on peut « remplir » dans un fichier Word ou PDF et la façon dont Google et Bing les lisent ou non (notamment les « Propriétés » ou « métadonnées » que l'on peut fournir sur ces deux formats pour mieux décrire les documents).

Tableau 6-1 Champs pris en compte par Google et Bing pour des fichiers PDF

	Google	Bing
Contenu textuel	OUI	OUI
Métadonnée Titre (`Title`)	OUI	NON
Métadonnée Sujet (`Subject`)	NON	NON
Métadonnée Auteur (`Author`)	NON	NON
Métadonnée Mots-clés (`Keywords`)	NON	NON
URL	OUI	OUI

Tableau 6-2 Champs pris en compte par Google et Bing pour des fichiers Word

	Google	Bing
Contenu textuel	OUI	OUI
Métadonnée Titre (`Title`)	OUI	NON*
Métadonnée Sujet (`Subject`)	NON	NON
Métadonnée Manager	NON	NON
Métadonnée Auteur (`Author`)	NON	NON
Métadonnée Compagnie	NON	NON
Métadonnée Category	NON	NON
Métadonnée Mots-clés (`Keywords`)	NON	NON
Métadonnée Commentaires (`Comments`)	NON	NON
URL	OUI	OUI

** Nous avons trouvé quelques cas isolés où Bing lisait la balise `<title>` du document mais la plupart du temps, ce n'était pas le cas... Notons également que ces données peuvent changer suite à l'accord entre Microsoft et Yahoo! (juillet 2009).*

La situation est donc, ici, assez simple en termes de lecture des contenus et des métadonnées par ces deux moteurs majeurs :

- Les deux moteurs lisent les contenus textuels des deux formes de fichiers.

- Ils détectent également les mots-clés dans les URL. Il sera donc important de soigner les intitulés des noms de fichiers en y incluant des mots-clés pertinents par rapport au contenu de l'article ou du document.
- Pour les métadonnées, Google ne lit que la balise `<title>` des fichiers PDF et Word. Il s'en sert d'ailleurs parfois pour l'afficher comme titre dans ses résultats. Ce moteur ne lit pas d'autres métadonnées.
- Pour ces métadonnées, Bing ne lit pas le `<title>` mais il lit les mots-clés (`keywords`) ajoutés au document PDF (mais pas Word). Il ne lit pas d'autres métadonnées.

Côté métadonnées, le travail sera finalement assez vite effectué sur vos fichiers : seule la balise `<title>` devra être remplie, voire également la balise `<keywords>` pour Yahoo! (bien qu'il y ait de fortes chances pour que ce moteur octroie à cette zone un poids bien faible...).

Contenu des snippets

Comment Google affiche-t-il ses résultats lorsqu'il s'agit de fichiers PDF ou Word, comme indiqué sur la figure 6-12 ?

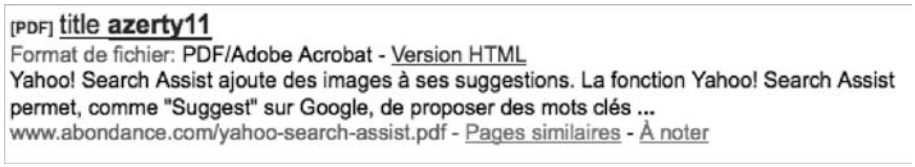


Figure 6-12

Affichage d'un fichier PDF dans les résultats de Google

Le titre du fichier (en bleu et souligné, à droite de la mention [PDF]) est constitué :

- Par le contenu de la métadonnée `title` s'il existe. Il semblerait que Google ne prenne pas en compte la balise `<title>` si son contenu se termine par une terminaison de fichier (`.doc`, `.eps`...). Il faut donc que cette zone contienne de « vrais mots » et non pas un nom de fichier.
- Par un titre trouvé dans le contenu textuel du document si la balise `<title>` n'est pas renseignée ou si elle propose un nom de fichier.

La mention « Format de fichier : PDF/Adobe Acrobat » est indiquée ensuite, suivie d'un lien vers une version HTML du document, permettant de visualiser rapidement le contenu du fichier.

Le texte descriptif reprend le début du contenu textuel du document (ici, « Yahoo! Search Assist ajoute des images à ses suggestions. La fonction Yahoo! Search Assist permet, comme « Suggest » sur Google, de proposer des mots-clés ») ou une zone de texte identifiée comme pertinente par Google.

À ce niveau, nous vous conseillons de créer des documents PDF dont la mise en page est très simple, surtout pour la première page : titre, chapô, texte, avec un titre et un chapô très descriptifs. La mise en page peut être plus sophistiquée par la suite, mais, si vous voulez maîtriser la façon dont Google « comprend » vos documents, un système de multicolonnes n'est par exemple pas approprié, pour le début du document en tout cas... Car dans ce cas, c'est le contenu de la première colonne qui risque d'être pris en compte par Google, ce qui ne donnera pas obligatoirement le résultat escompté...

Notons que les fichiers Word sont affichés, dans les grandes lignes, de la même façon sur Google que les PDF.

Quelques conseils d'optimisation

Pour bien optimiser vos fichiers PDF et Word, nous vous conseillons donc de suivre les conseils suivants :

- **Métadonnées** : remplissez la balise `<title>` qui est prise en compte par Google et parfois par Bing. Si vous avez un peu de temps, renseignez la balise `<keywords>` (n'y passez pas non plus beaucoup de temps, cela n'en vaut pas vraiment la peine)...
- **Contenu** : adoptez, pour la première page notamment, une mise en page simple et efficace (un titre en gras, en gros caractères, et un chapô de deux ou trois phrases, également en gras, le tout étant très descriptif et contenant les mots-clés importants pour comprendre le document). Et oui, on est ici très proche de l'optimisation d'une page HTML « classique »... Notez bien que les *footer* et *header* (haut de page et bas de page) sont compris comme du texte par les moteurs. Évitez de les afficher sur la première page car les moteurs risquent de lire le contenu du *header* en premier, ce qui peut ne pas être pertinent (notez bien que cela peut également l'être...).
- **Indexation** : bien sûr, proposez, depuis vos pages web, des liens vers vos documents PDF et Word, avec des intitulés de liens explicites, pour leur donner « bonne réputation » (évitez donc les phrases du type « Pour télécharger notre livre blanc sur le référencement, cliquez ici » et préférez des formulations du type « Téléchargez notre livre blanc sur le référencement »). N'hésitez pas, également, à proposer une page spéciale « Téléchargement de documents » listant, sous la forme d'un mini-annuaire, tous les fichiers disponibles sur votre site. Une bonne piste de départ pour les *spiders* des moteurs...
- **Nom du fichier** : utilisez des mots explicites pour nommer votre fichier (*assemblee-generale-2008.pdf*, *charte-experts-comptables.doc*, etc.) car ces mots sont lus et analysés par les moteurs.
- **Taille du fichier** : attention aux formats de fichiers, dépassant notamment le méga-octet. Des fichiers trop volumineux ne seront peut-être pas indexés en totalité. Préférez une suite de « petits fichiers » dans ce cas...
- **Accessibilité** : Word, notamment, propose de nombreuses fonctions visant à améliorer l'accessibilité des fichiers (ajout de l'attribut `alt` aux images, etc.). N'hésitez pas à vous en servir, les critères d'accessibilité sont toujours très proches de ceux du référencement...

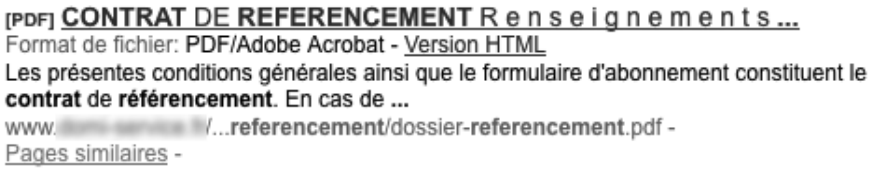


Figure 6-13

Un nom de fichier et une URL contenant des mots-clés précis sont deux points importants. La requête « filetype: » permet par ailleurs de trouver de nombreux documents très intéressants sur le Web (ici, des contrats types de référencement).

Voici donc pour ces quelques conseils qui, nous l'espérons, vous aideront à mieux optimiser vos fichiers pour les deux moteurs de recherche majeurs.

Pour en savoir plus

Les lignes que vous venez de lire sont basées sur nos propres tests. Notez bien que cette situation peut évoluer avec le temps. Par ailleurs, pour les conseils d'ordre général, nous avons parfois également puisé quelques informations dans les articles suivants que nous vous conseillons de lire avec la plus grande attention :

- *Optimiser le référencement des fichiers PDF* de Sébastien Billard
  http://s.billard.free.fr/referencement/?2007/09/14/422-optimiser-le-referencement-des-fichiers-pdf
- Traduction de l'article *Eleven Tips for Optimizing PDFs for Search Engines* de Galen DeYoung , dont l'original est disponible à l'adresse suivante :
  http://searchengineland.com/eleven-tips-for-optimizing-pdfs-for-search-engines-12156.php
- *Accessibilité et référencement des fichiers PDF*
  http://www.cmic.ch/2007/09/14/accessibilite-et-referencement-des-fichiers-pdf/
- *SEO Your PDF's* de Kevin Kantola
  http://www.searchengineguide.com/senews/002430.html
- *Optimizing PDFs for SEO* de Matt McGee
  http://www.smallbusinesssem.com/optimizing-pdfs-for-seo/288/

Mais il en existe bien d'autres... Bonne optimisation !

Référencement sur l'actualité et sur Google News

Google News (*http://news.google.fr/*) est, bien évidemment, l'un des principaux sites d'actualité sur le Web francophone. Pour un site qui traite d'informations et d'actualité « chaude », cet outil représentera parfois plus de la moitié de son trafic « moteurs de recherche ». Loin d'être négligeable... Voici donc quelques pistes de réflexion pour vous aider à mieux bâtir vos pages d'actualité afin de leur donner une meilleure visibilité sur Google Actualités...

Comment se faire référencer sur Google News ?

Dans un premier temps, avant d'apparaître dans les pages de résultats de l'outil, il faut que Google accepte le référencement de votre site d'informations. Si vous êtes un site à forte notoriété, cela ne devrait pas poser trop de problèmes, il y a même de fortes chances pour que cela se fasse sans que vous n'ayez rien à faire...

Si vous n'avez pas la chance d'être une source d'informations incontournable, il vous faut alors demander à être référencé sur le site. Vous devez passer par le formulaire idoine. Allez sur « Aide pour les éditeurs » en bas de la page d'accueil du site, puis « Envoi de votre contenu » et enfin « Google Actualités » pour cliquer sur « Envoyez-nous son URL » (*http://www.google.com/support/news_pub/bin/request.py?contact_type=suggest_content&hl=fr*).

La deuxième étape consiste à donner les indications (URL, nom, adresse e-mail, description) de la source d'informations à indexer.

Figure 6-14

Formulaire à remplir pour soumettre son site à Google News

N'oubliez pas de vérifier si le site n'est pas déjà référencé. En effet, il serait mal accepté par le moteur de soumettre un site qui se trouve déjà dans la base de Google... Vous feriez perdre leur temps aux personnes chargées de vérifier ces données...

Lors de la vérification, indiquez l'URL sous la forme la plus simple possible (abondance.com, libe.fr, etc.) et surtout sous le domaine où apparaissent les articles mis en ligne (n'indiquez pas *www.votresite.com* si vos articles sont accessibles sous *actu.votresite.com...*).

Sachez simplement que pour être intégré dans Google News, votre site devra répondre à plusieurs critères :

Critère numéro 1 : avoir une zone Actualités ou tout du moins une zone d'information mise à jour régulièrement. Il n'est pas nécessaire d'avoir une zone très fortement axée sur l'actualité « fraîche », mise à jour quotidiennement, voire plus souvent, pour être accepté. Un blog mis à jour chaque semaine peut faire l'affaire. En revanche, d'une façon ou d'une autre, un être humain, chez Google, va venir évaluer votre site. Évitez, par exemple, dans les jours qui suivent votre demande, l'autopromotion ou les fautes d'orthographe et de frappe sur vos pages... Bref, faites le forcing pendant cette période (mais pas uniquement) pour proposer de l'actualité qui « tienne la route » car vous pouvez être sûr que, bientôt, vous allez passer un « examen » à distance.

Critère numéro 2 : l'URL. Chaque page d'actualité, chaque information, chaque dépêche doit être accessible par l'intermédiaire d'une URL spécifique. Exemple :

http://actu.abondance.com/2009/07/yebol-un-nouveau-moteur-interessant.html

L'actualité comme elle était présentée en 2001 sur le site Abondance, sous la forme d'une seule page par semaine (suite de brèves affichées les unes en dessous des autres) regroupant toutes les dépêches, n'aurait pas été recevable :

http://actu.abondance.com/actu0141.html

Par ailleurs, si un article ne reste pas disponible en ligne durant les 30 jours pendant lesquels il restera accessible sur Google News, cela peut aussi poser quelques problèmes (cas des articles qui passent en zone d'archives payantes quelques jours après leur publication).

Critère numéro 3 : au moins trois chiffres dans l'URL. Critère assez étrange mais officiellement demandé par Google, les URL de vos pages d'actualité doivent contenir au moins trois chiffres. Exemples :

- *http://www.collectifvan.org/article.php?r=4&&id=4632*
- *http://www.agoravox.fr/article.php3?id_article=14384*
- *http://www.sports.fr/fr/cmc/scanner/football/200641/psg-halilhodzic-devra-payer-_109909.html?popup*

Vous pouvez vérifier que c'est effectivement le cas de tous les articles référencés sur l'outil... Ces chiffres peuvent désigner la date, le numéro de semaine ou un numéro d'articles, etc. Peu importe. L'essentiel est qu'il y ait ces trois numéros dans l'URL...

Ceci est rappelé dans le message que renvoie Google lorsque le site proposé ne propose pas ce type de particularité :

« Merci pour la demande d'ajout sur Google Actualités. Après investigation, nous avons constaté que notre système ne pouvait pas explorer certains de vos articles en raison du format de leurs URL. Pour que vos articles puissent être analysés par le système Google Actualités, leurs URL doivent contenir un nombre composé de trois chiffres au minimum. »

Vous savez donc ce qu'il vous reste à faire si cela n'est pas le cas... Notez cependant que, fin 2009, des rumeurs semblaient faire écho du fait que cette restriction pourrait être supprimée à moyen terme par Google. Mais il ne s'agissait encore, à ce moment-là, que de rumeurs...

Certains webmasters ont également reçu la réponse suivante suite à leur demande d'inclusion :

« Merci pour votre courrier électronique. Nous avons examiné le site *http://www.votre-site.com*, mais nous ne sommes pas en mesure de l'ajouter sur Google Actualités pour l'instant. Nous n'acceptons pas les journaux web (blogs) sur les actualités ni les sites d'information rédigés et actualisés par des particuliers. De même, nous ne pouvons pas inclure les sites pratiquant la publication ouverte sans processus formel de rédaction. Nous vous remercions d'avoir pris le temps de nous contacter et conserverons votre site afin de l'ajouter si nous modifions nos critères d'intégration. »

Il semble plutôt s'agir ici d'une façon polie de la part de Google d'indiquer que votre site web n'est pas assez « bon » dans son contenu pour être accepté. En effet, on trouve dans Google News de très nombreux blogs rédigés par des particuliers, sans réel processus « formel » de rédaction... Bref, si vous recevez ce type de message, il ne vous reste plus qu'à retravailler la qualité de vos articles et de retenter votre chance d'ici quelques semaines... Et oui, ce n'est jamais très agréable à lire... Cela dit, si vous « pompez » vos articles un peu partout sur le Web, ne vous étonnez pas de recevoir ce type de réponse...

Sachez enfin que le délai d'inclusion dans la base de Google, une fois la demande effectuée, est de un à deux mois... Ne vous impatientez donc pas si, une semaine après votre requête, vous n'êtes toujours pas indexé... Sachez également que Google News « crawle » les pages web de votre site. Lui proposer l'adresse d'un flux XML (RSS, Atom) ne servira à rien, ce n'est pas par ce biais que le moteur indexera votre contenu...

Comment assurer une indexation régulière des articles ?

Ce n'est pas parce que votre site est référencé en tant que source d'informations sur Google News que tous vos articles et/ou toutes vos dépêches vont être obligatoirement indexées et pris en compte. *A priori* (sans que Google ait jamais communiqué de façon officielle à ce sujet), cela semble dépendre de deux critères principaux :

- La taille de la zone éditoriale (l'article en lui-même) proposée : essayez de toujours dépasser les 150 à 200 mots et les 1 200 à 2 000 caractères (espaces compris) pour le corps de l'article, cela devrait fortement augmenter vos chances d'indexation et de

prise en compte. Google n'acceptera pas les articles trop courts (moins de 100 mots et/ou moins de 1 000 caractères)... Attention également à la taille du corps de l'article par rapport à la taille totale de la page (charte graphique, zones de navigation, etc.). L'idéal est que ce corps éditorial soit plus important (en termes de taille) que la moitié du contenu total du document... Bref, que l'aspect éditorial soit supérieur à l'aspect « look et navigation »...

- Le nombre d'articles indexés sur un sujet dans l'index du moteur. Si Google estime qu'il a suffisamment d'articles sur un sujet donné qui fait couler beaucoup d'encre (par exemple, l'accord « historique » entre Yahoo! et Microsoft en juillet 2009), il se peut qu'il « choisisse » les articles qu'il va indexer pour ne pas crouler sous de trop nombreuses pages. Si le sujet sur lequel vous écrivez un article est très « populaire », tentez de faire un article le plus long possible en termes de mots et de le publier le plus rapidement possible. Cela ne signifie pas non plus qu'il faille faire du « remplissage à la va-vite », n'oubliez pas que vos articles ont pour but d'être lus par des internautes... Une longueur « suffisante » fera peut-être en sorte que votre page sera « retenue » par l'outil mais cela n'est malheureusement pas une garantie...

Pas trop vite quand même...

La rapidité de réaction d'une source d'informations sur un sujet donné est un critère essentiel de Google News. À tel point qu'on voit des sites sportifs publier un article sur un match de football avant la fin de ce dernier, pour être sûr d'être le premier à en parler en ligne... Et tant pis si un but est marqué dans les arrêts de jeu !

N'hésitez pas non plus à agrémenter votre article d'une illustration (image, photo, graphique, etc.). Il se pourrait bien que cela aide à une meilleure indexation de vos pages (voir plus loin dans ce chapitre).

Comment apparaître sur la page d'accueil de Google Actualités ?

La page d'accueil du moteur (*http://news.google.fr/*) propose bon nombre d'articles sur des sujets considérés comme « populaires » (certainement le plus souvent cités dans les heures qui viennent de s'écouler), le tout dans plusieurs catégories : « À la une », « International », « France », etc.

Il semblerait que Google ait mis en place un système de « TrustRank » (différent de celui utilisé pour la recherche web) ou de « NewsRank », ou indice de confiance, fourni en partie par des êtres humains, permettant de donner des priorités à certaines sources d'informations considérées comme plus crédibles. Ainsi, on s'aperçoit rapidement que ce sont souvent les mêmes sources qui apparaissent en « Une » : Le Nouvel Observateur, Libération, Boursier.com, RFI, Le Monde, Les Échos, L'Express, etc. Bref, uniquement des sources « dignes de confiance », ayant « pignon sur rue » et contre lesquelles il sera difficile de lutter au niveau de la visibilité, certainement parce qu'elles ont reçu une sorte de « label de qualité » de la part de certains experts chez Google...

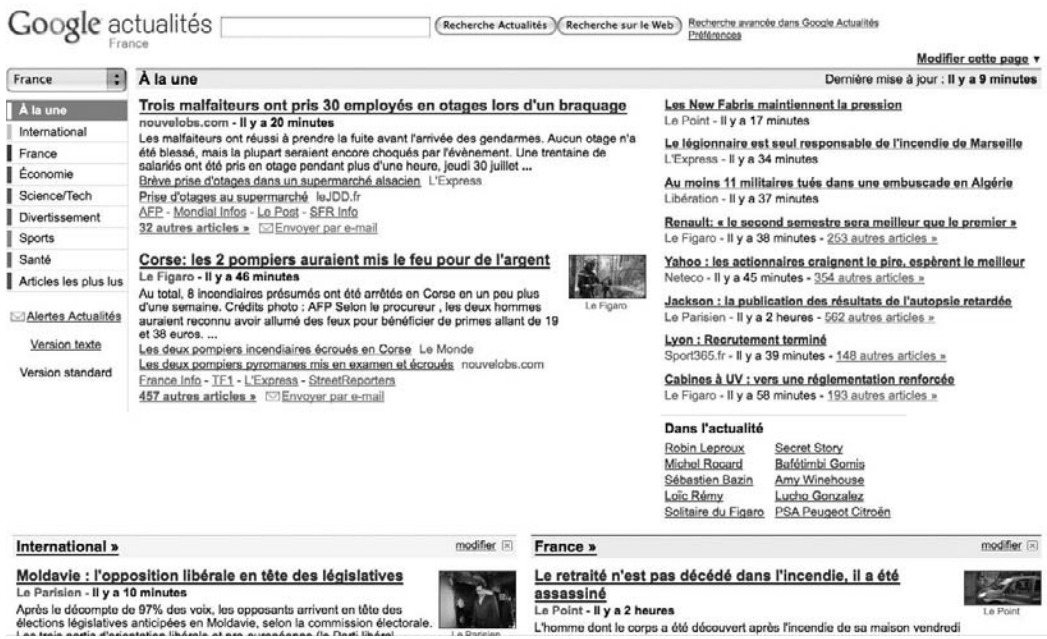


Figure 6-15

La page d'accueil de Google News propose de nombreux articles classifiés en différentes rubriques.

Mais il vous sera peut-être possible d'apparaître en page d'accueil sur des sujets plus pointus, moins généralistes, sur lesquels les « grands » n'ont pas obligatoirement écrit encore d'articles disponibles en ligne. On s'aperçoit rapidement que sur les « grands titres », il est quasiment impossible de « se battre » contre la presse « nationale ». Mais sur des thématiques plus ténues, vous avez vos chances... À vous, peut-être, d'être rapide et de proposer en ligne un article avant les « grands » pour voir celui-ci repris en une pendant quelques heures. Mais il ne semble pas y avoir de solutions miracles à ce sujet...

Voici, selon un brevet déposé par Google (dont nous avons déjà parlé au chapitre 5), quelques exemples de critères qui seraient pris en compte dans le TrustRank de Google News :

1. le nombre d'articles produit par la source ;
2. la longueur moyenne des articles ;
3. la « couverture » de la source ;
4. la réactivité de la source (*breaking score*) ;
5. un indice d'utilisation (en nombre de clics sur cette source) ;
6. une opinion humaine sur la source ;
7. une statistique extérieure d'audience telle que Media Metrix ou Nielsen Netratings ;

8. la taille de l'équipe ;
9. le nombre de bureaux ou agences différents de la source ;
10. le nombre d'entités nommées originales citées par la source (personnes, organisations, lieux) ;
11. l'étendue (*breadth*) et le nombre de sujets couverts par la source ;
12. la diversité internationale ;
13. le style de rédaction, en termes d'orthographe, de grammaire, etc.

Sources : Brevet *Systems and Methods for Improving the Ranking of News Articles* déposé par Google, cité par l'excellent blog « Technologies du Langage » (*http://aixtal.blogspot.com/2005/05/google-trustrank-beaucoup-de-bruit.html*). Certains de ces critères sont assez complexes à vérifier, il faut bien l'admettre, mais cela donne encore quelques explications et voies de réflexion... Tous ces critères sont certainement pris en compte dans l'algorithme global de pertinence...

Enfin, il est important de connaître un autre point : la quasi totalité de la gestion quotidienne des sites Google Actualités dans le monde est automatisée, sans aucune intervention humaine. Ce ne sont donc pas des éditeurs, contrairement à de nombreux autres sites similaires, qui effectuent des choix d'articles, d'images, etc. Ce sont des algorithmes. Seule une petite équipe de « googlers » (personnes travaillant chez Google) œuvrent sur l'outil Google News dans le monde... Impossible pour eux d'effectuer un traitement humain quotidien et un quelconque tri des informations qui serait effectué « à la main »... Seuls le choix et la « notation » des sources d'informations sont manuels au départ, tout le reste est automatisé.

Comment faire apparaître une image ?

Vous vous en êtes certainement rendu compte si vous utilisez souvent l'outil Google Actualités, certains articles sont agrémentés d'une image sur la gauche dans les pages de résultats, comme le montre la figure 6-16.

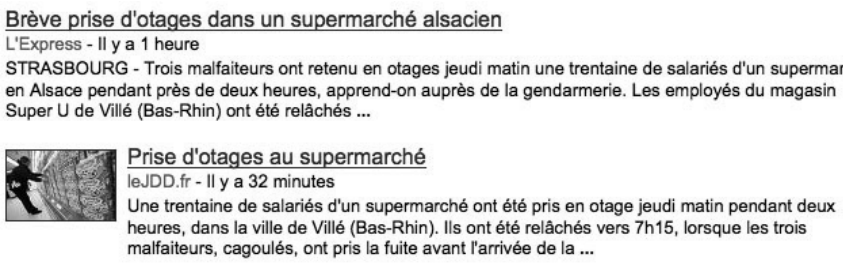


Figure 6-16

Une image est parfois affichée en face d'un article dans les pages de résultats.

Ce type d'image donne un « focus » indéniable à l'article qu'il souligne. Il y a fort à parier que les taux de clics sur les articles rehaussés par ce type de vignette doivent être bien plus forts que lorsqu'il n'y en a pas... Comment faire, alors, pour faire apparaître ces images ? Il s'agit là d'un mystère pas si simple à percer... Voici cependant quelques indications :

- La présence d'une image ne semble pas avoir de relation avec une quelconque notion de « TrustRank » du site. *A priori*, il ne semble pas que la présence d'une image ait non plus un rapport avec le mot-clé saisi sur Google News. Lorsqu'un article donné est complémenté par une vignette dans les pages de résultats, il semble l'être quel que soit le mot-clé saisi pour le trouver. Inutile donc d'insérer dans le nom de l'image ou dans l'option `alt` de la balise `<img>` (voir ci-après) certains mots-clés pouvant être saisis selon vous par les internautes pour trouver l'article en question. Préférez une bonne adéquation entre ces indications et le thème général de l'article, comme nous le verrons par la suite...
- Essayez plutôt d'afficher des images de « grande taille » (supérieures à 200 × 200 pixels), afin que Google puisse les réduire sous forme de vignettes sans trop y perdre en qualité.
- Préférez les formats GIF et JPEG.
- Faites en sorte que le nom de l'image (`xxxx.gif` ou `yyyy.jpg`) contienne un ou plusieurs mots explicites et correspondant au contenu de l'article (titre, corps du texte). Exemple : `anniversaire-ibm.gif` ou `jacques-chirac.jpg`.
- Renseignez l'option `alt` de l'image avec le titre de l'article ou sa légende. Par exemple, si le titre de l'article – et de la page – est « Corée du Nord : vers des sanctions à l'ONU », indiquez dans la balise `<img>` ces informations :

```
<img src="Images/coree-du-nord.jpg" alt="Corée du Nord : vers des sanctions à l'ONU"
➥height="265" width="230">
```

  Si la légende ou un texte proche de l'image contiennent la phrase « L'essai nucléaire nord-coréen a mis en émoi les grandes puissances mondiales », vous pouvez également indiquer ceci dans votre code HTML :

```
<img src="Images/coree-du-nord.jpg" alt="L'essai nucléaire nord-coréen a mis en émoi
➥les grandes puissances mondiales" height="265" width="230">
```

- Indiquez la largeur (`width`) et la hauteur (`height`) de l'image dans sa description, comme dans le code ci-dessus. Cela aidera éventuellement Google à la réduire sous la forme d'une vignette... À ce sujet, Google News semble bien apprécier les images qu'il peut réduire sous un format proche de 60 pixels (hauteur) sur 80 (largeur), ou le contraire pour une image en hauteur (80 × 60). Tentez de proposer des fichiers dont la taille en pixels correspond à un facteur multiplicateur de ces chiffres...
- Affichez votre image au tout début de votre texte, donc du corps de l'article, juste après le titre. L'alignement peut-être à droite, à gauche ou centré, cela ne semble pas poser de problème. En revanche, la présence de l'image au tout début du texte de l'article et après le titre semble essentielle pour qu'elle soit reprise par Google.
- Enfin, il semblerait que l'image ne doive pas être cliquable pour être retenue...

Tous ces conseils devraient vous aider à mieux optimiser la présence d'images extraites de vos articles sous la forme de vignettes dans les pages de résultats de Google Actualités. Il ne s'agit pas, là non plus, de recettes miracles, mais plutôt de « petites astuces » qui devraient améliorer votre situation à ce niveau. Sachant également que tout cela évolue à vitesse grand V et qu'une veille est obligatoire dans ce domaine...

Comment mieux positionner un article dans les résultats ?

Peu de surprises *a priori* pour une meilleure optimisation d'une page pour la voir apparaître plutôt en tête de classement sur la saisie d'un mot-clé. La plupart des critères pris en compte lors de l'élaboration d'une page web restent valables : bonne optimisation de la balise `<title>`, indication des mots-clés importants et descriptifs de l'article en haut de page (dans le titre de l'article – dans le corps de la page – et dans le chapô, voire le premier paragraphe), indication du titre dans une balise `<h1>`, etc. Le fait d'indiquer des mots-clés dans l'URL (*www.votresite.com/actu/elections-presidentielles.html*) peut jouer également...Bref, rien de bien nouveau à ce niveau-là. Vous connaissez ça parfaitement si vous avez lu les paragraphes précédents...

Un point nous a paru important dans nos investigations : il nous a semblé que, contrairement au moteur de recherche web, Google News accordait plus d'importance à la présence des mots dans le texte des liens sortants de la page. Si un mot est inclus dans le texte d'un lien, dans un document, cela semble donner un poids plus fort à ce dernier...

Comment faire pour ne pas être indexé par Google News ?

Il s'agit ici de la mauvaise nouvelle : il n'existe aucun moyen spécifique pour voir votre site non indexé par Google News tout en le restant sur Google Web. En effet, les systèmes de « barrage » proposés par Google pour votre site web sont communs à Google Web ET Google News. Si vous installez un fichier `robots.txt` ou une balise meta `robots` pour barrer l'accès aux robots nommés Googlebot, cela sera valable pour le moteur Web de Google mais aussi pour son moteur d'actualités.

Impossible, en d'autres termes, de barrer l'accès à Google News en laissant la porte ouverte à Google Web, de voir son site indexé sur le moteur web de Google et pas sur le moteur d'actualités (*http://googlepolicyeurope blogspot.com/2009/07/working-with-news-publishers.html*). Il est étonnant que le moteur leader ne propose pas ce type de possibilité et c'est peut-être là la source de certains de ses ennuis juridiques à ce niveau... Même s'il y a une certaine logique à ce comportement. En effet, on voit mal pourquoi un site accepterait l'indexation de ses dépêches d'actualités sur le moteur Web et pas sur le moteur de news... Ceci dit, le choix pourrait quand même être laissé à l'éditeur du site, ce qui n'est pas le cas à l'heure actuelle...

Si vous désirez donc voir votre site éliminé des résultats sur les deux outils de recherche, vous pouvez mettre en place un fichier `robots.txt` contenant les lignes suivantes :

```
User-agent: Googlebot
Disallow: /
```

ou utiliser les deux balises meta `robots` suivantes au choix :

```
<meta name="robots" content="noindex, nofollow">
<meta name="googlebot" content="noindex, nofollow">
```

L'emploi de la fonction `unavailable_after` est également possible, voir chapitre 9 à ce sujet.

Quelques liens sur le référencement dans Google News

Voici quelques articles au sujet de Google News et du référencement de sites web dans son index :

– Un article du site News Scientist sur le TrustRank :
http://www.newscientist.com/article.ns?id=mg18624975.900

– Des travaux de l'université de Stanford sur ce même sujet (*Combating Web Spam with TrustRank* ou la notion de « confiance » mise en équations) :
http://ilpubs.stanford.edu:8090/638/

– Un webmaster qui explique comment il a spammé Google News. À lire si vous avez envie d'être blacklisté ! :
http://www.zuneo.net/2005/10/jai-spamm-google-news.html

Le référencement local (Google Maps)

Le service Google Maps (*http://maps.google.fr/*) s'est largement déployé en France et devient de plus en plus intéressant dans le cadre de la recherche universelle. En effet, sur certaines requêtes, les résultats Google Maps se positionnent « dans le haut du panier » sur les résultats Google classiques (moteur web).

L'exemple suivant sur l'expression « pizzeria paris » dans Google sera plus parlant pour évaluer l'intérêt de Google Maps.

Figure 6-17

Requête « pizzeria paris » sur Google : le moteur fait la part belle aux résultats issus de Google Maps.

De plus, depuis avril 2009, il n'est plus nécessaire de saisir le nom d'une localité sur certaines requêtes (*http://actu.abondance.com/2009/04/les-recherches-de-plus-en-plus-locales.html*) : Google va géolocaliser votre ordinateur et ajouter automatiquement des résultats « locaux » au sein des liens proposés.

Figure 6-18

La simple saisie du mot-clé « pizzeria » implique l'affichage de résultats locaux proche de l'emplacement de votre ordinateur (ici, un internaute situé en région parisienne), géolocalisé par Google.

Difficile donc, dans les années qui viennent, de passer outre un bon référencement dans Google Maps pour les sociétés ayant une zone d'activité locale... Autre atout indéniable pour l'internaute, le service Google Maps permet de localiser de façon précise une entreprise sur une carte. Grâce à cette solution gratuite, il est possible de compléter les services apportés par un site web et de faciliter les rencontres directes avec les clients.

Une visibilité accrue en dehors du Web traditionnel, voilà ce que peut apporter Google Maps pour les entreprises, même si celle-ci reste limitée aujourd'hui à des recherches plutôt orientées « tourisme/hôtellerie/restauration » (mais il y a de fortes chances que cela évolue très rapidement)...

Le service Google Maps est également exportable : il est très simple d'ajouter ce type d'information sur son propre site web.

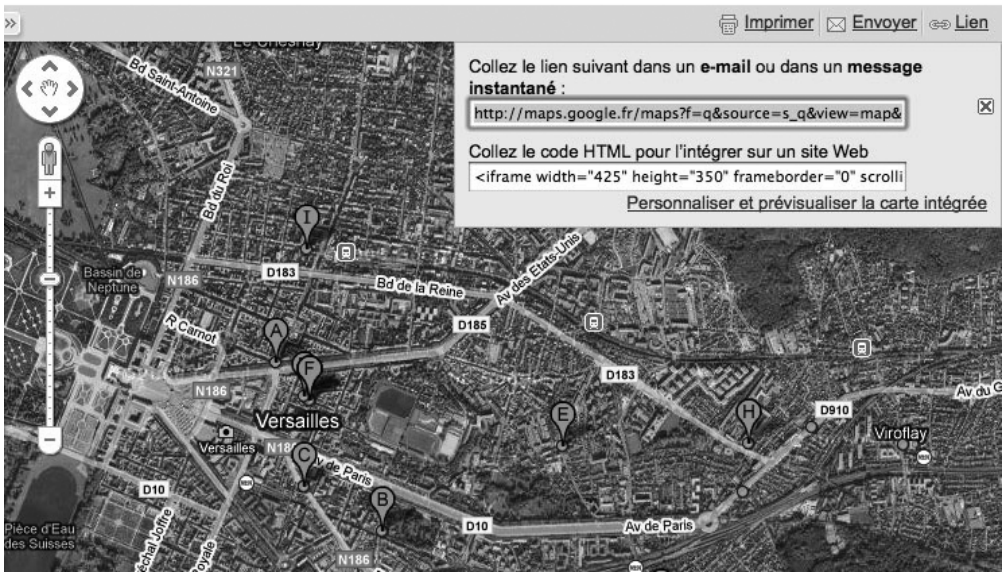


Figure 6-19

Une option permet d'obtenir un lien direct vers le plan proposé.

Google Maps est donc un outil très intéressant en termes de référencement, de visibilité et de services aux internautes, notamment pour des entreprises ayant des offres locales et une clientèle circonscrite à une zone géographique bien délimitée.

Suivre l'évolution de Google Maps

Pour conserver un œil sur l'évolution de cet outil, voici plusieurs ressources incontournables :

- Blog Google Maps France
  http://blogomaps.blogspot.com/
- Informations sur Google Maps et Google Earth (en anglais)
  http://google-latlong.blogspot.com/
- Plusieurs vidéos en français sur Google Maps
  http://www.google.fr/help/maps/tour/

Voyons maintenant comment intégrer et présenter votre entreprise sur Google Maps...

Le service « Local Business Center » est destiné aux entreprises et commerces, et leur permet de se faire référencer facilement (et gratuitement) dans Google Maps.

Voici ce que dit Google à ce sujet (*http://www.google.com/local/add/businessCenter?gl=fr&hl=fr*) :

« Utilisez le Local Business Center pour créer gratuitement votre fiche descriptive. Lorsque des clients potentiels rechercheront des informations locales sur Google Maps, ils trouveront votre entreprise, ainsi que votre adresse, vos horaires d'ouverture et même des photos de votre vitrine ou de vos produits. C'est facile, gratuit, et vous n'avez pas besoin de posséder un site web. »

Le Local Business Center

S'inscrire dans Local Business Center est donc simple et rapide, il suffit de posséder un compte Google et de se rendre à l'adresse suivante : *http://www.google.com/local/add/lookup?hl=fr&gl=FR*.

Figure 6-20

Interface de saisie du Local Business Center. Vous êtes prêt à décrire votre entreprise...

L'inscription repose sur des coordonnées postales et téléphoniques réelles, qui seront vérifiées par Google avant de finaliser l'inscription.

Il suffit de se laisser guider et de renseigner les différents champs proposés par Google. La plupart du temps, Google est capable de situer l'adresse indiquée sur une carte, de manière très précise.

Parmi les champs proposés, certains sont particulièrement importants :

- le numéro de téléphone principal sera utilisé par Google pour valider l'inscription ;
- l'adresse renseignée sera utilisée pour positionner l'entreprise sur une carte ;
- le nom de société sera affiché dans les résultats Google ;
- la catégorie servira à mieux indexer l'entreprise dans Google Maps (on peut choisir jusqu'à cinq catégories) ;
- la description apporte des informations complémentaires (elle joue le même rôle que dans Google Search).

D'autres champs de formulaire, comme les horaires d'ouverture ou les modes de paiement, sont plus anecdotiques en termes de référencement. Il s'agit d'informations qui peuvent intéresser les internautes mais qui ne serviront pas beaucoup au positionnement dans Google Maps (nous en reparlerons bientôt).

Il est également possible d'insérer des images pour illustrer son activité (voir consignes sur l'utilisation des photos : *http://maps.google.fr/help/photospolicy_maps.html*) ainsi que des vidéos issues de YouTube.

Le Local Business Center est très souple d'utilisation, car il permet aussi de créer ses propres informations complémentaires : prix des menus, transport en communs, parking... Vous avez carte blanche pour ajouter des informations susceptibles d'intéresser vos clients.

La validation d'une fiche Local Business Center est assez rigoureuse et elle comporte obligatoirement une phase de validation de l'adresse postale ou du numéro de téléphone par Google. Il peut s'écouler jusqu'à 6 semaines avant de voir sa fiche apparaître dans Google Maps (attention : chaque modification de fiche entraînera un nouveau délai de 6 semaines !).

Il est également possible de soumettre des fichiers de données dans Google Maps, en utilisant l'interface *http://www.google.com/local/add/uploadFeed?hl=fr&gl=FR*. Cette solution est particulièrement adaptée pour les entreprises qui possèdent de nombreux établissements.

Des instructions plus précises sont disponibles à ce sujet sur la page « Instructions pour le transfert groupé de coordonnées d'entreprise », disponible à l'adresse suivante : *http://www.google.com/local/add/helpFeeds?hl=fr&gl=FR*.

Se positionner dans Google Maps

Maintenant que nous avons vu comment s'inscrire dans Google Maps, nous allons voir comment être bien positionné. L'exercice est encore peu connu car Google ne communique

pas à ce sujet. Voici ce qui est dit dans l'aide officielle (*http://maps.google.fr/support/bin/answer.py?answer=7091&topic=13435*) :

« Comme pour tous les autres résultats de recherche Google, Google Maps organise ses listes de commerces et services en fonction de la pertinence des résultats par rapport aux termes recherchés, la distance géographique n'étant pas le seul critère pris en compte. Parfois, notre technologie de recherche estime qu'un commerce ou un fournisseur de service plus éloigné vous conviendra davantage qu'un commerce ou service plus proche. »

Voici quelques règles d'optimisation de vos descriptions dans le Local Business Center qui devraient vous permettre d'optimiser votre visibilité sur cet outil :

- Optimiser le titre avec une écriture du type : « expression clé + nom de société ». Remarques :
  - le nom de ville est optionnel, puisqu'il sera déjà inscrit dans l'adresse ;
  - le nom de société est important, c'est celui qui identifie l'entreprise. Ne pas l'omettre... ;
  - soyez succinct ! Les titres courts seront plus faciles à afficher.
- Optimiser la description : rédigez un petit texte de présentation de moins de 200 caractères, en utilisant l'univers sémantique approprié à votre domaine d'activité. Google semble être capable de comprendre une thématique et de positionner les fiches en fonction de celle-ci.
- Choisissez une catégorie appropriée et n'hésitez pas à compléter votre fiche avec des catégories personnalisées.
- Enrichissez votre fiche avec des images et des vidéos, bref tout ce qui peut plaire à l'internaute est utile pour favoriser le positionnement dans Google Maps.
- Si vous avez un site web, n'oubliez pas d'en renseigner l'adresse ! Cette information sera utilisée par Google lorsqu'il présentera les résultats dans Google Search.

Remarque complémentaire : attention au « mapspam » ! Cette nouvelle tendance consiste à spammer Google Maps en l'inondant de fiches optimisées sur des mots-clés. Début mai 2008, le blog Blumenthal (*http://blumenthals.com/blog/2008/05/02/google-mapspam-local-search-marketing-goes-wild/*) relevait ainsi quelques 4 000 annonces de la même société dans Google Maps ! De telles pratiques deviennent également courantes en France, hélas...

Il faut donc utiliser ce service de façon réfléchie, dans la perspective d'apporter un service complémentaire pour l'internaute. Il y a fort à parier qu'un spammeur dans Google Maps verra son site sanctionné d'une façon ou d'une autre par Google. Autant le savoir...

Enfin, sachez que les avis des utilisateurs sont également proposés dans Google Search... Mais cela ne dépendra pas de vous (même si de nombreux sites travaillent pour faire en

sorte que les commentaires que l'on trouve sur le Web au sujet de leur établissement soient, disons... plus positifs que la moyenne...) !

Figure 6-21

Commentaires affichés dans les résultats Google Maps

Conclusion : il est assez simple de se faire indexer dans Google Maps. Un bon positionnement s'obtient ensuite en suivant des règles de bon sens : proposer une description pertinente, un texte significatif, des informations susceptibles d'intéresser l'internaute, etc.

En effet, Google Maps est avant tout un service complémentaire aux sites web classiques, et c'est en ce sens qu'il doit être utilisé. L'aspect éditorial (texte soigné, images

attractives) comptera certainement plus que la technique (mots-clés utilisés, choix de la catégorie, etc.). Pensez à Google Maps comme un annuaire de type « Pages Jaunes » où vous souhaitez faire apparaître une annonce. Qu'est-ce qui pourrait inciter un internaute à cliquer sur votre fiche et à vous rendre visite ?

Ce type de démarche va de plus en plus caractériser le travail du référenceur, qui va devoir optimiser toutes sortes de contenus (cartes, images, vidéos...) et penser avant tout à proposer des services de qualité.

Se référencer aujourd'hui sur Google Maps, c'est aussi préparer demain et anticiper sur la recherche géolocalisée par GPS sur téléphone portable. En effet, ce type de service devrait voir le jour très bientôt, avec l'évolution des technologies et les moyens mis en œuvre par les gros acteurs de la recherche.

Le travail de référencement change donc et évolue... au bénéfice des internautes !

Le SMO (Social Media Optimization)

Les leviers à utiliser pour tirer le meilleur parti de son référencement se multiplient au fur et à mesure des années. Il suffisait, il y a peu de temps encore, de proposer quelques textes dans ses balises meta pour être positionné dans les moteurs de recherche « classiques » et le tour était joué. Aujourd'hui, la quantité de méthodes qu'il est possible d'utiliser pour favoriser son référencement est quasi infinie.

Et parmi les derniers leviers apparus, les réseaux sociaux prennent une place de choix. C'est qu'avec les dizaines de millions d'utilisateurs de Facebook ou les millions de visiteurs uniques de Digg, le marché est plus que tentant pour les responsables marketing de tous bords. Mais comment fonctionne au juste le monde social, et comment en tirer parti dans une stratégie globale d'acquisition de trafic et de référencement ? C'est le but du SMO (*Social Media Optimization*) qui est aux réseaux sociaux ce que le SEO (*Search Engine Optimization*) est aux moteurs de recherche. Indiquons-le cependant dès le départ, le SMO est une stratégie très « tendance », dans la mouvance du succès des réseaux sociaux, et sa réelle efficacité reste encore à démontrer de façon claire, ce qui n'est pas (plus) le cas du SEO, qui bénéficie de nombreuses années d'expériences. Pour le SMO, donc, pas d'affolement, ce qui ne signifie pas que ces stratégies ne sont pas intéressantes, loin de là.

Quels réseaux sociaux utiliser pour son référencement ?

Pour qui s'occupe uniquement de référencement naturel, et donc du positionnement et du trafic des moteurs de recherche – et qui est le sujet de cet ouvrage –, les réseaux sociaux sont avant tout un moyen d'asseoir sa popularité et de mettre en avant ses contenus. Dans ce cadre, c'est avant tout la recherche de liens populaires et qualifiés qui est recherchée. L'audience générée en propre par les réseaux sociaux importe, mais ce n'est pas l'objectif principal de cette approche. Il convient donc d'utiliser les réseaux permettant la création rapide de liens, et surtout de liens visibles par les moteurs de recherche.

Exit donc *a priori* tous les outils demandant une authentification de l'utilisateur pour être parcourus (Facebook – *http://www.facebook.com/*, Viadeo – *http://www.viadeo.com* ou LinkedIn – *http://www.linkedin.com/*, entre autres), l'action devra se concentrer sur les plates-formes de diffusion des liens tels Digg (*http://www.digg.com/*) et ses équivalents francophones (Wikio – *http://www.wikio.fr*, TapeMoi – *http://tapemoi.com/* ou Scoopeo – *http://www.scoopeo.com/*, et bien d'autres) ou sur les *bookmarks* sociaux dérivés de Del.icio.us, encore que ces derniers soient moins parcourus par les moteurs de recherche.

Le travail du référenceur, au sens stratégique, pourra cependant aller plus loin et commencer à prendre en compte le SMO dans un sens plus général, c'est-à-dire la diffusion d'informations et l'acquisition de trafic *via* les réseaux sociaux. Ici, on ne cherche plus seulement à créer des liens, mais on incite les visiteurs et les internautes à diffuser une information, à voter pour celle-ci et à partager un peu de la vie du site.

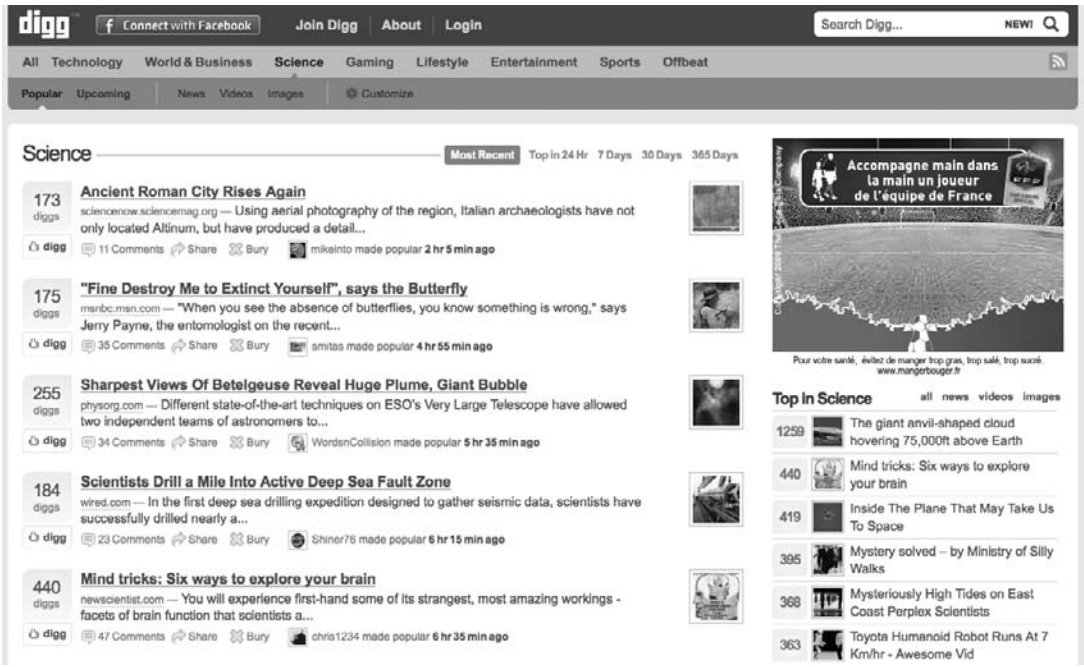


Figure 6-22

Avec ses millions d'utilisateurs, Digg est l'un des outils sociaux les plus populaires du Web.

Dans cette stratégie, les cibles ne s'arrêtent donc plus aux plates-formes ouvertes. Ce sont potentiellement tous les sites qui peuvent apporter un trafic, si possible qualifié, sur un site Internet. Les communautés d'intérêts et les réseaux plus personnels sont alors mis à contribution pour diffuser les dernières offres et les dernières informations d'une marque. C'est dans une stratégie de type SMO que l'utilisation du réseau social du

moment, Facebook, prend tout son sens. Twitter (*http://www.twitter.com/*), autre site star, en sera également un relais évident.

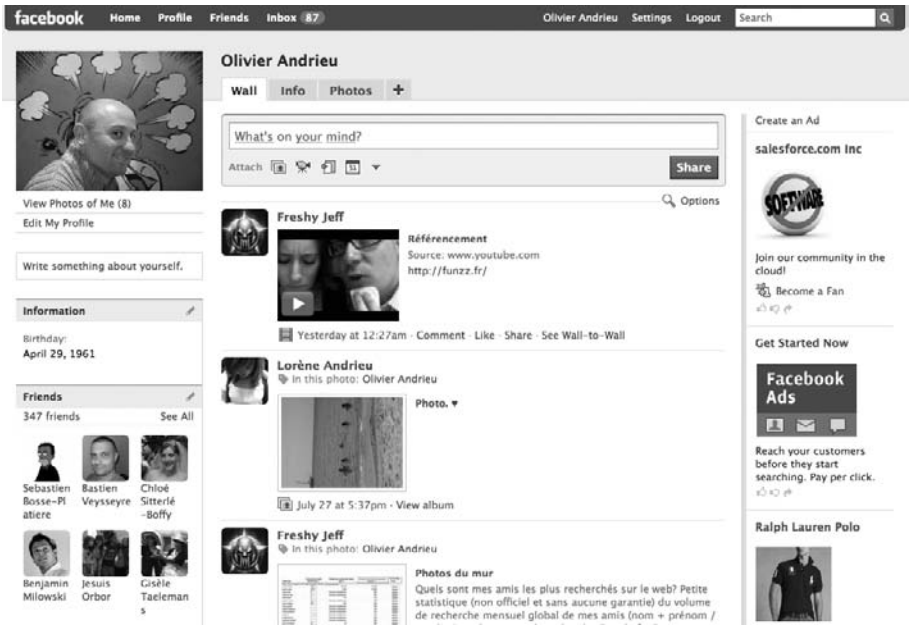


Figure 6-23

Avec son interface authentifiée, Facebook n'a que peu d'influence sur les résultats des moteurs de recherche, même si les profils des abonnés sont publics.

Figure 6-24

Les profils publics Facebook sont en effet indexés par Google, mais pas les informations personnelles des abonnés. Impact obligatoirement limité.

Ceci dit, dans un premier temps, nous envisagerons les réseaux sociaux avant tout comme générateurs de liens potentiels. Leur utilisation pour drainer de façon plus générale du trafic de qualité relève plus de l'e-marketing et moins de la stratégie de référencement proprement dite...

Où trouver des réseaux sociaux ?

Si vous souhaitez utiliser les réseaux sociaux, la toute première question qui vient à l'esprit devrait être : où trouver des réseaux sociaux exploitables, et mieux, intéressés par (ou intéressants pour) mon métier ? Historiquement, les premiers exemples d'annuaires sociaux sont très génériques et se sont plutôt adaptés au goût de leur public. Digg, le tout premier diffuseur massif de liens, reste encore aujourd'hui très généraliste même si sa cible principale est plutôt jeune et technophile. Les moteurs francophones qui sont apparus à sa suite (Scoopeo et Fuzz – avant son arrêt – principalement) ont gardé cette même orientation de par leur public. Sur ces moteurs généralistes, les articles insolites ou étroitement liés au monde d'Internet ont toutes leurs chances.

Des moteurs plus ciblés sont apparus ensuite et ont pris comme choix de restreindre leur activité à un sujet éditorial propre. On nommera ainsi l'américain YardBarker (*http://www.yardbarker.com/*) pour le sport, Sphinn (*http://www.sphinn.com/*) pour le marketing web, etc. Wikio (*http://www.wikio.fr/*) est fortement orienté sur les contenus d'actualité et dispose d'une jolie renommée sur Internet.

Le meilleur moyen d'identifier des moteurs sociaux reste encore de guetter ce que peuvent faire vos concurrents et de s'inspirer de leurs propres stratégies de soumission. Les blogs amateurs, s'il sont un tant soit peu sérieux et s'ils agissent dans votre domaine professionnel, peuvent également vous donner des pistes sur les outils à cibler. En dernier recours, un annuaire comme Digg-like.com (*http://www.digg-like.com/*) peut vous donner d'autres idées.

Un conseil toutefois : privilégiez les outils avec lesquels vous avez une affinité linguistique. Il est très difficile de percer sur un moteur anglophone avec un contenu exclusivement français et non visuel. Pas impossible, notez-le bien, mais difficile.

Avec quel contenu utiliser les réseaux sociaux ?

Il est également important de détailler la meilleure utilisation possible des Digg-like, c'est-à-dire des sites d'accès libre permettant la soumission de contenus et le vote des visiteurs sur la qualité de ces contenus. C'est là, en termes de référencement, le principal facteur de trafic qu'il est possible de rencontrer dans le monde des réseaux sociaux.

L'utilisation des Digg-like sera donc un facteur de génération de trafic et de liens important, certes, mais à quelques conditions dont la principale est avant tout la qualité du contenu soumis. Comme toujours quand on parle de référencement, un contenu de qualité est le prérequis indispensable à un positionnement et à une audience de qualité. Si les articles que vous publiez n'intéressent pas grand monde, et s'ils ne mettent pas en lumière un point de vue original, une information exclusive ou tout simplement intéressante, si vos articles sont mal écrits, alors ne comptez pas sur les internautes pour les plébisciter...

On le répète à l'envi depuis le début de cet ouvrage, c'est le contenu qualifié et original qui est roi (au sens roi = ROI = *Return on Investment* = retour sur investissement) sur Internet. Cette vérité est encore plus flagrante dans le domaine du SMO et des réseaux sociaux. Ici, ce ne sont plus des robots et des algorithmes qui vont évaluer la qualité de vos contenus, mais de véritables humains prompts à la critique. Ils seront d'autant plus difficiles à convaincre !

Oubliez donc cette voie si votre site n'est qu'un simple catalogue de produits ou qu'une simple présentation *corporate* de votre entreprise. Vous pouvez bien entendu soumettre vos pages aux différents moteurs sociaux existant, mais elles ne recevront *a priori* pas grand-chose d'autre que votre propre vote, et n'accéderont bien entendu jamais à la page d'accueil du moins populaire des outils sociaux. Inutile donc de perdre du temps et de l'énergie dans cette tâche, pour votre site, c'est le référencement naturel classique et les moteurs de recherche traditionnels qu'il faut viser !

En revanche, si vous possédez sur votre site un blog ou une plate-forme éditoriale propre, sur laquelle s'exprime, par exemple, votre service marketing ou des personnes influentes de votre entreprise, vous avez peut-être le potentiel pour « jouer » avec les Digg-like. À vous de (sa)voir, en étudiant chacun de vos articles et prises de position, s'ils sont suffisamment originaux pour déclencher l'enthousiasme de lecteurs extérieurs... Si c'est le cas, un petit conseil rapide qui peut s'appliquer également aux blogs : prenez garde aux titres que vous utiliserez à la fois sur les articles, mais également lors de la soumission des contenus. Ceux-ci devront être assez explicites pour que le visiteur sache de quoi vous allez parler. Sur un moteur comme Digg, on zappe rapidement le titre qu'on ne comprend pas. Le moteur brasse des milliers d'informations par jour et ses visiteurs sont là pour trouver un contenu original et éviter le plus possible de jouer aux devinettes. Soyez donc clair sur le sujet abordé, tout en mettant en avant le côté original de votre article. C'est à ce prix que vous attirerez les visiteurs, un équilibre que vous maîtriserez avec l'expérience. Cela tombe bien, la clarté des titres éditoriaux est l'un des *must* du référencement naturel, comme nous l'avons vu auparavant.

Figure 6-25

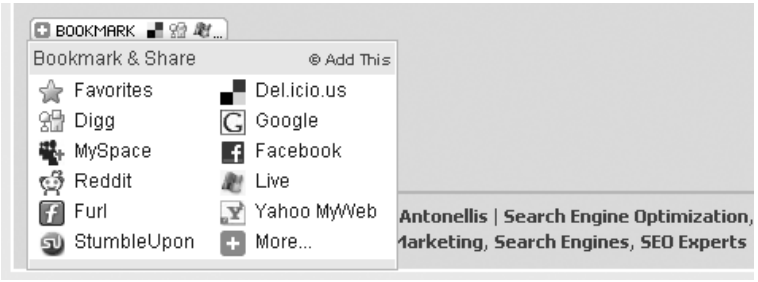
Les blogs sont à la base des réseaux sociaux, au point que Wikio leur consacre une section entière de sa plate-forme collaborative.

Soumettre ou ne pas soumettre ?

La question qui se pose ensuite, si vous disposez de contenu potentiellement intéressant, est celle de la soumission. Devez-vous soumettre vous-même vos contenus aux outils sociaux ou devez-vous laisser vos visiteurs s'en charger ? Biaisons un peu sur la réponse, car il n'y a pas de règle absolue dans ce domaine. Tout d'abord, mettez tous les moyens techniques de votre côté. Assurez-vous de proposer sur vos pages et sur chacun de vos articles un « bouton », ou mieux, un bloc complet, permettant la diffusion et la soumission de vos informations sur les principaux sites sociaux. Si votre blog fonctionne sur une plate-forme Open Source un tant soit peu populaire, il y a fort à parier qu'un plug-in existe déjà, prenant intégralement en charge ce point. Si votre site est développé de manière propriétaire, les moteurs comme Digg proposent dans leur aide un résumé des techniques d'intégration des boutons de vote et autres soumissions automatiques. Disposer de tels boutons, comme montré sur la figure 6-26, c'est s'assurer qu'ils pourront éventuellement être utilisés par des internautes zélés.

Figure 6-26

De nombreux plug-ins proposent un bloc complet de soumission aux outils sociaux pour les blogs et les sites de contenu.



Une fois cet outil en place, le reste des opérations tient plus de l'activité sociale que du référencement. Soumettez de vous-même en priorité les analyses les plus poussées et les contenus que vous avez le plus de raison de mettre en avant. Vous amorcerez ainsi la pompe et permettrez à votre article d'apparaître dans les *Fresh News* des sites sélectionnés. C'est un premier pas.

Le plus dur est ensuite d'obtenir le vote des autres internautes. Pour cela, deux techniques : mobilisez tout d'abord votre propre réseau et vos contacts personnels, s'ils utilisent Digg et les autres outils et demandez-leur de voter pour vos articles. C'est là une base de votes indispensable à la popularité de vos contenus. Cette stratégie est toutefois de moins en moins porteuse dans la mesure où l'algorithme d'un Digg est désormais capable d'identifier les « réseaux organisés » d'utilisateurs.

Une seconde stratégie consiste à acquérir les votes depuis votre site même. L'inclusion, comme on l'a vu précédemment, d'un bouton de soumission permet d'inciter à la publication de votre article. Placez de préférence celui-ci en fin d'article, afin que le vote soit la conclusion logique de la lecture. Et préférez un appel visuel, sous la forme d'une image (en utilisant les logos du Digg-like, exemple sur la figure 6-27) plutôt qu'une

incitation textuelle. Le vote est un acte spontané et instantané, et en tant que tel, il réagit mieux aux *stimuli* purement visuels.

Figure 6-27

Incitez au vote de manière visuelle, et de préférence en fin de lecture.

L'augmentation du nombre de votes *via* ce bouton ou un autre moyen, rapprochera votre article de la page d'accueil du service choisi, et donc de son trafic potentiel. Un bénéfice immédiat. Sur le long terme également, ce bon positionnement en page d'accueil d'un moteur social peut être positif pour votre référencement naturel, à condition toutefois que le Digg-like ciblé veuille bien jouer le jeu. En effet, tous ne servent pas les moteurs et des plates-formes comme Delicious (*http://delicious.com/*) ou autres agrémentent les liens qu'ils présentent d'un attribut `rel="nofollow"`, interprété par les moteurs de recherche comme une invitation à ne pas explorer les pages présentées (voir chapitre 5). Cette implémentation a permis à nombre de sites de réguler le spam dont ils faisaient l'objet mais l'intérêt en termes de référencement est, du coup, bien moindre. Vérifiez donc la présentation des liens sur les moteurs choisis si votre objectif est uniquement lié au référencement naturel.

Figure 6-28

Les liens servis par Delicious sont agrémentés d'un attribut nofollow, oubliez donc cette plate-forme pour le référencement naturel via le netlinking.

On aimerait concevoir la présence sur les réseaux sociaux comme une discipline à part entière, mais elle est indissociable des autres leviers de trafic. Les votants de votre site sont avant tout des lecteurs, et les lecteurs se trouvent *via* les moteurs de recherche ou les flux RSS... La meilleure méthode pour améliorer votre lectorat et votre positionnement sur les réseaux sociaux restera donc encore d'améliorer votre positionnement en dehors des réseaux sociaux. Ceux-ci ne sont pas un moyen d'exister pour des sites orphelins, ils constituent un moyen de compléter son audience sur le court terme ! Sur le Web, comme ailleurs, l'acquisition de trafic et de visibilité représentent un tout cohérent.

Référencement par les widgets

Les widget*s*, ces petits modules qui se glissent dans les pages web et sur les bureaux des systèmes d'exploitation, se sont multipliés de façon exponentielle, d'autant plus qu'il est relativement simple de créer son propre widget grâce à des applications en ligne. Mais comment profiter des widget*s* pour accroître les possibilités de référencement et quels sont les « bons » et les « mauvais » widget*s* pour les moteurs de recherche ? Voici quelques indications pour vous aider...

Widgets et popularité

Un widget n'est porteur de popularité que s'il propose des liens « en dur », pouvant être suivis par les moteurs de recherche. Ce principe permettra de développer considérablement le nombre de liens entrants, lorsque les webmasters insèreront un widget sur leur site web. Idéalement, il vaut donc mieux privilégier un bloc HTML renfermant des balises et des éléments utilisables par les moteurs (texte et image).

Cependant, les widget*s* sont souvent appelés au sein des pages par des codes JavaScript (voir chapitre 7), ce qui implique que la seule solution pour créer des liens utilisables par les moteurs est l'utilisation d'une balise `<noscript>`... Les widget*s* basés sur du Flash sont également globalement illisibles par les moteurs. Voici l'exemple d'un code appelant vos derniers *tweets* (interventions sur Twitter) et les affichant sur la plate-forme Blogger (*http://twitter.com/widgets/blogger*) :

```
<div id="twitter_div">
<h2 class="sidebar-title" style="display:none;">Twitter Updates</h2>
<ul id="twitter_update_list"></ul>
<a href="http://twitter.com/andrieu" id="twitter-link" style="display:block;
➥text-align:right;">Follow me on Twitter</a>
</div>
<script type="text/javascript" src="http://twitter.com/javascripts/blogger.js"></script>
<script type="text/javascript" src="http://twitter.com/statuses/user_timeline/
➥andrieu.json?callback=twitterCallback2&amp;count=5"></script>
```

On le voit, les liens proposés sur les derniers *tweets* sont insérés entre les balises `<script>` et `</script>` et ne sont donc pas lus par les moteurs...

Dans son guide en ligne (*http://www.google.fr/support/webmasters/bin/answer.py?answer=66353&topic=15263*), Google conseille ceci lorsqu'on utilise des commandes JavaScript :

« Placez le contenu JavaScript dans une balise `<noscript>`. Si vous utilisez cette méthode, assurez-vous que ce contenu est strictement identique à celui de JavaScript et qu'il est accessible aux visiteurs qui n'ont pas activé l'option JavaScript sur leur navigateur. »

Ceci laisse donc une certaine marge de liberté : un widget ne sera pas sanctionné s'il propose un lien dans un `<noscript>` ayant une fonction ou un contenu identique à celui qui est proposé à l'internaute dans le `<script>`.

Si on laisse de côté l'aspect popularité, il est quand même important de proposer des liens visibles et cliquables par les internautes, même s'ils sont faits en Flash ou en JavaScript. En effet, ces liens, même s'ils ne sont pas comptabilisés par les moteurs, sont susceptibles d'apporter du trafic vers un site. Or le trafic sur un site web est un critère pris en compte par de nombreux moteurs (dont Google) pour le positionnement d'un site.

En résumé, pour qu'un widget ait de l'impact sur le référencement, il faut toujours lui adjoindre un lien pointant vers le site web qui en est l'initiateur. Idéalement, ce lien devra être « en dur » de façon à ce qu'il soit suivi et comptabilisé par les moteurs de recherche.

Matt Cutts et les widgets

En juillet 2008, Matt Cutts (rappelons ici qu'il est le porte-parole « SEO » de Google) était interviewé par Éric Enge lors du salon SMX Advanced aux États-Unis. De nombreux sujets étaient abordés au sujet du linking, et on y parlait notamment de l'utilisation des widget*s* et de la pêche aux liens (*linkbait*) appliquée aux widget*s*.

Voici quelques extraits, d'après le transcript de l'interview proposé sur le site StoneTemple (*http://www.stonetemple.com/articles/interview-matt-cutts-061608.shtml*) :

« Le widgetbait ressemble au linkbait d'une certaine façon. Nous en avons parlé un peu avec Danny Sullivan [NDA : éditeur du site Search Engine Land (*http://www.searchengineland.com/*)] lors des sessions de questions-réponses de SMX Advanced, influencés par le fait que le premier *widgetbait* que nous avons vu était du spam dans un compteur web. Les gens avaient signé pour un compteur web et ils se retrouvaient avec des liens cachés à l'intérieur du compteur dont ils ignoraient l'existence.

Quelques-uns des critères à prendre en compte sont : Est-ce que les liens sont cachés ? Est-ce que l'image est cliquable ou est-ce que les liens sont enterrés dans un `noscript` ou quelque chose comme ça ? Si c'est le cas, cela ne sera pas très bon pour les utilisateurs. À quel point un widget est-il pertinent ?

Nous voulons que les liens soient comme ceux du linkbait habituel ; nous voulons que les gens soient informés de l'endroit vers où ils créent des liens et nous voulons que les liens soient éditoriaux. Et nous voulons savoir si quelqu'un s'est fait avoir en générant un lien, comme en s'inscrivant à un service sans réaliser qu'un lien allait être superposé à ce service.

Vous pouvez aussi prendre en compte des éléments tels que : Quelle est la cible du lien ? Est-ce qu'il pointe vers l'endroit où vous avez eu le widget ou est-ce qu'il part vers un site tiers complètement différent ?

Un site tiers est souvent hors sujet, et vous pouvez aussi regarder l'anchor text du lien lui-même. Si c'est juste un nom de société ou si c'est un anchor text bourré de mots-clés ou de spam. Et aussi, combien de liens il y a à l'intérieur du widget, et est-ce qu'il y a une tonne de liens enterrés à l'intérieur du widget.

Une chose également intéressante est la façon dont le webmaster a été informé lorsqu'il a placé le widget sur son site. En effet, nous avons déjà vu des widgets où il n'y avait pas vraiment d'informations, ou peut-être enterrées à la fin d'un accord de licence. »

Comme on vient de le voir, le porte-parole de Google est assez réticent à l'utilisation des widgets car il trouve qu'il existe de nombreux « gadgets » douteux qui circulent sur le Web.

Essayons d'en tirer un guide de bonne pratique sur les widgets, à prendre en compte pour le référencement dans Google et les autres moteurs...

Informer les internautes

D'après Matt Cutts, il est important d'informer les utilisateurs de widgets sur les fonctionnalités du mini-programme qu'ils vont installer : À quoi sert le widget ? Qui en est le propriétaire ? Quel en sera l'aspect ? Comment l'utiliser et le désinstaller ? etc. Ce sont quelques-unes des informations qui sont généralement proposées lorsque l'on télécharge un widget « sérieux ».

Voici, par exemple, sur la figure 6-29 les informations affichées sur la page de téléchargement du widget GameOne (*http://www.gameone.net/widget.asp*).

Figure 6-29

Indications fournies pour le widget de GameOne

La plupart des widgets proposent des liens pointant vers des services et informations sur Internet. C'est le cas ici avec une proposition d'actualités consultables sur GameOne. Les liens sont clairement affichés dans le widget, mis à jour toutes les 10 min et facilement consultables.

Ceci est donc un exemple de ce qu'il faut faire. *A contrario*, de nombreux widgets sont montrés du doigt car ils génèrent l'affichage d'annonces publicitaires à l'insu de l'internaute (voir cet article sur les widgets à éviter : *http://moderateur.blog.regionsjob.com/index.php/post/2007/11/22/Pop-up-pop-under-et-publicites-intempestives-%3A-les-widgets-a-eviter*). Ceci est tout aussi répréhensible que la présence d'*adwares* (programmes publicitaires) qui se dissimulent dans certains utilitaires et qui ne sont évidemment pas signalés dans le traditionnel accord d'utilisation.

Éviter le spam dans les widgets

Matt Cutts est assez catégorique dans son intervention : Google n'aime pas les liens cachés et la surabondance de mots-clés qui peuvent être utilisés dans les anchor texts des widgets. Logique...

En ce domaine, il faut considérer un widget comme un morceau de code source : les techniques de « triche » habituelles (*keyword stuffing*, texte caché, voir chapitre 8) sont donc susceptibles d'être pénalisées. On espère seulement que les webmasters utilisant des widgets créés par d'autres, bourrés de spam et téléchargés en toute bonne foi, ne seront pas sanctionnés !

Quid donc des liens contenus dans des balises `<noscript>` ou autres éléments dissimulés dans le code source du widget ? Comme on l'a vu précédemment, il semble y avoir une certaine tolérance de la part de Google, à condition que ces liens reflètent exactement ce qui est affiché dans le widget.

En résumé, tous les liens sortants d'un widget doivent être clairement identifiables par l'internaute, qui doit à tout moment savoir ce qu'il se passe lorsqu'il utilise un widget. Les visites générées « à l'insu du plein gré » de l'internaute sont à proscrire absolument.

Privilégier les liens éditoriaux

Google n'aime pas les liens cachés ; il n'aime pas non plus les blocs de liens et autres éléments qui n'apportent aucune information pertinente pour les internautes. Ce qui vaut pour une page web vaut également pour un widget : il vaut mieux privilégier des liens basés sur du texte pertinent plutôt que des liens basés sur des images ou des successions de mots-clés.

Pour avoir une idée de ce qu'il faut faire, on peut se reporter au gadget Google News pour page web (*http://googleblog.blogspot.com/2009/02/add-google-news-to-your-website.html*).

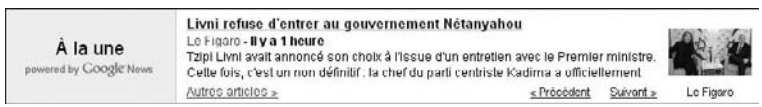


Figure 6-30

Widget proposé par Google News

Il s'agit d'un petit module basé sur une *iframe*, qui affiche en temps réel les informations de Google News. Non seulement les liens sont en dur, mais ils sont également « éditoriaux » car basés sur des textes pertinents (titres d'articles). Le code en est le suivant :

```
<iframe frameborder="0" width="300" height="250" marginwidth="0" marginheight="0"
➥src="http://www.google.com/uds/modules/elements/newsshow/iframe.html?q=google%2
➥Creferencement&ned=fr&rsz=small&hl=fr&format=300x250"><br /></iframe>
```

La limitation de ce système est celle de la taille (en pixels) du widget : il s'agit généralement d'une petite zone de texte qui ne permet pas d'afficher beaucoup d'informations à la fois, à moins d'obliger les internautes à des opérations de *scroll* pas très pratiques. Autre limitation, bien sûr : les liens étant dans une *iframe*, l'impact sur le référencement naturel (en termes de popularité et de réputation) est assez limité... Mais il est possible d'indiquer dans le widget un lien en dur vers le site source :

```
Toute l'actualité avec <a href="http://news.google.fr">Google News</a> :<BR>
<iframe frameborder="0" width="300" height="250" marginwidth="0" marginheight="0"
➥src="http://www.google.com/uds/modules/elements/newsshow/iframe.html?q=google%2
➥Creferencement&ned=fr&rsz=small&hl=fr&format=300x250"><br /></iframe>
```

Le lien vers Google News sera alors repris par les moteurs sur toutes les pages affichant ce widget... Rien ne vous empêche de faire la même chose avec votre site web.

Privilégier les liens thématiques

Dans son intervention, Matt Cutts pointe du doigt les liens hors sujet, qui partent depuis un widget vers un site tiers, qui n'a rien à voir avec la thématique abordée.

Bien évidemment, on parle ici de la thématique abordée par le widget, il serait illusoire de demander aux webmasters d'utiliser des widgets en accord avec la thématique de leur propre site ! Quoique ce serait certainement bien plus bénéfique pour le référencement... Google ne semble pas apprécier qu'un widget pointe vers une page hors sujet par rapport au contenu du widget (comme un widget météo qui pointe vers un site d'utilitaires web).

Lors de la création d'un widget, il faut donc penser à la thématique abordée par son site et créer un gadget qui exploite cette thématique.

C'est un peu comme si l'on proposait une bannière publicitaire à insérer sur un site partenaire : il ne viendrait à l'idée de personne d'utiliser un lien du type « voiture pas chère » pointant vers un site d'actualité cinématographique...

La création d'un widget impose donc une certaine réflexion sur les services et pages web qui vont être mises en valeur. Est-ce qu'il s'agit de cibler tout ou partie de son site ? Quel anchor text (texte du lien qui pointe vers le site) ? Le contenu des pages ciblées est-il amené à changer ?

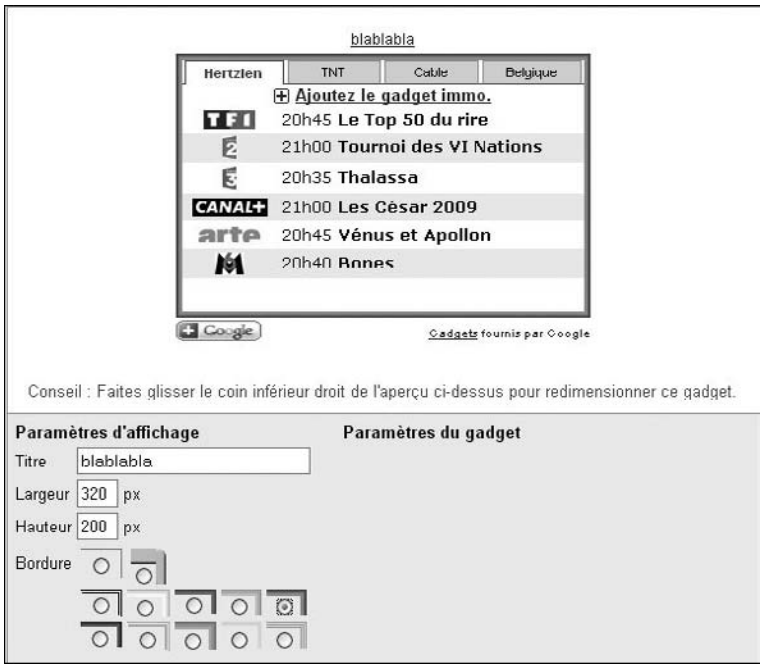
Permettre la personnalisation des widgets ou pas ?

Ce dernier point est évidemment intéressant pour les webmasters voulant maîtriser complètement les fonctionnalités d'un widget. De cette façon, ils pourront choisir d'éliminer ou de modifier les liens sortant qui ne leur conviennent pas, mais ceci est également dangereux pour le concepteur du widget.

Ainsi, la bibliothèque de gadgets Google (*http://www.google.com/ig/directory?hl=fr&synd=open*) propose de nombreux widgets qui sont facilement personnalisables, grâce à une interface graphique. Prenons l'exemple d'un widget de Programme TV (*http://www.programme.tv/*) : en quelques secondes, l'anchor text « Programme TV » du lien pointant vers le site *www.programme.tv* a été remplacé par « blablabla ».

Figure 6-31

Widget du site Programme.tv

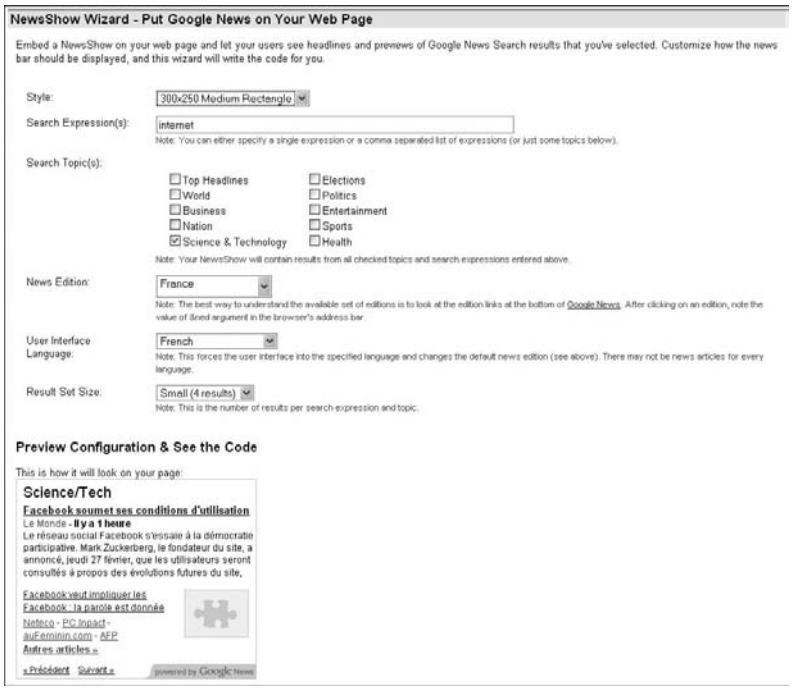


Voici comment casser complètement la pertinence d'un lien, en générant un anchor text complètement hors-sujet. Prenons un autre exemple : la façon dont on peut personnaliser le gadget Google News (*http://code.google.com/apis/ajaxsearch/documentation/newsshow/wizard.html*) est à notre avis beaucoup plus pertinente. Il s'agit cette fois non pas de modifier l'anchor text d'un lien, mais de choisir quels liens seront affichés par le widget.

À l'aide de quelques manipulations, il est par exemple aisé de choisir un affichage en bloc plutôt qu'en bande, de puiser dans les actualités Science et technologie, et de privilégier l'actualité française centrée sur Internet.

Figure 6-32

Interface de configuration du widget Google News



Voici donc un exemple de personnalisation « intelligente » : le webmaster peut choisir l'apparence de son widget (ce qui devrait toujours être le cas) et privilégier l'affichage de certains types de liens (par exemple, des liens susceptibles d'intéresser les visiteurs fréquentant son site).

Si l'on se base sur cet exemple, la stratégie à mettre en œuvre est donc de laisser une certaine liberté aux webmasters, afin de leur permettre d'adapter un widget à leurs besoins. Il ne s'agit pas pour autant de leur laisser modifier en profondeur le widget, ce qui risque de détruire le travail de conception.

Idéalement, pour qu'un widget ne soit pas modifié par un webmaster, il faut qu'il soit appelé depuis une source externe (JavaScript, iframe, Flash) mais dans ce cas, *quid* des liens « en dur » dans le code source, apporteurs de popularité ? Un choix devra être fait en ce sens.

Il faut bien le dire, une certaine anarchie semble régner actuellement sur le monde des widgets et les moteurs de recherche observent désormais de près leurs fonctionnalités et les dérapages éventuels. L'utilisation de widgets est également complexe dans le cadre d'une stratégie de référencement, du fait que les liens qu'ils vont proposer sont le plus souvent invisibles pour les moteurs...

Un widget doit être considéré à la fois comme un mini-programme et comme un morceau de code HTML, à placer sur une page web. Avec quelques règles essentielles à suivre :

- les conditions d'utilisation doivent être clairement définies ;

- un utilisateur de widget doit comprendre tout ce qui se passe quand il s'en sert ;
- le widget ne doit pas renfermer de spam et de texte caché ;
- il faut privilégier des liens thématiques entre le widget et les pages web qu'il cible, et si possible utiliser des liens éditoriaux, et insérer dans ses codes des liens en dur, lisibles par les moteurs.

Ces conseils basés sur du simple bon sens devraient permettre de créer des widgets « SEO Friendly », susceptibles de plaire à la fois aux utilisateurs et aux moteurs de recherche.

Référencement sur les mobiles

Le web mobile est la nouvelle terre promise des créateurs de sites. Ce vaste territoire encore peu colonisé offre de nombreuses possibilités en matière de positionnement marketing, mais la médiatisation autour des mobiles de dernière génération ne doit pas nous faire oublier que l'équipement des « mobinautes » ne leur permet généralement pas d'accéder à des contenus très sophistiqués.

Selon une étude de marché datant de février 2008 (source JDNET : *http://www.journaldunet.com/cc/05_mobile/mobile_internet_fr.shtml*), seulement 20 % des abonnés à des téléphones mobiles sont des mobinautes, c'est-à-dire des utilisateurs du Web grâce à leur terminal mobile. Le marché de l'Internet mobile est donc encore assez limité.

Cela ne nous empêchera pas de vous donner ici quelques conseils pour réussir son site mobile et se faire connaître auprès des mobinautes, même si la notion de « référencement mobile » n'en est qu'à ses prémices....

Faire un site « mobile friendly »

Un site web pour mobile n'est pas conçu de la même façon qu'un site web traditionnel. En effet, il existe un certain nombre de limitations techniques et de normes à utiliser pour rédiger le contenu destiné au mobinaute.

Les *smartphones* sont encore peu répandus en France (même si la percée de l'iPhone et des terminaux sous Androïd change peu à peu la donne), et il faut donc concevoir son site pour des mobiles qui ne possèdent pas de grandes capacités d'interprétation. Oubliez le Flash, les effets dynamiques et autres éléments susceptibles de saturer les ressources d'un navigateur mobile !

Dans ce contexte, plusieurs directives ont été proposées dans le cadre d'une norme W3C pour mobile (*http://www.w3.org/TR/mobile-bp/*).

Il sera, comme d'habitude, intéressant de se référer aux conseils donnés par Google dans son guide en ligne à l'attention des webmasters (*http://www.google.com/support/webmasters/bin/answer.py?answer=40348*) ainsi qu'aux conseils prodigués par le site dev.mobi (*http://mobiforge.com/starting/story/a-beginners-guide-mobile-web-development*).

Voici quelques éléments techniques conseillés lors de la création de votre site :

- utiliser le langage de balisage XHTML Basic 1.1 (*http://www.w3.org/TR/xhtml-basic/*), particulièrement adapté au marché mobile européen ;
- utiliser l'encodage de caractères UTF-8 ;
- prévoir une largeur de page de 120 pixels ;
- utiliser des images GIF ou JPEG ;
- pas plus de 20 Ko par page ;
- se limiter aux styles CSS1 et de préférence les intégrer dans la page ;
- pas de redirection automatique ou de rafraîchissement des pages (meta `refresh`, etc.) ;
- pas de *frameset* ou *iframe* ;
- pas de pop-up ;
- remplacer les tableaux par des calques (`div`) ;
- pas de JavaScript, Flash ou autres éléments ayant besoin de plug-in ;
- utiliser un système de navigation basé sur une liste ordonnée ;
- pas de cookies.

Et oui, il ne reste pas grand-chose...

Validateurs de sites pour mobiles

Il existe plusieurs outils incontournables pour tester son site mobile depuis un PC ou en obtenir une visualisation à partir d'un émulateur :

– Validateur W3C

http://validator.w3.org/mobile/

– Émulateur de mobile

http://mtld.mobi/emulator.php

– Ready.mobi : émulation mobile et correction du code

http://ready.mobi/launch.jsp?locale=en_EN?

Optimiser un site mobile

Les spécialistes s'accordent à dire qu'il est généralement illusoire de vouloir retranscrire le contenu d'un site web sur un site mobile. En effet, l'interface d'un téléphone mobile est encore spartiate et impose des restrictions importantes au niveau de la taille du contenu et de sa sophistication.

Par ailleurs, les offres de téléphonie en France sont encore assez limitées au niveau des échanges de données et de la navigation web : un mobinaute ne va sans doute pas surfer des heures à la recherche d'une information intéressante. Selon les chiffres donnés par

IAB France – Opinion Way (source JDNET, 18 mars 2008 : *http://www.journaldunet.com/ebusiness/mobile/chiffre/080318-iab-opinionway-internet-mobile/index.shtml*), 83 % des mobinautes trouvent l'Internet mobile trop cher pour être utilisé !

Un chiffre est également significatif : selon une étude réalisée en février 2008 (source JDNET : *http://www.journaldunet.com/cc/05_mobile/mobile_internet_fr.shtml*), 66 % des mobinautes ne visitent qu'une page sur un site web mobile.

Il faut également bien saisir qu'un mobinaute ne recherchera pas le même type de contenu qu'un internaute : il sera intéressé par des informations synthétiques et des services pratiques, susceptibles de l'aider ou de le divertir lors d'un déplacement (par exemple, adresse de restaurant, lieux à visiter, actualité du moment, jeux en ligne...)

Si l'on se réfère à nouveau aux chiffres IAB, on voit que l'utilisation de l'Internet mobile correspond aux services mail (63 % des mobinautes), au suivi de l'actualité (60 %), à la consultation de la météo (52 %), à la consultation des informations relatives au trafic (45 %) ou des horaires de transport (42 %).

Les principes de rédaction du contenu peuvent donc se résumer ainsi :

- proposer un contenu synthétique ;
- proposer un contenu facilement accessible et optimiser la navigation ;
- proposer des services et des informations adaptés aux mobinautes.

Figure 6-33

mobile.google.com : un site optimisé pour les mobinautes

En ce qui concerne l'optimisation du site web lui-même et de ses pages, on peut utiliser les mêmes principes que pour le référencement de site web classique, à quelques différences près :

- Créer des titres et des descriptions courtes, pour faciliter l'affichage dans les navigateurs mobiles.
- Ne pas hésiter à utiliser des expressions concurrentielles : en effet, le marché est encore assez ouvert en matière de sites web mobile, autant en profiter...

Les noms de domaine en *.mobi* sont fortement conseillés pour la création d'un site mobile. Ces extensions sont ouvertes à tous depuis 2006, pour une durée minimale de 2 ans.

Concernant le nom de domaine et les URL, il faut faire très simple : en effet, il est beaucoup plus difficile de taper une URL sur son mobile que sur son ordinateur et il faudra éviter les URL « à rallonge », même si elles renferment des mots-clés pertinents.

Soumettre son site dans les moteurs mobiles

Comme pour un site web classique, un bon référencement passe d'abord par la soumission dans les principaux moteurs de recherche.

Dans ce cadre, il peut être intéressant de créer un Sitemap mobile (*http://www.google.com/support/webmasters/bin/answer.py?hl=en&answer=34627*) et de le soumettre à l'aide des interfaces prévues à cet effet.

La norme Sitemap mobile est une extension du protocole Sitemap, qui peut utiliser les langages de marquage XHTML, WML ou CHTML (voir également le chapitre 8 à ce sujet).

Mais l'utilisation efficace des liens reste encore – et toujours – un bon moyen de se faire connaître sur le web mobile.

En effet, un lien bien placé et facilement cliquable depuis un site bien positionné sur le web mobile peut apporter à la fois de la notoriété et du trafic.

Dans ce contexte, il faudra privilégier les liens bien placés et bien visibles, car la navigation est encore assez limitée sur un site web mobile. Les solutions de type « footer » seront certainement moins efficaces qu'un lien placé dans le corps du texte ou même au niveau du menu de navigation.

Les annuaires et autres portails joueront certainement un rôle important pour la notoriété et le trafic sur un site, du moins dans un premier temps. En effet, les possibilités proposées par les téléphones mobiles sont encore limitées (tout le monde ne possède pas un iPhone) et un mobinaute sera plus enclin à consulter un annuaire qu'à faire des recherches compliquées dans un moteur mobile.

Concernant les annuaires proprement dits, il en existe très peu (peu d'équivalents d'un Dmoz sur mobile, par exemple), en dehors du portail Gallery proposé par de nombreux opérateurs mobiles (pour plus d'informations, consultez le portail Pro Gallery à l'adresse suivante : *http://www.pro.gallery.fr/fr/Solution_Gallery/qu_est_ce_que_c_est/index.jsp*).

Quelques annuaires de sites web mobile

Voici quelques annuaires mobiles qui proposent une inscription gratuite :

- *http://www.surfsurmobile.com/*
- *http://www.wmob.fr/*
- *http://www.netoo.com* (annuaire mobile sur *http://m.netoo.com/*)
- *http://www.mobilocom.com*

En conclusion, on peut dire que le Web mobile est encore loin de ressembler au Web traditionnel, à cause des limitations imposées par le support mobile : la taille de l'écran, le débit, les limites du processeur, les langages de balisage supportés sont autant d'éléments qui brident la créativité des webmasters.

Il faudra donc trouver l'équilibre entre les techniques classiques de référencement (présence de mots-clés dans le texte de la page, optimisation des balises, URL pertinentes, linking développé) et les limites du support. Pas question pour le moment de développer un texte conséquent autour des mots-clés !

Le meilleur moyen de référencer son site mobile restera donc essentiellement de proposer un contenu pertinent et approprié pour les visiteurs. Avant de penser à vous positionner dans les moteurs mobiles, pensez aux mobinautes ! Un site fonctionnel et agréable à consulter ressortira de lui-même sur la Toile mobile, surtout s'il est promu sur différents supports (réseaux sociaux, blogs, annuaires et sites web partenaires). La concurrence n'est pas encore aussi forte dans ce domaine que sur le Web « classique ».

Pour terminer sur une note rassurante, l'équipement des mobinautes progresse régulièrement en France, et des offres de navigation et d'échanges de données illimités leur permettront bientôt de surfer facilement sur le Web mobile. D'ici quelques années, le référencement mobile rejoindra le référencement web. Pour le meilleur et pour le pire, bien sûr...

Le référencement audio

Pour conclure ce chapitre sur les différents types de référencements (multimédias, multisupports), nous ne pouvions pas faire l'impasse sur le référencement des sons, puisque les travaux dans ce sens se sont accélérés depuis 2008.

En effet, en septembre de cette année-là, Google a annoncé le lancement de son premier portail consacré à la recherche de documents sonores, baptisé Gaudi pour Google Audio Indexing (*http://labs.google.com/gaudi*). En juillet de la même année, Google avait déjà frappé un grand coup en pleine période électorale américaine en proposant un service baptisé « Speech Recognition », destiné à rechercher des mots-clés dans les discours des politiques américains (voir le communiqué : *http://googleblog.blogspot.com/2008/07/in-their-own-words-political-videos.html*).

À l'époque, il s'agissait d'un « gadget » destiné aux possesseurs d'un compte Google et pouvant enrichir la recherche de vidéos YouTube. Ce module marquait la création de l'équipe « Google Speech Team ». Le service s'est ensuite généralisé vers un portail vidéo dédié, sans que l'utilisateur ait besoin d'installer un module complémentaire. Gaudi (rien à voir, donc, avec le célèbre et talentueux architecte) fonctionne comme un moteur de recherche vidéo classique... sauf que les mots-clés recherchés se trouvent à l'intérieur même de la vidéo, au sein de ce qui est dit dans la bande son. Pour le moment le moteur de recherche interroge exclusivement les « YouTube Political Channels » ou discours des hommes politiques.

Figure 6-34

Google Audio Indexing permet d'effectuer des recherches dans les discours des hommes (et femmes) politiques américains.

Le système est finalement assez simple dans son concept : un algorithme de reconnaissance vocale traduit le son en texte. Une fois les discours disponibles au format textuel, il est « facile » de repérer les mots qui y sont énoncés et de positionner la vidéo à cet endroit-là lorsqu'on recherche un terme donné. Notez bien que ce système est loin d'être nouveau puisque AltaVista, un dinosaure des moteurs de recherche, proposait déjà une fonction assez similaire il y a bien des années de cela...

Comme souvent chez Google, l'interface est pensée en termes d'ergonomie et d'efficacité, et la présence de marqueurs temporels intégrés à la vidéo en rend l'utilisation immédiatement attractive.

Plusieurs informations intéressantes sont proposés par Google dans l'aide en ligne : *http://labs.google.com/gaudi/static/faq.html.*

On y apprend ainsi que l'indexation audio est un étape supplémentaire dans le grand projet de Google : organiser l'information mondiale et la rendre universellement accessible de façon pratique. On s'en serait un peu douté...

Dans ce cadre, les discours des hommes politiques ont été choisis comme terrain d'expérimentation, car des informations clés étaient transmises dans les élocutions. Mais comme l'affirme Google, le domaine des élections américaines n'est qu'une première étape. À terme, on peut imaginer un système qui se généralise à de nombreuses chaînes YouTube.

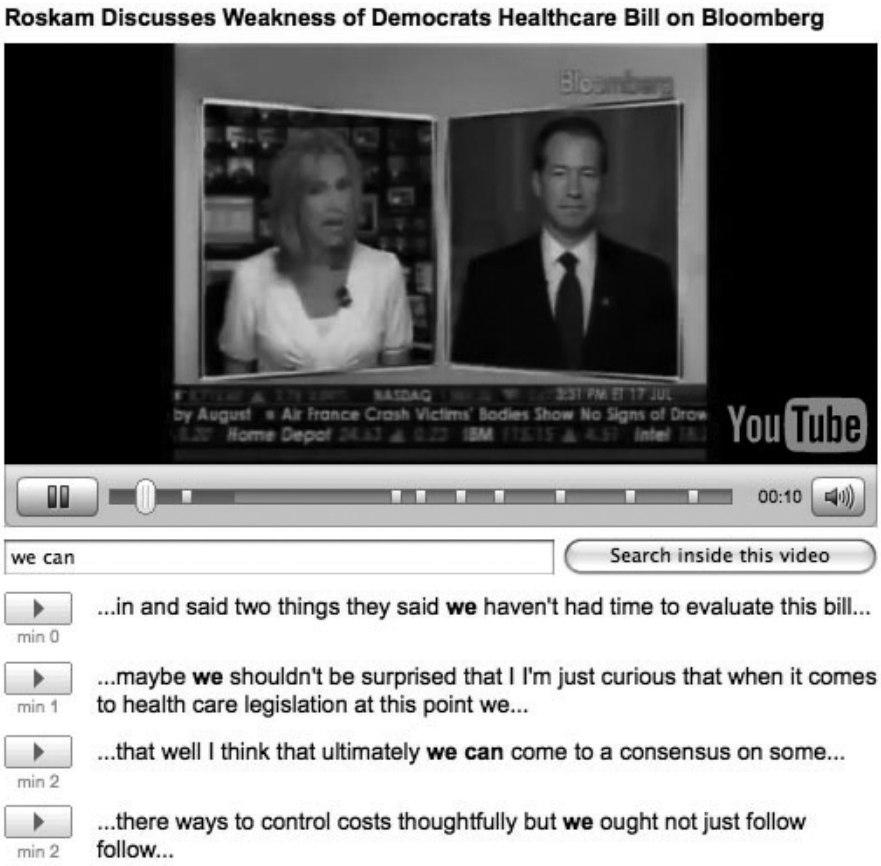


Figure 6-35

Les petites marques jaunes indiquent les emplacements, dans la vidéo, où les mots recherchés sont énoncés.

Sur le fonctionnement de la technologie et du classement des résultats, Google dit simplement qu'il utilise un outil de reconnaissance du langage parlé et que les résultats sont classés d'après le contenu audio, les métadonnées et la fraîcheur.

Blinkx, autre technologie de recherche majeure

Google n'est pas le seul à avoir pensé à la reconnaissance vocale, loin de là : le portail vidéo Blinkx (*http://www.blinkx.com/*) s'y est intéressé dès 2004.

Récompensé en 2008 par le magazine Speech Technology, ce pionnier de la recherche vidéo utilise en effet sa propre technologie pour rechercher des expressions clés dans n'importe quelle langue.

Par exemple, la requête « référencement google » met en avant une vidéo où un artisan apporte son témoignage sur Internet et sur la façon dont il s'est positionné dans Google (figure 6-36).

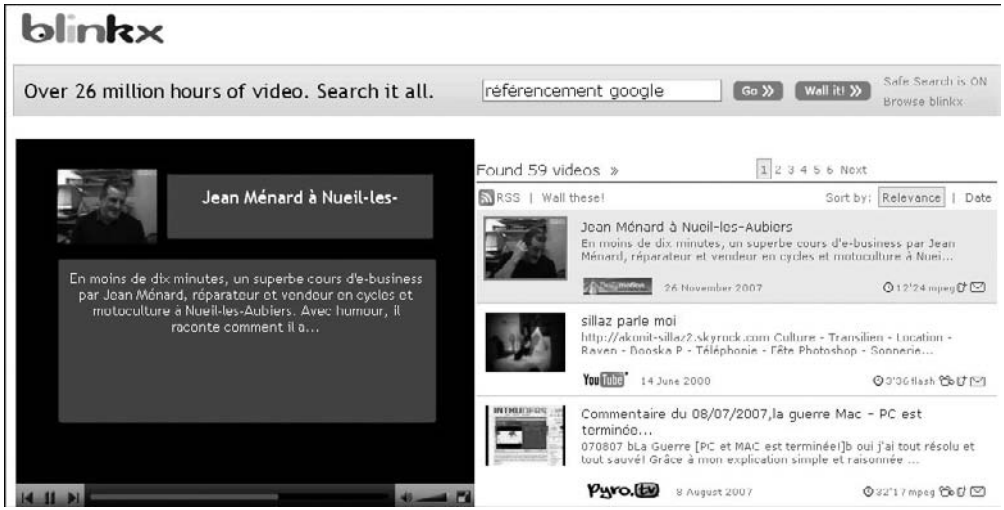


Figure 6-36

Exemple de requête sur le site Blinkx

Blinkx propose donc une technologie qui a fait ses preuves et qui semble assez efficace... D'ailleurs, la société a noué des partenariats avec de nombreux diffuseurs tels que CNN ou la BBC et également avec Microsoft, Lycos, Ixquick ou AOL.

Le point fort de Blinkx est sans conteste sa capacité à interpréter différentes langues (Blinkx gère l'anglais, l'allemand, le français et l'espagnol) et à trouver des vidéos basées non pas sur un mot-clé mais sur une thématique, en explorant le champ sémantique.

Podscope/TVEyes

Podscope, disponible à l'adresse *http://www.podscope.com/,* s'intéresse plus particulièrement aux flux audios présents dans les podcasts.

Lancé en 2005 par TVEyes, il a fait preuve de son efficacité et a été retenu par AOL pour son portail radio (*http://music.aol.com/radioguide/bb*) ainsi que par la société Evoca (*http://www.evoca.com*) consacrée à l'animation audio de sites web (technique dite de *phone-to-web*).

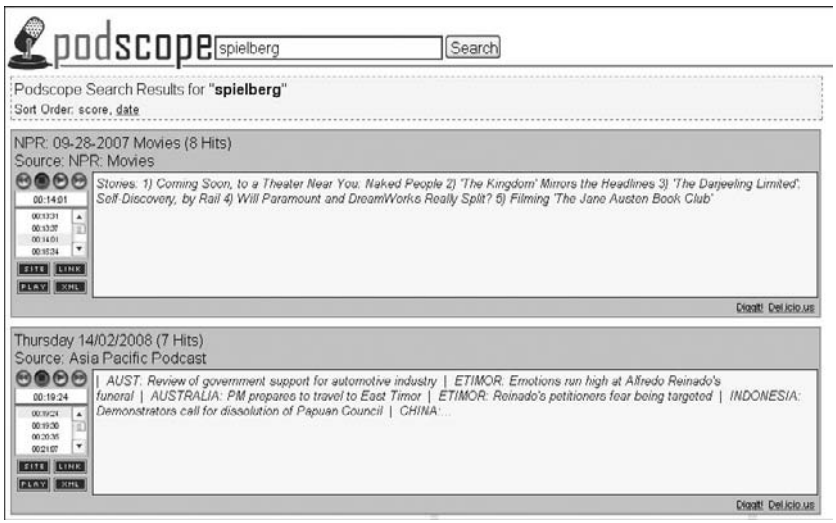
À noter que TVEyes connaît bien son affaire puisque cette société indexe les émissions radio et TV anglophones depuis 1999. (voir interview de David Ives de TVEyes :

http://www.masternewmedia.org/audio_search/audio_search_engines/Podscope_and_TVEyes_search_audio_video_content_20051005.htm)...

Podscope se présente comme un moteur de recherche classique ; une liste d'émissions radio renfermant le mot-clé est proposée à l'internaute, avec des marqueurs temporels. Par exemple, sur la requête « spielberg », on obtient une sélection d'émissions où les projets du cinéaste sont évoqués.

Figure 6-37

Exemple de requête sur Podscope



En théorie, la technologie peut s'appliquer à de nombreux domaines, et non seulement aux flux audio. Ainsi, en avril 2008, TVEyes a annoncé que Podscope pouvait traiter les vidéos, ce qui va ouvrir de vastes possibilités en matière d'indexation. Il est fort possible que la technologie soit bientôt utilisée sur un portail vidéo, qui deviendrait alors un concurrent non négligeable pour Blinkx et YouTube.

L'avenir du référencement audio

Nous venons de voir trois exemples de technologies intéressantes en termes de reconnaissance audio et de transcription textuelle. Il en existe bien d'autres et nombreuses sont celles qui sont encore en test et en gestation dans de nombreux laboratoires, notamment en France... Pour le moment il n'existe aucun guide permettant de faire de l'optimisation audio pour le référencement mais quelques tests sur les outils vont permettre de faire ressortir des principes de base.

Critère numéro 1 : la bande son. Tout d'abord, il est certain que la qualité de la bande son et l'absence de « bruit de fond » seront primordiales. Plus le son est intelligible et la voix reconnaissable (imaginez un système de reconnaissance vocale tentant de comprendre ce que dit une personne située à côté d'un marteau-piqueur...), meilleures seront les

chances de voir le contenu analysé avec efficacité... Pensez-y au moment de l'enregistrement de vos vidéos...

Critère numéro 2 : les occurrences des mots-clés. Gaudi est clairement basé sur les occurrences de mots-clés dans la vidéo, même s'il n'est pas systématiquement vérifié que le fichier renfermant le plus d'occurrences de mots-clés soit placé systématiquement en première position. Il s'agit cependant d'un critère de pertinence majeur pour cette technologie.

Critère numéro 3 : l'ancienneté est également un facteur important parmi les critères qui déterminent l'affichage des résultats. La vidéo la plus récente est le plus souvent placée en première position (et pour le moment ce système de classement n'est pas paramétrable comme sur YouTube).

Critère numéro 4 : la thématique du contenu. La technologie utilisée par Blinkx est un peu plus subtile que celle de Google. En effet, ce moteur ne prend pas seulement en compte les occurrences de mots-clés, mais il évalue également la thématique de la vidéo et la présence de mots-clés connexes. C'est donc l'univers sémantique du domaine traité qui doit être privilégié pour un contenu audio de qualité.

Critère numéro 5 : l'optimisation du discours. Il est tout d'abord évident qu'un mot-clé sera d'autant mieux pris en compte qu'il est facilement intelligible dans la vidéo. Si un mot ou une expression est mal comprise, il y a peu de chances d'obtenir des positions !

On ne connaît pas encore très bien le fonctionnement des logiciels d'analyse de la parole et de reconnaissance vocale, mais il est probable qu'il faille privilégier des expressions bien articulées, avec la prononciation adéquate, et pourquoi pas, un changement de ton permettant de mettre en relief telle ou telle expression. La gestion du phrasé aura sans doute le même effet que les balises `<h1>` ou `<strong>` dans un texte web.

Il est sûr qu'il faudra adapter les discours au fonctionnement des logiciels de reconnaissance vocale pour y insérer des mots-clés pertinents, exactement comme on le fait pour une page web. Les figures de style et jeux de mots risquent de ne pas être bien compris par les robots analyseurs, et que se passera-t-il si un orateur possède un accent marseillais ou alsacien très prononcé (l'auteur de ce livre est originaire du Sud et vit en Alsace : curieux mélange pour une bonne optimisation...) ?

L'internaute aura-t-il le dernier mot ?

Avant de se focaliser sur l'aspect purement technique, il faut aussi penser aux internautes. Ce qui fait le succès d'une vidéo dans YouTube, ce n'est pas sa qualité technique, mais plutôt son originalité et la façon dont elle interpelle l'internaute.

On peut très bien imaginer un classement basé sur le comportement des internautes qui viendrait compléter l'optimisation technique d'une vidéo. Des portails comme Dailymotion ou YouTube ont déjà mis en place un système de vote, qui permet aux internautes d'intervenir eux-mêmes dans le classement.

Dans ce cadre, il ne faut pas imaginer l'indexation audio comme un nouveau système de classement à part entière mais plutôt comme un outil permettant de classer et d'identifier les vidéos. Ce sera ensuite l'internaute qui prendra le relais.

En définitive, il est vraisemblable que l'on va retrouver en audio la même approche que pour une page web-clés et accrocheur pour l'internaute.

Les méthodes ne changeront donc pas en profondeur mais la modification du support fera intervenir de nouveaux spécialistes, tels que des ingénieurs du son, des comédiens, des réalisateurs... Le référencement a encore de beaux jours devant lui !

7

Les contraintes : obstacles ou freins au référencement ?

Votre site est développé en Flash ? Il utilise le langage JavaScript pour sa navigation ? Il s'affiche sous des URL exotiques contenant les caractères `?` ou `&` à en perdre haleine ? Il est structuré en frames (ou cadres) ? Si c'est le cas, vos pages présentent quelques-uns des freins technologiques qui peuvent, aujourd'hui encore, être éventuellement synonymes de blocage pour les moteurs de recherche.

Rassurez-vous, dans la plupart des cas, vous pourrez trouver une solution à vos problèmes. Nous avons essayé, dans ce chapitre, de lister tous les soucis pouvant intervenir lors de l'optimisation d'un site web et de mettre en regard les solutions adaptées.

Nous essaierons de prendre en compte la meilleure optimisation possible des pages du site lui-même et donc la solution technique à mettre en place pour corriger d'éventuelles contraintes… lorsque c'est possible ! C'est heureusement le cas la plupart du temps.

Les frames

Les frames, ou cadres en français, consistent en une technique qui a longtemps été très utilisée pour créer des sites web en plusieurs fenêtres indépendantes dans le navigateur.

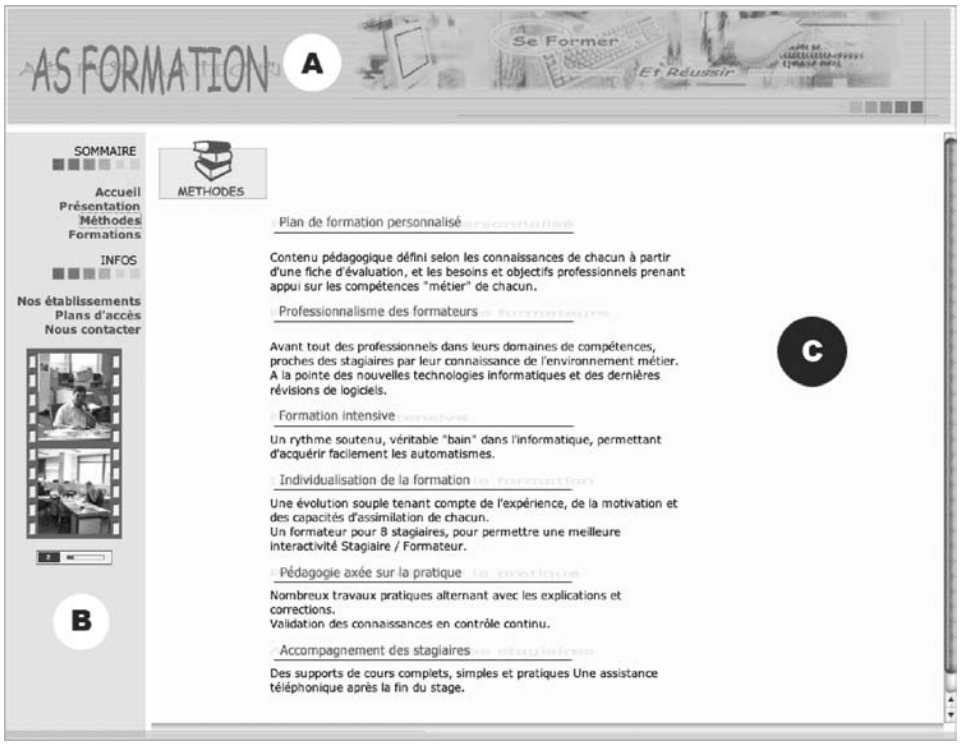


Figure 7-1

Exemple de site web créé avec des frames

Le site représenté à la figure 7-1 est un exemple de cette technique. Il est constitué de trois frames : A et B qui sont des frames de navigation et C qui est la frame de contenu (on voit son ascenseur spécifique à sa droite). Le tout est encapsulé dans une frame dite « mère » qui décrit la taille, l'emplacement et diverses informations sur ces trois frames « filles ».

De façon assez générale, les frames sont très souvent considérées comme un réel obstacle pour les moteurs de recherche et donc pour le bon référencement d'un site qui prendrait en compte cette technique de subdivision de l'écran en différentes fenêtres indépendantes. Ce n'est qu'à moitié vrai : nous verrons plus loin que les frames peuvent même être utilisées pour obtenir un meilleur positionnement (même si elles sont fortement déconseillées par le W3C).

Avant de voir comment les moteurs réagissent lorsqu'ils arrivent sur une page ainsi bâtie, il est nécessaire de dire deux mots de la réalisation de ce type de page en HTML. Imaginons une page web (page mère) qui aurait pour nom `frames.html` et dont le code HTML serait le suivant :

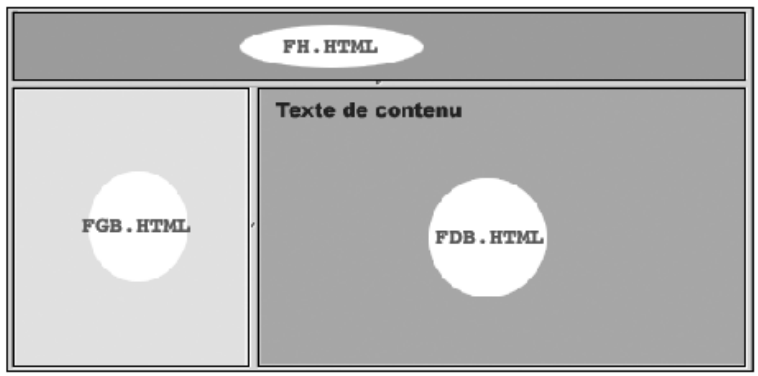
```
<frameset rows=20 %,80 %>
  <frame src="fh.html" name="haut">
  <frameset cols=*,2*>
    <frame src="fgb.html" name="gauchebas">
    <frame src="fdb.html" name="droitebas">
  </frameset>
  <noframes>
    Cette page a &eacute;t&eacute; r&eacute;alis&eacute;e avec des frames.
  </noframes>
</frameset>
```

La balise `<frameset>` … `</frameset>` permet de définir les cadres qui rempliront l'écran du navigateur et d'indiquer quels fichiers HTML (ici `fh.html`, `fgb.html` et `fdb.html`) seront affichés à l'intérieur de ces cadres. Notez bien que le fichier `frames.html`, que nous appellerons fichier mère, sert uniquement à la description des zones de découpage de l'écran et ne contient aucune indication sur le texte ou les images qui y seront affichées.

Les fichiers `fh.html`, `fgb.html` et `fdb.html`, que nous appellerons fichiers filles, contiennent pour leur part les informations à afficher dans chaque partie d'écran. Le fichier mère `frames.html` décrit donc la façon dont les fichiers filles seront affichés sur l'écran (voir figure 7-2).

Figure 7-2

Représentation dans un navigateur du code html indiqué ci-dessus



Mettons-nous maintenant à la place du spider du moteur de recherche qui, dans un premier temps, arrive sur une page de type mère. Il peut avoir trois réactions différentes.

1. Il ignore complètement la page web et ne l'indexe pas, car il a décidé (enfin, ses concepteurs ont décidé pour lui) de ne pas prendre en compte les pages avec frames. Ce type de cas n'existe quasiment plus aujourd'hui sur le Web (mais cela est arrivé dans le passé, sur le moteur Excite, par exemple).

2. Le spider indexe uniquement la page mère et ignore délibérément les fichiers filles en ne suivant pas les liens présents dans les balises `<frame>`. Là encore, ce type de comportement est plutôt rare. De plus, il n'est pas cohérent puisque le « vrai » contenu se trouve dans les pages filles...
3. Le spider indexe les fichiers mère et filles, puis il les considère tous comme des pages web distinctes, sans rapport les unes avec les autres. Si un mot-clé est trouvé, par exemple, dans la page `fdb.html`, le moteur proposera un lien direct vers ce document et non pas vers la page mère `frames.html`. La page fille `fdb.html` s'affichera alors seule dans le navigateur. Le moteur n'a pas pu reconstituer le lien entre le fichier fille `fdb.html` et la page mère `frames.html`. Le contexte des frames est ainsi perdu. L'internaute, qui a cliqué sur un lien dans la page de résultats du moteur, se retrouve avec une page extraite de son contexte de cadres. Pour tout dire, c'est assez gênant et, malheureusement, très courant sur les moteurs de recherche comme le montre la figure 7-3.

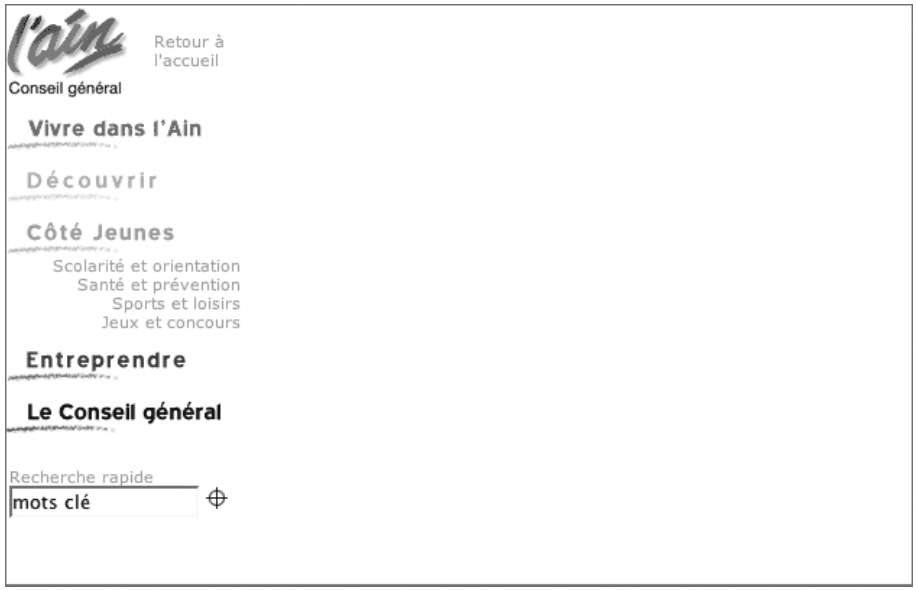


Figure 7-3

Exemple de page trouvée sur un moteur de recherche : la barre de navigation gauche d'un site web, soit une page « fille » faisant partie d'un environnement à frames et ayant perdu ses repères avec sa page de contenu et sa page mère.

Conclusion simple, mais bien souvent irrémédiable : réfléchissez bien avant d'utiliser des frames dans vos pages. Et comme il existe de moins en moins de sites réalisés avec des frames (visualisez les 50 premiers sites mondiaux en termes d'audience et vous verrez vite qu'aucun n'utilise plus cette technique qui n'est pas recommandée par le W3C), il y a peu de chances que les moteurs fassent quoi que ce soit pour mieux les prendre en compte à l'avenir.

Cependant, si votre site est ainsi réalisé, il existe des solutions « miracle », grâce à la balise `<noframes>` et au JavaScript, pour faire en sorte que votre site, bien que réalisé avec des frames, soit bien pris en compte et correctement affiché par les moteurs de recherche.

Optimisation de la page mère

Plusieurs palliatifs peuvent atténuer le fait qu'un site est mal pris en compte lorsqu'il est bâti sur la base de frames : l'emploi, dans la page mère, de bons titres et de balises `<meta>`, qui s'avèrent ici intéressantes (même si ces balises sont aujourd'hui moins bien prises en compte que par le passé par les moteurs majeurs, on pense que dans certains cas extrêmes, comme pour les sites contenant des cadres, elles peuvent avoir une utilité non négligeable), et l'utilisation de la balise `<noframes>` … `</noframes>`, qui permet d'indiquer un texte, à l'origine destiné aux navigateurs n'acceptant pas cette fonctionnalité.

Soignez le texte que vous allez indiquer ici, car il y a de fortes chances, si le moteur ne prend pas en compte les balises meta, pour que seules ces lignes soient affichées dans la description du fichier mère sur la page de résultats du moteur. Si, comme dans l'exemple ci-dessus, votre page a été réalisée avec le code suivant :

```
<noframes>
  Votre navigateur n'accepte pas les frames.
</noframes>
```

voici ce qui s'affichera dans la page de résultats du moteur :

> Chaussures de sport Stela
>
> Votre navigateur n'accepte pas les frames.
>
> http://www.stela.com/index.html – size 1K – 20-Mar-2007

On peut rêver mieux comme description de document, non ? Soignez donc les textes introduits dans la balise `<noframes>` pour qu'ils décrivent parfaitement votre site et les pages en question !

Voici également une astuce qui devrait vous être d'un grand secours lors de la réalisation de vos pages : insérez un lien dans la balise `<noframes>` vers les documents filles qui affichent les liens de navigation internes à votre site.

Pour être plus explicite, reprenons la page de début, intitulée `frames.html`, et adaptons ce fichier à une entreprise fictive appelée Stela. Le code de cette page est le suivant :

```
<frameset rows=20 %,80 %>
  <frame src="fh.html" name="haut">
  <frameset cols=*,2*>
    <frame src="fgb.html" name="gauchebas">
    <frame src="fdb.html" name="droitebas">
  </frameset>
  <noframes>
    Cette page a &eacute;t&eacute; r&eacute;alis&eacute;e avec des frames.
  </noframes>
</frameset>
```

La balise <noframes> est alors remplie avec un texte de remplacement tout à fait commun. Modifions maintenant cette balise comme suit :

```
<noframes>
  <a href="fh.html">Stela</a>, sp&eacute;cialiste de la vente de <a href="fgb.html">
  ➥chaussures de sport</a>, bas&eacute; &agrave; <a href="fdb.html">Paris, France
  ➥</a>.<br />
</noframes>
```

Que se passe-t-il pour le robot qui, dans un grand nombre de cas, ne connaît que le contenu de cette balise <noframes> ? Il va indexer la description fournie (Stela, spécialiste […] France), puis il va suivre les liens proposés vers les fichiers fh.html, fgb.html et fdb.html. Or, ces fichiers contiennent des liens vers les autres parties du site. Vous avez gagné : le spider va alors visiter les pages importantes de votre site et les indexer. En revanche, elles seront enregistrées et visualisées ensuite sans les frames sous la forme de fichiers filles, mais la situation est tout de même bien meilleure que précédemment, où votre site devenait souvent entièrement transparent pour le moteur de recherche.

Soignez le contenu de la balise <noframes>

Tenez compte des mots et expressions importants pour insérer des liens dans la balise <noframes>. Dans l'exemple précédent, nous avons choisi les termes « Stela », « chaussures de sport », « Paris » et « France » pour y insérer les liens, car les moteurs privilégient les termes situés entre les balises <a> et dans leurs calculs de pertinence (voir la notion de réputation au chapitre 5). Alors, autant y proposer des termes descriptifs et représentatifs de votre activité.

Optimisation des pages filles

Il existe également une façon de contourner le fait que les moteurs peuvent proposer un lien vers une page fille devenue orpheline, perdant ainsi le contexte de cadres. Dans le code HTML de chacun de ces fichiers filles (dans la balise <head>), insérez le code JavaScript suivant :

```
<script type="text/javascript">
  <!—
    // Test d'affichage sans l'environnement frames
    if (parent.frames.length==0)
    {
      parent.location.href="pagemere.html";
    };
    // -->
</script>
```

Ce code recrée obligatoirement l'environnement de la page avec des frames en rappelant la page mère. Ainsi, si un internaute tente d'afficher la page fille, ce code appelle la page mère et recrée l'environnement sous forme de cadres de la page. Vous avez gagné ! Indiquez, à la place du nom pagemere.html, le nom de la page mère (ici, frames.html, par exemple) correspondant à chaque page fille au sein desquelles vous allez insérer ce bout de code.

Du travail en plus…

Ce travail – auquel vous n'aviez pas forcément pensé – peut s'avérer considérable si vous avez de très nombreuses pages filles à modifier, mais il sera nécessaire si vous désirez que votre site réalisé avec des frames soit bien référencé. Mais un conseil : à la prochaine refonte de votre site, abandonnez les frames !

Enfin, n'oubliez pas de donner un titre explicite à toutes vos pages filles ainsi qu'aux autres documents qui seront indexés par le spider, car cela constituera une zone de mots-clés importante. De même, insérez des balises meta spécifiques (au moins des balises meta `description`) à chacune des pages filles, puisqu'elles seront traitées comme des pages à part entière par les moteurs. En définitive, traitez les pages filles exactement comme si elles n'étaient pas des parties de frames mais de simples pages HTML. Là aussi, cela peut être synonyme de beaucoup de travail…

Si vous ne voulez (pouvez) pas insérer le code JavaScript de reconstitution de l'environnement sous forme de cadres, et si vous voulez éviter que les pages filles soient indexées par les moteurs (seules seraient prises en compte, dans ce cas, les pages mères), il vous faudra insérer des balises meta `name="robots"` et `content="none"` dans chacun de ces documents. Voir le chapitre 9 à ce sujet. Mais ce serait dommage de ne pas vouloir indexer vos contenus textuels qui se trouvent le plus souvent dans vos pages filles...

Utiliser les frames pour être mieux référencé

Un avantage inattendu des pages réalisées avec des frames, et notamment des documents mères, est que le peu d'éléments graphiques qu'ils contiennent en font des pages souvent très optimisées. C'est d'ailleurs pour cette raison qu'elles sont parfois classées en tête dans les résultats des moteurs. À quelque chose malheur est bon…

La technique du « Frame 100 % » est, d'ailleurs, assez souvent utilisée par certains référenceurs pour pallier un facteur bloquant sur un site web. Exemple : des URL dynamiques qui pourraient être refusées par les moteurs.

Pour illustrer ce fait, prenons l'exemple d'une page répondant à l'adresse *http://www.monsite.com/prod/art/search/?sp-a=00051053-sp00000002&sp-q=stylo&sp-q2=encre&sp-q3=bleu&sp-q4=France/*.

Cette URL étant considérée comme dynamique (elle contient un caractère `?` et plus de trois esperluettes (`&`) – voir la gestion des sites dynamiques plus loin dans ce chapitre), elle risque donc de ne pas être acceptée sur certains moteurs de recherche.

Une astuce consiste alors à l'intégrer à l'intérieur d'une URL statique avec une frame 100 %. Illustration :

```
<!DOCTYPE html …>
<html>
  <head>
    <title>Titre de la page (à choisir avec soin)</title>
    <meta name="description" content="descriptif à choisir avec soin">
    <meta name="keywords" content="mots-clés à choisir avec soin">
```

```
  </head>
  <frameset cols="100 %" border="0" framespacing="0" marginheight="0" frameborder="no">
    <frame src="http://www.monsite.com/prod/art/search/?sp-a=00051053-sp00000002
    ➥&sp-q=stylo&sp-q2=encre&sp-q3=bleu&sp-q4=france" name="main" marginwidth="0"
    ➥marginheight="0" border="0" frameborder="no" framespacing="0">
  </frameset>
  <noframes>
    <body>
      Texte de la balise noframes (à choisir avec soin)
    </body>
  </noframes>
</html>
```

Finalement, que contient cette page ? Un titre, des balises meta et un texte qui ne s'affichera pas sur l'écran du navigateur. Bref, tout l'attirail du parfait référenceur/positionneur, pour peu que son site ait un bon indice de popularité.

En jouant finement sur ces champs, vous devriez arriver à bien référencer et positionner la page web dynamique qui, au départ, aurait posé de nombreux problèmes, ne serait-ce que pour son intégration dans les index des moteurs. Ceci dit, une option de réécriture d'URL sur les adresses posant problème (voir plus loin dans ce chapitre) serait certainement plus efficace...

Attention, cependant, que cela ne vous pousse pas à tenter de spammer les moteurs sous prétexte que le contenu des balises `<noframes>` ne s'affiche pas sur l'écran du navigateur ! En effet, le moteur détectera la supercherie et pourrait vous le faire chèrement payer. En revanche, si vous proposez un bon titre et surtout un contenu textuel pour la balise `<noframes>` très descriptif, contenant des liens, des mots importants mis en exergue (gras, balises `<h1>`, etc.) mais surtout ayant un rapport avec le contenu du site affiché, cela ne devrait pas poser de problèmes. Mais le tout avec modération...

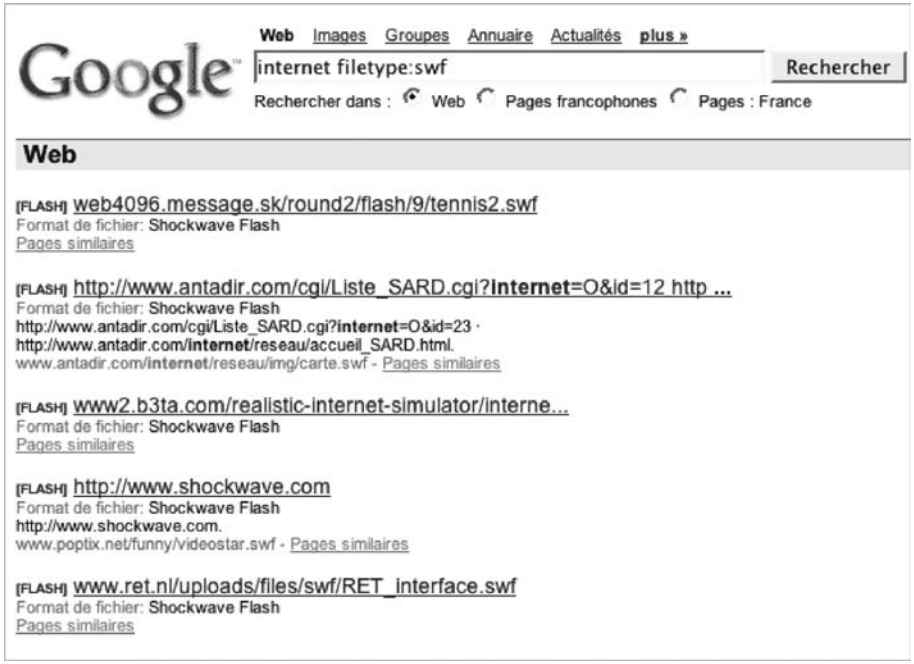
Site 100 % Flash

La problématique des sites réalisés en Flash est finalement assez proche de celle des sites élaborés à l'aide de frames. On le sait, le format Flash de Macromedia/Adobe (fichiers d'extension `.swf`) représente encore un obstacle pour les moteurs de recherche. Il ne s'agit pas obligatoirement d'une problématique axée sur l'indexation. En effet, Google, par exemple, indexe de nombreux fichiers à ce format. Testez la requête « internet filetype:swf » sur ce moteur et vous trouverez près de 1,4 millions d'animations Flash (elles sont indiquées par le moteur de recherche grâce à la mention [FLASH] à gauche du titre – voir figure 7-4).

Même si on peut légitimement penser que Google n'est pas exhaustif – loin de là – sur ce type de fichier, on se rend bien compte ici que leur indexation est techniquement possible. Le problème est plus sur le positionnement. Si, comme nous, vous êtes des utilisateurs assidus des moteurs de recherche, vous n'avez certainement jamais vu une animation Flash dans la première page de résultats sur les requêtes que vous tapez habituellement. De plus, les moteurs de recherche peuvent faire baisser le classement de ces fichiers de façon artificielle dans les pages de résultats pour privilégier les pages au format HTML.

Figure 7-4

Google sait indexer les animations Flash.



Ainsi, en ce qui concerne le Flash, le problème devient de moins en moins d'indexer (référencer) ces fichiers, mais plutôt de les optimiser pour les moteurs de recherche afin de bien les positionner. Il semble bien que cela soit quasi impossible aujourd'hui (et ce malgré la communication importante faite à ce sujet par plusieurs moteurs dont Google).

Google indexe-t-il les fichiers au format Flash ?

Google communique beaucoup sur sa faculté à indexer et analyser du mieux possible le Flash. Vous trouverez ci-dessous quelques liens à ce sujet, mais en pratique, ce format d'animation pose encore de nombreux problèmes en termes de référencement et de positionnement :

- *http://actu.abondance.com/2008/07/google-apprend-mieux-indexer-le-flash.html*
- *http://googlewebmastercentral.blogspot.com/2008/06/improved-flash-indexing.html*
- *http://googleblog.blogspot.com/2008/06/google-learns-to-crawl-flash.html*
- *http://www.adobe.com/aboutadobe/pressroom/pressreleases/200806/070108AdobeRichMediaSearch.html*
- *http://actu.abondance.com/2009/06/google-indexe-les-contenus-externes.html*
- *http://googlewebmastercentral.blogspot.com/2009/06/flash-indexing-with-external-resource.html*
- *http://googlewebmastercentral.blogspot.com/2008/06/improved-flash-indexing.html*

En conclusion, il faut pour l'instant abandonner toute idée de prendre en compte les animations Flash dans votre stratégie de référencement de façon approfondie. Bien sûr,

pour les sites comprenant seulement quelques animations Flash, le mieux sera de ne pas tenir compte de ces fichiers `.swf` et d'optimiser de façon classique les pages web au format HTML du site : titres, textes, liens, etc. On revient ici à une optimisation normale du site, excluant le Flash.

Cependant, si votre site est majoritairement constitué d'animations Flash, comment faire ? Dans un premier temps, voyons comment est structuré un fichier Flash. Lorsque vous créez une animation, vous obtenez un fichier nommé, par exemple, `anim.fla` (l'extension `.fla` est caractéristique du format Flash). Pour afficher ce fichier dans une page web, il est nécessaire de l'exporter au format Shockwave Flash (extension `.swf`). C'est ce fichier, une fois exporté, que vous allez utiliser pour votre site web.

Si l'animation réalisée contient du texte, celui-ci ne sera pas – ou mal – pris en compte par le moteur de recherche. Mais le fichier HTML qui lance l'animation Flash est, lui, pris en considération. Dans ce cas, le remède sera très proche de celui déjà indiqué pour les frames (voir précédemment) : utilisation optimisée des balises `<title>`, meta et, dans notre cas, `<noembed>` (de la même façon qu'avec la balise `<noframes>`). Dans la balise `<noembed>`, essayez d'insérer le plus de texte possible (sans spammer), afin que celui-ci soit « aspiré » par les spiders des moteurs. Nous reviendrons sur ce point plus en profondeur dans le paragraphe suivant.

La situation idéale consistera, sinon, à développer un site en HTML conjointement à la version Flash ou, au moins quelques pages pour contenter les robots. Dans ce cas, faites attention à ce que le lien sur la page d'accueil vers cette version HTML puisse être suivi par les robots (ne l'insérez pas dans l'animation Flash, pas de JavaScript, etc.). Il serait dommage de développer ces pages pour rien.

Des « rustines » pour mieux indexer le Flash ?

On l'a vu, un objet Flash (extension `.swf`) est par nature indéchiffrable (ou mal compris) par les moteurs de recherche, même si certains d'entre eux (par exemple, Google) utilisent l'outil Search Engine SDK Technology (*http://www.adobe.com/devnet/flashplayer/articles/swf_searchability.html*) développé par Adobe pour extraire des informations clés de ce type de fichier.

La plupart du temps, quand on parle de référencement Flash, on parle donc d'optimiser le contenu des pages web (au format HTML) « hors Flash » et de proposer ainsi des éléments lisibles par les moteurs de recherche.

L'enjeu d'un bon référencement de ce type de site est donc de proposer un contenu lisible par les moteurs tout en conservant l'attractivité d'une animation Flash, susceptible d'attirer les internautes. Pas si simple...

Techniques problématiques

Pourtant, il existe de nombreux moyens de proposer un contenu accessible uniquement aux moteurs de recherche, tout en conservant les animations Flash sur le site. Le problème est que la plupart des techniques sont sanctionnées par les moteurs de recherche.

Insérer du texte « caché »

Par définition, un texte présent dans le code source mais non visible par l'internaute est un texte caché. Son existence peut donc compromettre la prise en compte par les moteurs de recherche, car il s'agit souvent d'une technique de « triche ».

Par définition également, une animation Flash est contenue dans un fichier SWF. Celui-ci est intégré dans le code source par le biais d'une balise `<object>` ou `<embed>` (remarque : la balise `<object>` est actuellement conseillée par le W3C).

Exemple d'intégration Flash dans une page HTML :

```
<object type="application/x-shockwave-flash" data="anim.swf" width="550" height="400">
  <param name="play" value="true" />
  <param name="movie" value="anim.swf" />
  <param name="menu" value="false" />
  <param name="quality" value="high" />
  <param name="scalemode" value="noborder" />
</object>
```

Les techniques suivantes sont souvent conseillées pour insérer des éléments textuels compréhensibles par les moteurs (voir articles aux adresses suivantes : *http://www.journaldunet.com/solutions/0601/060120_referencement-flash.shtml* et *http://www.webrankinfo.com/referencement/contenu/flash.php*) : insérer du texte à l'intérieur de la balise `<object>` ou dans une balise `<noembed>`.

Exemple d'optimisation à l'intérieur de la balise `<object>` :

```
<object type="application/x-shockwave-flash" data="anim.swf" width="550" height="400">
  <param name="play" value="true" />
  <param name="movie" value="anim.swf" />
  <param name="menu" value="false" />
  <param name="quality" value="high" />
  <param name="scalemode" value="noborder" />
  texte de présentation de mon animation
</object>
```

Exemple d'optimisation dans une balise `<noembed>` :

```
<embed src="moviename.swf" width=100 height=80
pluginspage="http://example.com/shockwave/download/">
</embed>
<noembed>
Texte de remplacement pour mon animation
</noembed>
```

Il est pourtant difficile de résumer le contenu d'une animation Flash en quelques lignes et l'on peut donc se retrouver avec un paragraphe conséquent de texte caché dans le code source, ce qui est toujours dangereux car cela peut être pris pour de la fraude. Ce type d'élément peut être pénalisé par les moteurs, à moins qu'ils reconnaissent le bien fondé de la technique. Mais il existe potentiellement un risque de sanction.

Extrait du guide de qualité Google (*http://www.google.fr/support/webmasters/bin/answer.py?answer=66353*) :

« Si votre site contient du texte et des liens cachés conçus pour induire les moteurs de recherche en erreur, votre site peut être retiré de l'index Google et ne plus être affiché dans les pages de résultats de recherche. Lorsque vous évaluez votre site afin de vérifier s'il contient du texte ou des liens cachés, recherchez tout ce qui n'est pas facilement affichable par les visiteurs. Existe-t-il du texte ou des liens accessibles uniquement aux moteurs de recherche et non aux visiteurs ? »

Il s'agit donc d'une technique à utiliser avec parcimonie. Les contenus de type « liste de liens » ou « succession de mots-clés » sont, par exemple, à proscrire.

Dans tous les cas, la question suivante doit être posée : pourquoi ne pas proposer les éléments directement dans la page plutôt que sous une forme alternative ? Par exemple, un menu HTML en pied de page est de loin préférable à l'utilisation d'éléments alternatifs à un menu Flash.

Faire du cloaking

Le cloaking est une technique consistant à proposer aux moteurs de recherche une page différente de celle qui est vue par les internautes. Nous en reparlerons dans ce chapitre lorsque nous aborderons les sites dynamiques.

Dans le cadre du référencement Flash, une méthode consiste alors à détecter la capacité du visiteur à lire les animations (détection du plug-in navigateur, par exemple) et à le rediriger automatiquement vers une page textuelle s'il n'est pas capable de lire l'animation.

Une autre méthode consiste à insérer l'animation Flash à l'aide d'un script JavaScript. Le script n'étant pas utilisable par les moteurs, ces derniers verront une page purement textuelle.

Dans tous les cas, il y a bien présentation d'un contenu différent aux internautes et aux moteurs de recherche.

Extrait du guide de qualité Google (*http://www.google.fr/support/webmasters/bin/answer.py?answer=66355*) :

« Le cloaking est la pratique qui consiste à présenter aux utilisateurs des URL ou un contenu différents de ceux destinés aux moteurs de recherche. En raison de la présentation de résultats différents selon le *user-agent*, votre site peut être considéré comme trompeur et être retiré de l'index Google.

Exemples de cloaking :

- Présentation d'une page de texte HTML aux moteurs de recherche, mais affichage d'une page d'images ou Flash aux utilisateurs.
- Présentation aux moteurs de recherche d'un contenu différent de celui destiné aux utilisateurs.

Si votre site contient des éléments non explorables par les moteurs de recherche (par exemple, fichiers Flash, scripts JavaScript ou images), vous ne devez pas leur fournir de contenu masqué. »

Pour les moteurs, il n'existe donc pas de « bon cloaking » et de « mauvais cloaking », un internaute, ou un moteur, doit toujours avoir la possibilité d'accéder à la version de son choix.

Plutôt que de faire du cloaking, il est préférable de créer un lien `<a href...>` visible et pointant vers une version HTML du site. Voici un exemple sur le site Mobalpa (*http://www.mobalpa.fr/fr*).

Figure 7-5

La page d'accueil du site propose deux versions à l'internaute et au spider : une en HTML et une en Flash. Intéressant même si l'on rajoute un clic pour accéder à la véritable page d'accueil.

Utilisation du script sIFR

Google recommande notamment l'utilisation du script sIFR, un projet Open Source qui permet aux webmasters de remplacer des éléments textuels par des équivalents Flash (*http://www.google.fr/support/webmasters/bin/answer.py?answer=72746*).

Le portail Wiki sIFR (*http://wiki.novemberborn.net/sifr/What+is+sIFR*) et un article de Mike Davidson (*http://www.mikeindustries.com/sifr*) apportent des informations complémentaires à ce sujet.

Le principe du sIFR (pour *Scalable Inman Flash Replacement*) est de remplacer les éléments textuels des pages par des éléments Flash. Il s'agit donc de l'ajout d'une couche technologique par-dessus le code source, laissant la possibilité aux moteurs d'accéder au contenu de base.

Ce script est utilisé principalement pour la mise en forme particulière de contenu texte (utilisation d'une police spéciale, par exemple) sous forme d'animation Flash.

Le processus est le suivant :

- une page HTML ou XHTML est chargée par le navigateur ;
- un JavaScript détecte si le player Flash est installé ;
- si le player Flash n'est pas installé ou si le navigateur ne supporte pas JavaScript, la page web se charge normalement et présente un contenu texte ;

- si le Flash est supporté, le script insère des animations Flash par-dessus les éléments de la page, en récupérant les données texte. Les animations affichent le texte à l'aide d'un code ActionScript.

Le résultat peut être observé sur la page *http://www.mikeindustries.com/blog/files/sifr/2.0/* où l'on peut activer (figure 7-6) ou désactiver (figure 7-7) le script sIFR.

Figure 7-6

Activation sIFR

Figure 7-7

Désactivation sIFR

Le script sIFR offre donc des performances intéressantes : le remplacement de police est totalement transparent et l'internaute a toujours la possibilité de sélectionner le texte. En effet, il va choisir en réalité le contenu texte qui se trouve sous le Flash.

Il est également possible de substituer des images à certaines animations, comme le montrent d'autres exemples donnés sur *http://wiki.novemberborn.net/sifr/Examples*.

Pourquoi la méthode sIFR est-elle privilégiée par Google ? Probablement parce qu'elle est totalement transparente pour l'utilisateur et affiche exactement le même contenu pour un internaute et un moteur. De plus, il ne s'agit pas d'optimiser une animation Flash présente dans un code source mais de superposer du Flash à un contenu texte qui se trouve sur le site. Les grands principes de l'accessibilité (notamment la notion d'enrichissement progressif) sont ainsi respectés.

Remarque : Que se passe-t-il si un internaute utilise des modules de blocage de Flash (comme Flashblock pour Mozilla Firefox) ? Dans ce , le script sIFR est désactivé et le site s'affiche comme si l'internaute ne pouvait pas voir le Flash. Il n'y a donc aucune pénalisation au niveau de l'affichage de la page web.

swfIR pour les images

Notons également le format swfIR, pour *swf Image Replacement*, qui est l'équivalent du format sIFR pour les images... Pour plus d'informations, consultez le site suivant : *http://www.swfir.com/*.

La méthode sIFR est particulièrement adaptée aux animations « basiques » et montre donc ses limites pour des animations plus complexes, à base de cinématiques. Mais, Google ayant clairement indiqué qu'il s'agit là d'une technique « permise », elle est à prendre en compte lors du développement d'un site Flash. En espérant que des webmasters peu scrupuleux ne tentent pas de la détourner en allant trop loin pour être mieux positionné...

SWFObject

Une autre méthode pour proposer des versions textuelles d'animations Flash est la fonction JavaScript `SWFObject` (*http://code.google.com/p/swfobject/*) qui détecte la version du plug-in Flash utilisé par le navigateur de l'internaute et envoie, en fonction de celle-ci, un contenu différent qui peut être textuel. Mais n'est-ce pas du cloaking ? La question reste ouverte mais le projet est disponible sur le site Google Code...

Langages JavaScript, Ajax et Web 2.0

Le JavaScript peut servir à de nombreuses possibilités « cosmétologiques » pour agrémenter l'aspect visuel d'un lien : *roll-over*, menus déroulants (voir figure 7-8), etc. De nombreux sites développés en Ajax utilisent notamment ce langage. On a tendance à dire, un peu rapidement parfois, que les liens JavaScript ne sont pas pris en compte par les robots des moteurs de recherche. Ce n'est pas réellement exact. En fait, les liens écrits en JavaScript doivent surtout être « spider compatibles ».

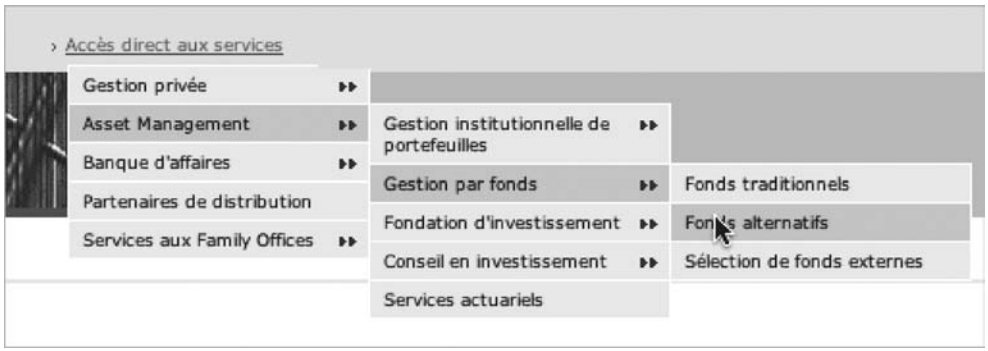


Figure 7-8

Exemple type de menu de navigation écrit en JavaScript

Un lien écrit en JavaScript compatible pour les moteurs, sera suivi par les robots des moteurs. N'oubliez pas que les attributs `title` des liens (balises `<a>`), que certains emploient pour pallier ces problèmes, ne sont pas pris en compte par les moteurs (voir fin du chapitre 4). En revanche, un lien JavaScript non compatible hypothèquera grandement

l'exhaustivité de l'indexation de vos pages par les moteurs car les spiders ne les reconnaîtront pas. Ne l'oubliez pas !

Ainsi, un spider comme Googlebot (le robot de Google), lors de son arrivée sur votre page d'accueil, va tenter de suivre les liens qui y sont présents pour découvrir d'autres pages afin de les indexer également. Si le lien est classique, c'est-à-dire de la forme :

```
<a href="http://www.votresite.com/page-distante.html">
  Texte du lien
</a>
```

cela ne lui posera aucun problème. Il suivra fidèlement ce lien pour indexer la page distante. Tout ira pour le mieux dans le meilleur des mondes.

En revanche, tout se complique si le lien est créé à l'aide d'un code JavaScript. Notez qu'il existe plusieurs façons de décrire un lien en utilisant le langage JavaScript. En voici quelques exemples, parmi de nombreux autres :

```
<a href="javascript:window.open('http://www.votresite.com/page-distante.html',
➥'newWindow')">Texte du lien</a>

<a href="#" onClick="javascript:toto()" >Texte du lien</a>

<a href="javascript:fonctionlambda()">Texte</a>
```

La page pointée dans le premier exemple ci-dessus, présente à l'adresse *http://www.votre-site.com/page-distante.html/*, ne sera donc pas visitée par les spiders… pas par ce biais tout du moins (même s'il se murmure depuis bien longtemps que Google serait en train de développer un robot capable de suivre des liens dans du JavaScript). Par ailleurs, dans la mesure où les liens constituent la meilleure façon d'indexer une page sur un moteur, de façon bien plus efficace qu'à travers un formulaire de type « Add URL » des outils de recherche (voir chapitre 8), on se rend compte des soucis que peuvent causer les liens JavaScript dans le cadre de la bonne indexation d'un site.

Comment faire du JavaScript « spider compatible » ?

Heureusement, il est possible de créer des liens JavaScript qui soient bien interprétés par les robots. Par exemple, voici le même lien que précédemment, mais rendu compatible :

```
<a href="page-distante.html" onclick="window.open(this.href); return false;">
  Texte du lien
</a>
```

ou :

```
<a href="http://www.votresite.com/page-distante.html" onclick="window.open(this.href);
➥return false;">
  Texte du lien
</a>
```

Le fait que l'adresse de la page distante se trouve maintenant dans la zone `href` permet au robot de la reconnaître et de la suivre pour indexer le document. En revanche, lorsque l'internaute cliquera sur le lien, c'est l'action JavaScript (`onclick`) qui sera prise en compte et qui se déroulera.

Les choses évoluent petit à petit

En juin 2009, Google a fait savoir qu'il travaillait sur l'analyse et la compréhension de certains liens JavaScript contenant l'événement `onclick` et notamment ce type de lien :

- `<div onclick="document.location.href='http://foo.com/'">`
- `<tr onclick="myfunction('index.html')"><a href="#" onclick="myfunction()">new page</a>`
- `<a href="javascript:void(0)" onclick="window.open ('welcome.html')">open new window</a>`

Pour plus d'informations, consultez le site suivant : *http://actu.abondance.com/2009/06/google-sur-le-point-de-reconnaitre-les.html*.

Il est également plus rapide d'écrire `this.href`, option qui permet de simplifier l'écriture et la maintenance puisque `this` représente l'objet courant, donc la balise `<a>`. `this.href` est alors égal à l'URL indiquée juste à gauche. On aurait pu écrire :

```
<a href="http://www.votresite.com/page-distante.html" onclick="window.open
➥('http://www.votresite.com/page-distante.html'); return false;">
  Texte du lien
</a>
```

mais cela aurait été plus long…

Votre code devient ainsi compatible à la fois pour l'internaute et les robots. N'hésitez pas à regarder à quoi ressemblent vos liens et à modifier leur forme si cela vous est possible. Vous faciliterez ainsi grandement la vie des spiders de Google, Yahoo! et autres Bing.

Il est malheureusement très complexe de prendre en compte toutes les façons de créer un lien en JavaScript (notamment en fonction du type d'affichage que vous désirez obtenir en ligne), mais retenez que les codes HTML des liens optimisés pour les moteurs doivent proposer à la fois :

- un attribut `href` contenant l'URL de destination pour les robots ;
- une zone JavaScript propre à l'action que vous désirez créer en cas de clic ou de survol par la souris.

Si ces deux conditions sont réunies, on peut penser que tout devrait bien se passer pour vos pages… et pour votre référencement !

Créer des menus autrement qu'en JavaScript

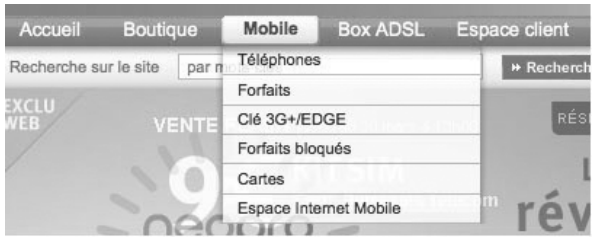
Pour créer des menus déroulants sympathiques et faciles à utiliser pour les internautes, il est loin d'être nécessaire de les réaliser en JavaScript. La tendance actuelle, largement

suivie par la majorité des développeurs, est de suivre d'autres voies qui, quant à elles, sont tout à fait compatibles avec les moteurs de recherche et leurs spiders. Regardons cela au travers de quelques exemples...

Prenons le site de Bouygues Telecom (*http://www.bouyguestelecom.fr/*), qui propose des menus déroulants tout à fait *user friendly* (voir figure 7-9).

Figure 7-9

Menu déroulant sur le site Bouygues Telecom



Le code pour réaliser ce menu est le suivant :

```
<li><a class="v2_menuLevel1" href="http://www.laboutique.bouyguestelecom.fr/
➥nos-offres.html"
onclick="xt_med('C','1','Hub::Header::Mobile','S')" xtclib="HeaderMobile">Mobile</a>
<ul>
<li><a href="http://www.laboutique.bouyguestelecom.fr/nos-telephones-mobiles.html"
➥onclick="xt_med('C','1','Hub::Header::Mobile::Telephones','S')" xtclib="HeaderMobile">
➥T&eacute;l&eacute;phones</a></li>
<li><a href="http://www.laboutique.bouyguestelecom.fr/1-forfait-formule.html"
➥onclick="xt_med('C','1','Hub::Header::Mobile::Forfaits','S')" xtclib="HeaderMobile">
➥Forfaits</a></li>
<li><a href="http://www.laboutique.bouyguestelecom.fr/recherche-telephones-cle--
➥cle+3gplus/edge.html" onclick="xt_med('C','1','Hub::Header::
➥Mobile::Cle3GEdge','S')" xtclib="HeaderMobile">Clé 3G+/EDGE</a></li>
<li><a href="http://www.laboutique.bouyguestelecom.fr/3-universal-mobile-formule.html"
onclick="xt_med('C','1','Hub::Header::Mobile::ForfaitBloque','S')" xtclib
➥="HeaderMobile"> Forfaits bloqu&eacute;s</a></li>
<li><a href="http://www.laboutique.bouyguestelecom.fr/2-carte-nomad-formule.html"
➥onclick="xt_med('C','1','Hub::Header::Mobile::Carte','S')" xtclib="HeaderMobile">
➥Cartes</a></li>
<li><a href="http://www.internetmobile.bouyguestelecom.fr/"
onclick="xt_med('C','1','Hub::Header::Mobile::EspaceInternetMobile','S')" xtclib
➥="HeaderMobile">Espace Internet Mobile</a></li>
</ul>
</li>
```

On le voit, ce code est tout à fait *spider friendly* puisque, pour chaque choix du menu, l'option `<a href="http://www.laboutique.bouyguestelecom.fr/3-universal-mobile-formule.html"...>Forfaits bloqu&eacute;s</a>` est indiquée. Une adresse URL valable est indiquée dans l'attribut `href`, donc le lien sera suivi par les robots des moteurs (les événements `onclick` ou `xtclib` ne gênent en rien ces derniers).

Regardons maintenant les menus déroulants du site Amazon.fr (*http://www.amazon.fr/*), représenté sur la figure 7-10.

Figure 7-10

Menu déroulant sur le site Amazon.fr

Le code suivant est proposé (extrait simplifié) :

```
<div class="navLeftNavTitle">
<h4 style="margin: 0px; font-size: 13px">Livres</h4></div>
<div class="leftNav"> <a href="/livre-achat-occasion-litterature-roman/b/ref
➥=sa_menu_lv0_w?ie =UTF8&node=301061&pf_rd_p=218610591&pf_rd_s=left-nav-1&p
➥f_rd_t=101&pf_rd_i=405320&pf_rd_m=A1X6FK5RDHNB96&pf_rd_r=1MKKR1HRNC3Y6F7KFTW8">
➥Tous les livres</a> </div>
<div class="leftNav"> <a href="/livres-anglais-computers-business-used/b/ref
➥=sa_menu_enlv0_w?ie=UTF8&node=52042011&pf_rd_p=218610591&pf_rd_s=left-nav-1&p
➥f_rd_t=101&pf_rd_i=405320&pf_rd_m=A1X6FK5RDHNB96&pf_rd_r=1MKKR1HRNC3Y6F7KFTW8">
➥Livres en anglais</a> </div>
<div class="leftNav"> <a href="/Nouveaut%C3%A9s-para%C3%AEtre-Livres/b/ref
➥=sa_menu_nfp0_w?ie=UTF8&node=112828011&pf_rd_p=218610591&pf_rd_s=left-nav-1&p
➥f_rd_t=101&pf_rd_i=405320&pf_rd_m=A1X6FK5RDHNB96&pf_rd_r=1MKKR1HRNC3Y6F7KFTW8">
➥Nouveautés et À paraître</a> </div>
<div class="leftNav"> <a href="/Chercher-Coeur-Livres/b/ref=sa_menu_si0_w?ie
➥=UTF8&node=306966011&pf_rd_p=218610591&pf_rd_s=left-nav-1&pf_rd_t=101&p
➥f_rd_i=405320&pf_rd_m=A1X6FK5RDHNB96&pf_rd_r=1MKKR1HRNC3Y6F7KFTW8">Cliquez pour
➥feuilleter</a> </div>
</div>
```

Le cas est identique à celui de Bouygues Telecom avec un code de lien tout à fait compatible avec les moteurs de recherche :

```
<a href="/livre-achat-occasion-litterature-roman/b/ref=sa_menu_lv0_w?ie=UTF8&node
➥=301061&pf_rd_p=218610591&pf_rd_s=left-nav-1&pf_rd_t=101&pf_rd_i=405320&pf_rd_m
➥=A1X6FK5RDHNB96&pf_rd_r=1MKKR1HRNC3Y6F7KFTW8">Tous les livres</a>
```

Si l'URL indiquée n'est pas optimisée pour les moteurs de recherche (ce qui est un autre problème), le lien, quant à lui, sera suivi sans problème puisque l'adresse de la page de destination apparaît dans l'attribut `href`.

Dernier exemple avec le système d'onglets du site Abondance.com (*http://www.abondance.com/*), représenté à la figure 7-11.

Figure 7-11

Navigation par onglets sur la page d'accueil du site Abondance.com

Dont voici le code HTML :

```
<ul id="onglet">
<li class="active"><a href="http://actu.abondance.com/"> Actualit&eacute; </a></li>
<li><a href="http://docs.abondance.com/"> Articles </a></li>
<li><a href="http://blog.abondance.com/"> Blog </a></li>
    <li class="abonnes"><a href="http://abonnes.abondance.com/"> Abonn&eacute;
    ➥s </a></li>
<li><a href="http://outils.abondance.com/"> Outils </a></li>
<li><a href="http://www.forums-abondance.com/"> Forums </a></li>
<li><a href="http://lettres.abondance.com/"> Newsletters </a></li>
<li><a href="http://livres.abondance.com/"> Etudes/livres </a></li>
<li><a href="http://emploi.abondance.com/"> Emploi </a></li>
<li><a href="http://ressources.abondance.com/"> Ressources </a></li>
<li><a href="http://www.boutique-abondance.com/"> Boutique </a></li>
</ul>
```

Vous l'avez compris, pour les mêmes raisons que précédemment, ces liens seront suivis sans aucun souci par les robots (adresse valable dans la zone `href` du lien). Il est donc tout à fait possible, au travers de balises `<div>` ou `<ul><li>`, et en gérant bien les feuilles de styles (CSS) correspondantes, de réaliser des systèmes de navigation qui ne poseront aucun problème aux spiders de Google et consorts. Bonne nouvelle. Pourquoi s'en priver ?

Dernier petit « truc » : pour savoir si vos menus sont compatibles avec les moteurs de recherche, observez-les sur la version textuelle du cache de Google (liens « En cache>Version en texte seul ») : s'ils apparaissent, c'est qu'ils sont compatibles ! Exemple en figure 7-12 pour le site Abondance.

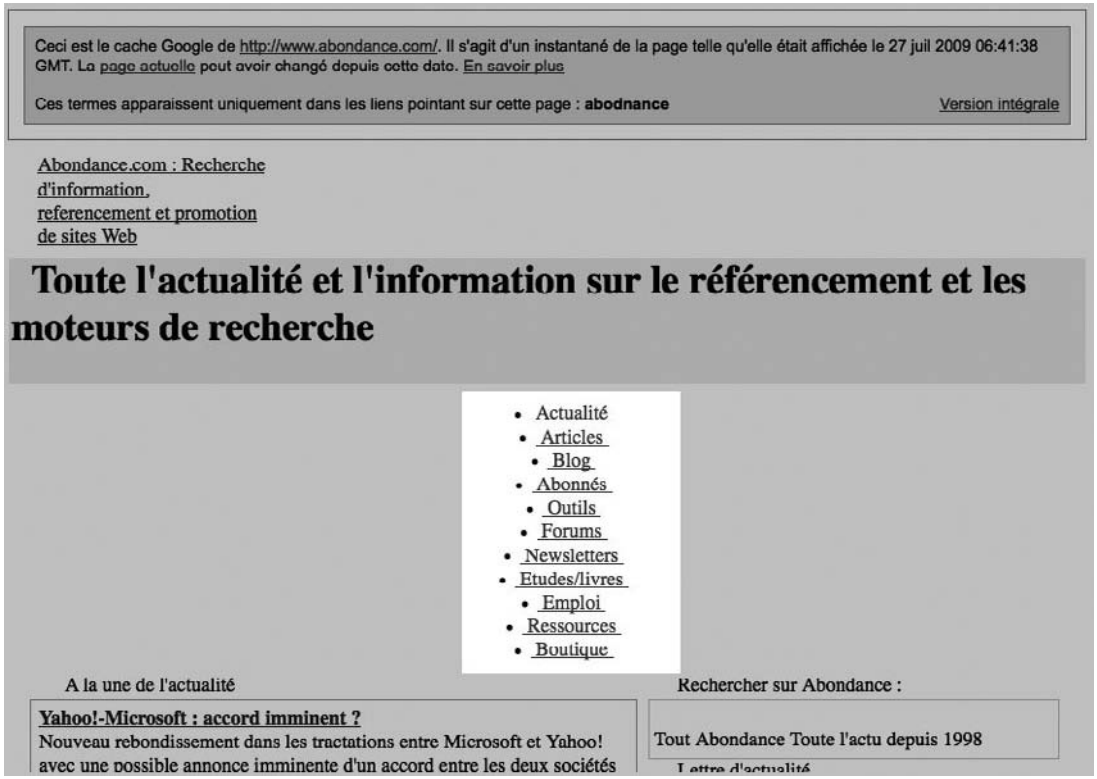


Figure 7-12

Les liens du menu apparaissent sur la version textuelle du cache : ils sont donc suivis par les spiders du moteur !

Webographie spider friendly

Voici quelques liens pour vous aider dans vos créations de menus en CSS :

– Créer des menus simples en CSS

http://www.alsacreations.com/tuto/lire/574-Creer-des-menus-simples-en-CSS.html

– De façon plus générale, cet excellent site :

http://www.alsacreations.com/tutoriels/

– CSS Menus

http://www.cssmenus.co.uk/

– Menu déroulant en JavaScript

http://astuces-webdesign.com/xhtml-html-css/menu-deroulant-en-javascript-398.html

La problématique des sites Web 2.0 et Ajax

Les sites Web en Ajax ou dans la mouvance Web 2.0 posent quelques problèmes aux moteurs de recherche pour deux raisons majeures :

- Ils contiennent une bonne dose de JavaScript et certains de leurs liens ne sont pas conçus pour être compatibles avec les moteurs de recherche.
- De nombreux contenus, notamment textuels, sont compris dans des scripts (entre les balises `<script>` et `</script>`), qui sont des zones non lues par les moteurs de recherche, des *terra incognita* pour leur spider.

Pour pallier ces inconvénients, il vous faudra éviter le langage JavaScript le plus possible, bien que Google améliore sa compréhension de ce langage. Dans certains cas (on vient de le voir), il est de toute façon possible de faire autrement. Dans certains cas, non.

Il vous faudra alors, dans la mesure du possible, « extraire » le contenu textuel des scripts afin qu'ils soient lus par le moteur. Là encore, la vision d'une page de test dans le cache textuel de Google vous donnera beaucoup d'informations... Bref, il faut que votre site reste consultable même si vous désactivez JavaScript sur votre navigateur. Si votre code Ajax propose le chargement à la volée de petits bouts de contenu insérés dans des scripts, il y a de fortes chances pour que le problème soit pour l'instant insoluble pour les moteurs de recherche...

Une question est souvent posée également au sujet des propriétés `display:none` et `visibility:hidden`. Certains webmasters se posent la question sur le fait d'employer ces fonctions pour rendre invisibles certains contenus (assez souvent utilisées, par exemple pour cacher le contenu d'onglets) et le risque que les moteurs de recherche prennent cette technique pour du spam. En fait, la situation est assez simple : si vous utilisez ces propriétés pour des raisons de charte graphique et d'affichage conditionnels, par exemple, cela ne posera aucun problème. Si vous les utilisez pour cacher du texte aux internautes mais le rendre visible aux moteurs pour un meilleur référencement, alors il y a de fortes chances pour que cela soit puni... Comme toujours, on ne fait pas le procès d'un marteau sous prétexte qu'il a servi à taper sur la tête d'une personne...

Résumons donc la situation en quelques mots : Ajax et moteurs de recherche ne font pas vraiment bon ménage pour l'instant. Autant le savoir au moment où les choix technologiques pour votre site web seront déterminés...

Ajax et moteurs de recherche

Un billet d humeur intéressant de Jérôme Charron, disponible sur le site Moteurzine, traite de la façon dont les moteurs de recherche « digèrent » les sites en Ajax :

http://www.moteurzine.com/archives/2006/moteurzine127.html#2

Nous vous invitons également à consulter les liens suivant :

– Optimisation du référencement d'un site en Ajax

http://www.webrankinfo.com/actualites/200711-referencement-site-ajax.htm

– A Spider's View of Web 2.0 (par Google)

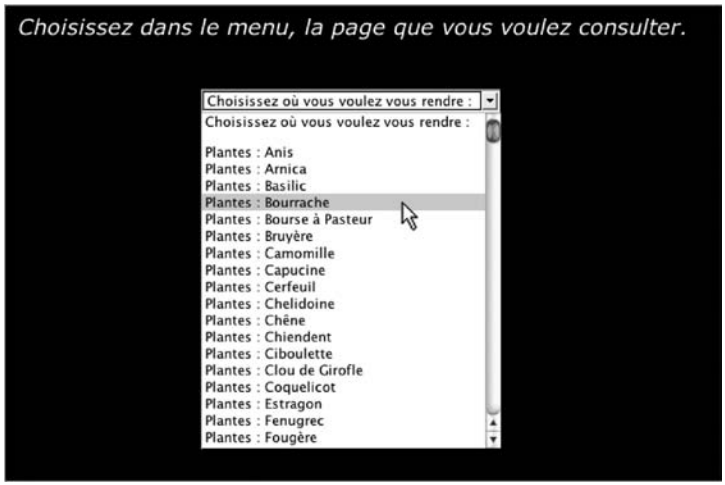
http://googlewebmastercentral.blogspot.com/2007/11/spiders-view-of-web-20.html

Menus déroulants et formulaires

Si vous proposez sur votre site une navigation basée sur les menus déroulants, comme sur la figure 7-13, sachez que si les textes compris dans ces menus sont bien pris en compte par les moteurs comme du texte visible, les robots ne suivent pas ce type de lien. Il vous faudra donc les doubler au travers de liens textuels ou compatibles avec les robots, sur la page elle-même ou sur le plan du site (voir plus loin pour ce dernier cas).

Figure 7-13

Une navigation par menus déroulants ne sera pas comprise des robots…



Idem si vos « fiches produits » sont accessibles au travers d'un moteur de recherche, comme dans l'exemple de la figure 7-14.

Figure 7-14

Les spiders des moteurs ne savent pas remplir des champs de formulaires pour trouver des pages web.

Les robots des moteurs ne savent pas cliquer dans des « boîtes à cocher », entrer des données dans un formulaire, choisir dans des menus déroulants et valider une recherche. Ils ne savent que suivre des liens (même si Google travaille sur le sujet : *http://actu.abondance.com/2008/04/les-robots-de-google-veulent-explorer.html*)… Donc si seul ce type d'accès est disponible, vos produits ne seront jamais indexés par Google et consorts. Là encore, il faudra fournir aux robots, toujours de façon visible et loyale (oubliez toute velléité de cacher ce type de lien dans vos pages HTML, d'une façon ou d'une autre) des liens compatibles avec les robots. Dans ce cas, seule la page « Plan du site » (voir la fin de ce chapitre) ou des options de type Sitemaps (voir chapitre 8) seront envisageables.

Sites dynamiques et URL « exotiques »

Un site dynamique est ainsi appelé par opposition à un site statique. Le site statique gère des pages créées au préalable. Ces pages HTML sont dites « statiques » et sont affichées telles quelles dès qu'un internaute les demande. Les pages sont donc créées à l'aide d'un éditeur HTML, puis stockées pour être affichées sous leur forme initiale.

Le site dynamique, pour sa part, puise ses informations dans une base de données (qui peut être d'origines diverses) et crée des pages à la volée, en fonction d'une action ou d'un événement. Par exemple, pour une saisie effectuée par un internaute, le moteur de recherche est l'exemple type de site dynamique.

En effet, un internaute, lorsqu'il arrive sur un moteur, saisit une requête dans un formulaire et l'outil, sur la base des mots-clés demandés, va créer une page de résultats « sur mesure ». Bien entendu, cette page n'existe pas en tant que telle sur le disque dur du moteur, elle est donc créée à la volée. Un moteur de recherche est donc un site dynamique.

Il en sera de même avec des sites web d'e-commerce, par exemple dans le cadre d'un catalogue en ligne, mais également la consultation d'archives de presse, etc.

Ce qui bloque le plus souvent les moteurs de recherche est représenté par l'URL des pages, qui contient, pour ce type de site, deux caractères spécifiques et représentatifs des sites dynamiques : le point d'interrogation (`?`) et l'esperluette (`&`).

Si votre site est dynamique, c'est-à-dire construit sur une base de données consultée à la volée, cela peut poser des soucis aux moteurs de recherche en fonction de la structure des URL de vos pages. Exemple :

http://www.sitedynamique.com/v2/V2_liste_produit.asp?id1=2155&id2=2159&id3=2177&inftyp=1&niv=3&marque=no&num=3

Cette URL contient un caractère `?`, synonyme de « passage de paramètres d'une page à l'autre » et des esperluettes (`&`) qui séparent les différents paramètres entre eux. Ici, ils ont pour nom : `id1`, `id2`, `inftyp`, `niv`, `marque`, `num`, etc.

La situation aujourd'hui est assez claire : si les URL de vos pages contiennent jusqu'à trois paramètres (soit deux esperluettes), votre site ne doit pas poser de problèmes aux moteurs de recherche (hors souci spécifique comme les cookies ou les identifiants de

session - voir plus loin). Si les URL de votre site contiennent quatre paramètres (soit trois esperluettes) et plus, il aura du mal à être indexé, et ce, quel que soit le moteur. Son référencement deviendra plus aléatoire. Nous allons développer ce point dans ce chapitre.

Un décalage entre les technologies de création de site et leur prise en compte par les moteurs

Il existe, et cela est vrai depuis que les moteurs de recherche existent, un certain décalage entre le moment où les techniques de création de sites web sont utilisées et la façon dont les moteurs de recherche les indexent.

Cela s'est vérifié pour les frames (souvenez-vous d'Excite qui ignorait totalement les sites ainsi réalisés), puis pour le Flash ou le JavaScript, par exemple. Cela se vérifie encore avec les sites web dynamiques, qui ont longtemps représenté un obstacle rédhibitoire pour les moteurs. La situation semble s'améliorer aujourd'hui, mais elle n'est pas encore parfaite, loin de là.

Format d'une URL de site dynamique

L'URL d'une page émanant d'un site dynamique est le plus souvent affichée sous une forme du type :

http://www.sitedynamique.com/prog.cgi?kw=motcle&langue=fr&zone=france&encodage=ISO-8859-1

Cette adresse peut s'interpréter ainsi : « sur le site *www.sitedynamique.com*, on a lancé le programme nommé `prog.cgi` en lui passant comme paramètres les variables `kw` (de valeur `motcle`), `langue` (de valeur `fr`), `zone` (de valeur `france`) et `encodage` (de valeur `ISO-8859-1`).

Il en est exactement de même sur Google. Si vous allez sur le site *http://www.google.fr/* et que vous tapez le mot-clé « abondance », l'URL de la page de résultats aura comme intitulé :

http://www.google.fr/search?q=abondance&ie=ISO-8859-1&hl=fr&btnG=Recherche+Google&meta=

Sur Google, c'est le programme nommé `search` qui a été lancé, avec pour paramètres :

- `q` = `abondance` (le mot-clé) ;
- `ie` = `ISO-8859-1` (l'encodage des caractères) ;
- `hl` = `fr` (la zone linguistique) ;
- `btnG` = `Recherche Google` (le nom du bouton de validation de Google) ;
- `meta` = (autre information – vide dans ce cas – pour le moteur).

Les méthodes GET et POST

Google utilise pour son formulaire de recherche la méthode `GET` (passage de paramètres dans l'URL) contrairement à un moteur comme celui de Free, par exemple, qui utilise, sur sa page d'accueil, la méthode `POST`. Dans ce cas, la page de résultats a une URL identique quel que soit le mot-clé recherché (*http://search.free.fr/google.pl*). La méthode `POST` est rédhibitoire pour l'indexation des pages dynamiques puisqu'une seule URL est proposée aux robots pour chaque page. L'adresse des documents n'est donc plus différenciatrice de leur contenu.

Voici quelques exemples (réels) d'URL dynamiques :

- *http://www.nova-cinema.com/main.php?page=search.en.htm*
- *http://canadapost.internic.ca/search.asp?lang=fr*
- *http://www.medbioworld.com/MedBioWorld/TopicLinks.aspx?type=Reference%20Tools&&category=(All)&&concept=Medicine*

Le plus souvent, les sites dynamiques sont créés sur la base de technologies de programmation comme PHP, ASP ou CFM. Mais ils peuvent également être conçus grâce à des produits propriétaires (qui poseront plus ou moins de problèmes supplémentaires) comme Lotus Notes, Vignette, BroadVision, etc.

Pourquoi les moteurs de recherche n'indexent-ils pas – ou mal – les sites dynamiques ?

Le fait que les URL dynamiques aient un format spécifique ne nous explique pas pourquoi elles sont refusées ou mal comprises par les moteurs de recherche. Il y a en fait plusieurs explications à cela :

- Le nombre de pages créées à la volée par un site dynamique peut être quasi infini. En effet, prenez un catalogue du type de ceux d'Amazon ou de la Redoute, multipliez le nombre d'articles par le nombre d'options possibles (délai d'envoi, couleur, taille des vêtements, etc.) et vous obtenez rapidement, pour un seul site, plusieurs centaines de milliers, voire millions de pages web potentielles présentant chaque produit de façon unique. Difficile, pour un moteur, de les indexer toutes ou, en cas contraire, de savoir où s'arrêter.
- Un site web dynamique a la possibilité de créer, en quelques secondes, des milliers de pages à la volée. Il s'agit également là d'un système à haut risque pour ce qui concerne le spam contre les moteurs. Dans ce cas, ces derniers se méfient et, parfois, optent pour l'option la moins risquée. Ils préfèrent ne prendre en compte aucune page plutôt que de courir le risque de devenir un réservoir à spam au travers de techniques de création incessante de pages un peu trop optimisées.
- Une même page, proposant le même contenu, peut être accessible à l'aide de deux URL différentes (ce problème est notamment crucial en ce qui concerne les identifiants de session comme nous le verrons plus loin). Cela risque d'être problématique pour un moteur, qui devra alors mettre en place des procédures de dédoublonnage (duplicate content, voir plus loin dans ce chapitre) qui peuvent s'avérer complexes.
- La longueur excessive de certaines URL, passant de nombreux paramètres, peut également poser des problèmes aux moteurs. Par ailleurs, certains caractères apparaissant dans ces adresses (`#`, `{`, `[`, `|`, `@`, etc.) peuvent également être bloquants parfois, tout comme les lettres accentuées, peu fréquentes dans les URL statiques, qui peuvent causer des soucis de codage.
- Certains problèmes posés par les sites web dynamiques sont appelés *spider traps* : il s'agit de pages mal reconnues par les spiders des moteurs, qui s'y perdent parfois dans des boucles infinies et indexent alors des milliers de documents différents représentatifs de quelques pages web uniquement.

Quels formats sont rédhibitoires ?

Comment un moteur de recherche réagit-il face à une page dynamique ? Il y a de cela quelques mois, voire quelques années, elles étaient purement et simplement ignorées. Pour certains moteurs, les pages en PHP, ASP ou CFM étaient bannies, quelle que soit leur forme. Heureusement, cette période est aujourd'hui révolue. Le simple fait d'avoir été créée dans l'un de ces langages de programmation n'est plus rédhibitoire.

En effet, à l'heure actuelle, les moteurs de recherche reconnaissent de façon bien plus optimale les pages dynamiques. Mais la situation n'est pas encore idéale et certains blocages sont encore présents. Globalement, il en existe deux très importants : le nombre de paramètres passés dans l'URL et l'identifiant de session (que nous étudierons plus loin dans ce chapitre).

Dans un premier temps, il semblerait que les URL contenant jusqu'à trois paramètres ne posent pas – ou plus – de problèmes aux moteurs. Exemples d'adresses aujourd'hui acceptées par ces derniers :

- *http://www.sitedynamique.com/search.cgi?kw=motcle*
- *http://www.sitedynamique.com/search.cgi?kw=motcle&langue=fr*
- *http://www.sitedynamique.com/search.cgi?kw=mc&langue=fr&zone=france*

Ce fait est avéré sur des moteurs comme Google et Yahoo!, par exemple.

En revanche, jusqu'en 2009, ce type d'URL (passage de plus de trois paramètres dans l'adresse) était refusé ou tout du moins pris en compte de façon aléatoire :

http://www.sitedynamique.com/search.cgi?kw=mc&langue=fr&zone=france&codage=ISO-8859-1

Il semblerait, cependant, que la situation s'améliore de ce côté. On voit de plus en plus de pages possédant quatre paramètres, voire plus, dans leur URL et néanmoins présentes dans les index respectifs de Google et de Yahoo!. Cependant, même si cette situation est meilleure aujourd'hui, elle reste encore bloquante dans de nombreux cas. Il vous faudra donc en tenir compte lors la mise en place de votre site afin de passer le moins de paramètres possible dans vos adresses. Allez au strict minimum. Pour l'instant, on peut encore estimer que le chiffre de trois paramètres est un maximum. Au-delà, il vous faudra envisager une solution technique adéquate comme la réécriture d'URL que nous étudierons très bientôt…

De plus, la présence de mots-clés dans les URL est un critère important pour les moteurs, ce qui est rarement le cas dans une URL dynamique par défaut. Dans tous les cas, vous devrez donc certainement passer par ces solutions de réécriture de vos URL dynamiques pour optimiser votre site.

Yahoo! pense aux sites dynamiques

À noter qu'en août 2007, Yahoo! a rajouté sur son site Site Explorer, destiné aux webmasters (*http://siteexplorer.search.yahoo.com/*), une fonction visant à mieux référencer les sites dynamiques. Plus d'informations ici : *http://actu.abondance.com/2007/08/yahoo-facilite-le-rfrencement-des-sites.html.*

Les pages satellites

Il s'agit là de la solution la plus souvent proposée sur le marché jusqu'à fin 2005 par les sociétés de référencement. Elle est simple dans sa conception et revient à dire : « Votre site est dynamique et peut présenter des blocages pour les moteurs ? Pas de problème, nous allons créer de nouvelles pages, spécialement construites pour le référencement et ce sont sur celles-ci que nous allons travailler. »

Avantages de la page satellite :

- Le site web en lui-même n'est pas modifié. Le client garde donc la totale maîtrise de son contenu et de ses pages web et sous-traite totalement l'aspect référencement. Les pages satellites peuvent même être hébergées chez le référenceur, comme cela arrive parfois.
- La page satellite n'étant pas vue par l'internaute, le référenceur peut réellement travailler de façon pointue sur son contenu et son optimisation.

Inconvénients de la page satellite :

- Son contenu appartient parfois au référenceur, notamment s'il héberge lui-même les pages satellites de son client. Certains contrats peuvent ne pas être très clairs à ce sujet. Si le client change de référenceur, il lui faudra peut-être recommencer le travail à zéro. Si vous désirez passer par ce type de prestation, assurez-vous bien que les pages satellites développées pour votre société restent votre propriété en fin de contrat.
- Il est toujours plus difficile d'obtenir un bon PageRank (la popularité selon Google) sur une page satellite que sur une page réelle du site. En effet, il est rare que les sites qui pointent vers vous aient établi des liens vers vos pages satellites. Même si un maillage complexe et élaboré de différentes pages satellites entre elles permet de pallier le plus souvent ce problème, on ne sait pas vraiment jusqu'à quand cela va fonctionner.
- Les moteurs de recherche indiquent clairement dans leurs conditions générales d'utilisation qu'ils sont contre ces techniques considérées comme de la fraude. Cela règle le problème.

Nous en avons déjà parlé dans cet ouvrage : la page satellite est une technique devenue obsolète et à abandonner. Passons donc à la suivante…

Le cloaking

La technique du cloaking (ou *IP delivery*) est finalement assez proche de celle de la page satellite et nous en avons déjà parlé à plusieurs reprises dans cet ouvrage. Imaginons que votre page d'accueil s'intitule `index.html`. Vous allez créer, dans un premier temps, une copie de cette page. La première sera la page originelle du site, la seconde sera optimisée pour les moteurs.

Les systèmes de cloaking sont ensuite installés sur le serveur, sous forme de logiciels spécifiques, et tentent d'identifier qui arrive sur vos pages :

- soit c'est le robot d'un moteur (reconnaissable par son user-agent et/ou son adresse IP, dont des bases de données sont aisément identifiables sur le Web) et, dans ce cas, le système lui fournit la page optimisée (l'équivalent de la page satellite) ;
- soit c'est un internaute et, dans ce cas, le serveur lui envoie la page « normale ».

L'analogie avec la page satellite n'est pas fausse puisque, finalement, seule la fonction de redirection est différente. Dans un cas, elle est mise en place au moment du clic (page satellite), dans l'autre cas, au moment de la visite du robot.

Que penser de cette technique ? Un peu la même chose qu'avec les pages satellites. Elle constituerait, après tout, une bonne façon de pallier les problèmes techniques posés par les sites web dynamiques. Sauf que… la plupart des gros moteurs de recherche, dont Google, ont indiqué par le passé qu'ils refusaient ce type de solution, considérée comme du spam. Google est très clair à ce sujet sur son site (*http://www.google.com/support/webmasters/bin/answer.py?hl=fr&answer=66355*) :

« Le cloaking est la pratique qui consiste à présenter aux utilisateurs des URL ou un contenu différents de ceux destinés aux moteurs de recherche. En raison de la présentation de résultats différents selon le user-agent, votre site peut être considéré comme trompeur et être retiré de l'index Google. »

Avantages du cloaking :

- peu complexe à mettre en œuvre, la plupart des informations et des outils sont disponibles en ligne ;
- les mêmes avantages que pour la page satellite.

Inconvénients du cloaking :

- les gros moteurs de recherche n'apprécient que modérément ce type de pratique. Risque de liste noire assez important ;
- les mêmes inconvénients que pour la page satellite.

Il sera donc difficile de prendre en compte le cloaking dans le cadre du référencement d'un site web dynamique.

Cependant, il semblerait que les moteurs de recherche nuancent, petit à petit, leur opinion sur le cloaking. En effet, on a vu Google, dans le passé, mettre en liste noire un site ayant utilisé ce type de pratique (*http://actu.abondance.com/2004-21/whenu.html*) mais également accepter officiellement qu'un autre le fasse (*http://searchenginewatch.com/sereport/article.php/3360681*). Le discours n'est donc pas totalement radical. En tout état de cause, l'idéal serait de contacter les moteurs de recherche pour leur exposer votre cas si vous désirez utiliser du cloaking dans le cadre de votre référencement (en espérant qu'ils vous répondent...). Sinon, le risque reste aujourd'hui élevé…

La recopie de site web

Autre solution parfois proposée par les sociétés de référencement, notamment les plus importantes : la recopie pure et simple du site web de leur client sur leurs serveurs afin de le rendre statique (proposant des URL compatibles avec les contraintes des moteurs).

Le principe est simple : le référenceur effectue une copie conforme, un clone de votre site sur son serveur. Ce sont alors ces pages (dont les URL auront été rendues « spider compatibles ») et le contenu optimisé, qui seront prises en compte pour le référencement. Par ce biais, le référenceur rend statique un site dynamique.

Avantages de la recopie de site web :

- toutes les pages peuvent ainsi être rendues statiques et être alors optimisées pour le référencement, puisque le référenceur ne travaille que sur un clone du site réel.

Inconvénients de la recopie de site web :

- la mise à jour des pages doit être prise en compte (par exemple, les prix des articles ou leur description ou les nouveaux articles). Si le site est mis à jour quotidiennement, la « moulinette » peut vite ressembler à une usine à gaz ;
- la solution peut s'avérer complexe à mettre en œuvre et donc onéreuse. Cette solution sera peut-être réservée aux sites importants, disposants de budgets de référencement conséquents ;
- la copie du site se trouve parfois sur le site du référenceur. Si vous changez de prestataire, vous devrez reprendre le référencement depuis le début.

Conclusion : une solution qui peut s'avérer intéressante mais qui n'est pas à la portée de toutes les bourses.

Création de pages de contenu

Autre solution proposée par les sociétés de référencement : la création de pages web, sans redirection, hébergées sur votre site ou sur le serveur du référenceur, proposant un contenu spécifique et, bien sûr, une URL « propre » pour les moteurs.

L'idée, là encore, est simple : il s'agit de créer des pages, disposant d'une adresse statique, et qui, à l'instar des pages satellites, serviront en priorité au référencement de votre site. Mais, là où les pages satellites redirigent l'internaute vers une page réelle du site, la page de contenu propose un contenu réel, basé sur celui de votre site originel et respectant donc votre charte graphique.

Le référenceur travaille avec le client et crée des pages en piochant du contenu sur le site web de départ, contenu qui lui sert à proposer du code HTML optimisé (le premier paragraphe, par exemple, sera plus particulièrement mis en valeur avec les mots importants en gras, de grande taille et cliquables, etc.). Cette page proposera également, comme toutes les pages du site, une barre de navigation permettant au visiteur de cliquer sur les autres rubriques.

Avantages des pages de contenu :

- bonne pérennité de la solution puisqu'il s'agit là de mettre à la disposition de l'internaute un contenu de qualité, adapté à la requête demandée. Difficile de penser que les moteurs voient ce type de possibilité d'un mauvais œil à l'avenir…

Inconvénients des pages de contenu :

- cela demande un vrai travail éditorial dont la mise en œuvre peut s'avérer longue et onéreuse ;
- les options techniques retenues par le client pour la charte graphique de son site peuvent, dans certains cas, être un obstacle au référencement (Flash, menus en JavaScript) et augmenter le travail d'optimisation ;
- le client doit pouvoir contrôler du mieux possible le contenu mis en ligne par le référenceur.

La création de pages de contenu est une solution qui se développe au fur et à mesure de l'abandon des pages satellites, et qui permet d'obtenir très souvent de bons résultats pour un rapport qualité/prix intéressant. Elle demande cependant un travail important et un contrôle strict de la part du client sur les informations mises en ligne par le référenceur.

Optimisation des pages non dynamiques

Une solution que l'on oublie souvent consiste à optimiser uniquement les pages web d'un site ne représentant pas d'obstacle pour les moteurs : page d'accueil, plan du site, présentation de la société, etc. Toutes ces pages n'ont pas, *a priori*, besoin de passage de paramètres dans leurs adresses et il devrait être possible de faire en sorte qu'elles soient disponibles grâce à des URL propres (sans `?`, ni `&`, ni identifiant de session, etc.). Si ce n'est pas le cas par défaut, peut-être cela vaut-il la peine de faire en sorte de simplifier les adresses de ces documents, ce qui les rendra immédiatement indexables par les moteurs.

Peut-être que, pour certains sites, vous pouvez faire l'impasse sur l'indexation des pages web purement dynamiques, c'est-à-dire affichant des informations issues d'une base de données, et de ne prendre en compte pour le référencement que des pages statiques du site. C'est en tout cas une solution très économique en temps… et en argent, même si elle n'est clairement pas la plus optimisée !

Avantages de l'optimisation des pages statiques :

- peut être rapide et peu onéreuse.

Inconvénients de l'optimisation des pages statiques :

- il n'est pas toujours techniquement possible de rendre statiques les pages connexes d'un site ;
- cette solution ne permet pas de référencer les pages utilisant des informations issues d'une base de données, comme un catalogue produits, qui sont souvent les plus importantes et celles à qui on veut donner la meilleure visibilité.

Offres de référencement payant et de liens sponsorisés

Dernière solution : prendre en compte les offres XML (autrement appelées « référencement payant », « trusted feed », « paid inclusion », « feed XML »…) de moteurs comme Voila/Wanadoo et Yahoo!, qui vous permettent, contre rémunération, d'indexer facilement vos pages web dynamiques.

Vous disposez alors d'une garantie d'indexation et de rafraîchissement (toutes les 48h le plus souvent), mais pas de garantie de positionnement (même si cela reste à démontrer). Ask Jeeves, qui proposait une solution de ce type, l'a abandonnée. Google, pour sa part, n'en propose pas et clame haut et fort qu'il ne mettra jamais ce type de solution à son catalogue.

- Voila/Wanadoo proposait jusqu'en 2006 une offre baptisée « URL Express » qui permettait de soumettre des pages pour des tarifs commençant à 15 € par page. Elle ne semblait plus disponible en 2009, de façon officielle tout du moins.
- Pour ce qui est de Yahoo!/Overture, l'offre ne semble pour l'instant disponible qu'aux États-Unis (*http://search.yahoo.com/info/submit.html*). Il vous faudra donc contacter directement Yahoo! outre-Atlantique ou passer par une société française ayant passé un contrat avec Yahoo! US.

Nous reparlerons – rapidement – de ces offres au chapitre 8.

Avantages de l'offre de référencement payant :

- rapidité ;
- indexation d'à peu près n'importe quel type de page.

Inconvénients de l'offre de référencement payant :

- payant ;
- peut s'avérer rapidement onéreux en fonction du nombre de pages à indexer ;
- les modèles économiques proposés (CPM, coût pour mille affichages – on paye à chaque fois qu'une page est affichée, ou CPC, coût par clic – on ne paye que lorsque les pages sont cliquées) sont différents d'une offre à l'autre et ne permettent pas toujours de définir un budget précis sur une durée donnée ;
- une fois que vous ne payez plus, vous n'avez aucune garantie que les pages en question restent dans l'index.

Enfin, il vous restera toujours la solution des liens sponsorisés Microsoft adCenter, Google Adwords et Yahoo! Search Marketing, entre autres, pour mener à bien vos campagnes publicitaires.

L'URL Rewriting

Il s'agit ici de la solution qui est certainement la plus efficace pour un site dynamique. Le but est de définir des règles de réécriture pour des adresses de pages web dynamiques puisque ce sont elles qui, dans la plupart des cas, bloquent les moteurs.

Par exemple, des URL du type :

http://www.sitedynamique.com/prog.php?kw=motcle&langue=fr&zone=france&encodage=ISO-8859-1

pourront être réécrites en :

http://www.sitedynamique.com/prog.php/motcle/fr/france/ISO-8859-1

Cette adresse ne pose plus de problème au moteur de recherche. Le tour est joué et le site web dynamique devient indexable sans souci par les robots.

Ces URL facilitent l'indexation des sites dynamiques, et donc leur référencement dans les moteurs.

En plus de cet avantage indéniable, la réécriture d'URL permet également de renforcer la sécurité du site en masquant les noms des variables passées dans l'URL. Si l'extension des URL propres est neutre (par exemple, `.html` ou `.htm`), il est même possible de masquer le langage utilisé sur le serveur (PHP dans notre exemple).

Mettre en place une politique d'URL Rewriting ne pose pas de difficultés majeures, même si cela peut prendre du temps et demande surtout beaucoup de rigueur organisationnelle. Mais, une fois que cela est fait, vous n'avez plus à vous en occuper, sauf modification majeure dans la structure de votre base de données.

Avantages de l'URL Rewriting :

- permet de rendre compatibles avec les moteurs de recherche bon nombre de sites web dynamiques ;
- règles de réécriture d'adresses établies une seule et unique fois ;
- il est pris en compte par tous les moteurs de recherche.

Inconvénients de l'URL Rewriting :

- la mise en place des règles de réécriture peut parfois être assez longue et fastidieuse ;
- certaines configurations techniques (serveurs/solutions propriétaires) ne proposent pas de telle solution.

Principe de l'URL Rewriting

Le principe de la réécriture d'URL est donc de mettre en place un système sur le serveur pour qu'il sache interpréter ce nouveau format d'URL. Par exemple, quand un visiteur accède à la page *http://www.notre-site.com/articles/article-12-2-5.html*, le serveur doit renvoyer exactement la même chose que si le visiteur avait demandé à accéder à la page *http://www.notre-site.com/articles/article.php?id=12&page=2&rubrique=5*.

La correspondance entre les deux schémas d'URL est alors décrite sous forme de règles de réécriture. Chaque règle permet de décrire un format d'URL. Dans l'exemple ci-dessus, la règle de réécriture va indiquer au serveur de prendre le premier nombre comme numéro d'article, le deuxième comme numéro de page et le troisième comme numéro de rubrique.

La technique de réécriture d'URL la plus connue est celle disponible sur les serveurs Apache, le plus souvent utilisés avec le langage PHP. Sauf mention spéciale, tous les exemples de ce chapitre seront donc consacrés au langage PHP et au serveur Apache.

Mise en place de l'URL Rewriting

Si vous avez déjà un site dynamique en ligne, voici les étapes à suivre pour mettre en place la réécriture d'URL :

1. Vérifier que votre hébergeur permet l'utilisation de l'URL Rewriting.
2. Identifier les pages dynamiques dont l'URL comporte des paramètres, et choisir un nouveau schéma d'URL propre.
3. Écrire les règles de réécriture dans le fichier `.htaccess` adéquat.
4. Changer tous les liens vers chaque fichier dont l'URL a changé.
5. Mettre à jour votre site et vérifier que tout fonctionne.

Examinons en détail chacune de ces étapes.

Vérification de la comptabilité de l'URL Rewriting avec votre hébergeur

La première chose à faire est bien évidemment de s'assurer que le serveur qui héberge votre site permet d'utiliser la réécriture d'URL. Tout dépend, dans un premier temps, du type de serveur utilisé. L'objet de ce chapitre n'étant pas de passer en revue tous les types de serveurs, voici un résumé des possibilités de réécriture d'URL sur les serveurs web les plus courants.

Tableau 7-1 Différentes possibilités d'URL Rewriting sur les serveurs principaux

Serveur web	Support de la réécriture d'URL	Détails

Apache	Géré par le module `mod_rewrite`, module standard d'Apache.	Le module `mod_rewrite` doit être actif. Le fichier de configuration d'Apache (`httpd.conf`) doit contenir cette ligne :
		`LoadModule rewrite_module libexec/mod_rewrite.so`
		ainsi que celle-ci :
		`AddModule mod_rewrite.c`

IIS (Microsoft)	En ASP : réécriture possible par des filtres ISAPI, commercialisés par diverses sociétés (payants).	Le paramétrage des règles de réécriture est spécifique à chaque composant.
IIS (Microsoft)	En ASPX (.NET), sur tous les serveurs supportés : des fonctions sont disponibles comme `RewriteURL()`, etc., qui prennent en charge la réécriture d'URL.	Des codes prêts à être compilés pour exploiter ces capacités sont fournis par Microsoft : *http://msdn2.microsoft.com/en-us/library/ms972974.aspx* ou *via* des projets Open Source comme : http://www.codeproject.com/KB/aspnet/urlrewriter.aspx Aucune méthode standard n'a été prévue pour définir les règles de réécriture mais une utilisation pratique consiste à les paramétrer directement dans le `web.config` (fichier de configuration de l'application ASP.Net, présent notamment à la racine du site), qui est standardisé en XML.

Si votre site est hébergé sur un serveur dédié, vous avez vous-même accès à la configuration du serveur. Dans le cas d'un serveur Apache, vous pouvez donc modifier le fichier de configuration afin d'activer le support de la réécriture d'URL. Pensez à redémarrer Apache après avoir modifié le fichier de configuration.

Mais ce n'est pas tout. Si votre site est hébergé sur un serveur mutualisé, il n'est pas certain que votre hébergeur ait activé le support de la réécriture d'URL, principalement pour des raisons de sécurité.

Enfin, si votre site est fourni par un hébergeur gratuit, il y a peu de chances pour que la réécriture d'URL soit possible. Nous vous conseillons fortement d'investir dans un hébergement payant (en plus d'un nom de domaine), les avantages sont réellement nombreux pour effectuer un bon référencement.

Définition des schémas d'URL

Reprenons notre exemple d'un site qui dispose d'une base de données d'articles, et dressons la liste des types d'URL. En voici quelques exemples :

- *article.php?id=12&rubrique=5*
- *article.php?id=12&page=2&rubrique=5*

Pour simplifier la lecture, nous n'avons listé ici que des URL concernant le même article, mais dans la pratique, quand vous essayez de dresser la liste des types d'URL sur votre site, vous pouvez tomber, par exemple, sur :

- *article.php?id=182&rubrique=15*
- *article.php?id=36&page=5&rubrique=3*

Le principe de l'URL Rewriting consiste à trouver les schémas des URL à partir de leurs formes communes. Dans notre exemple, les articles sont accessibles selon deux types d'URL (id+rubrique ou id+rubrique+page), suivant que le numéro de page est précisé ou non.

À partir du moment où vous avez identifié ces « schémas d'URL », vous devez choisir un nouveau format d'URL (l'URL propre). En général, on fait apparaître un nom de fichier avec l'extension `.html` (ou `.htm`) mais sachez que vous pouvez mettre ce que vous voulez, cela n'a aucune incidence sur la prise en compte des pages par les moteurs. En effet, quelle que soit l'extension que vous aurez choisie, la page restera une page respectant la norme HTML.

Le nom du fichier sera formé d'un préfixe et/ou d'un suffixe, et des valeurs des variables (que ce soient des chiffres ou des lettres). Profitez de cette étape pour bien réfléchir en fonction du référencement, car vous pouvez utiliser des mots-clés dans les URL de vos pages, qui soient plus parlants pour les internautes et sans doute pris en compte par les moteurs de recherche.

Voici les nouveaux formats d'URL que nous avons choisis pour chacune des URL des exemples précédents :

- *article-12-5.html*

- *article-12-2-5.html*
- *article-182-15.html*
- *article-36-5-3.html*

Pour séparer les différentes parties de l'URL, vous devez choisir un séparateur. Ici, nous avons choisi d'utiliser uniquement des tirets : il est plus efficace pour le référencement de choisir un caractère qui soit considéré comme un séparateur de mots par les moteurs de recherche.

Vous pouvez néanmoins également utiliser les caractères suivants :

- le tiret (`-`) ;
- la virgule (`,`) ;
- le point (`.`) ;
- la barre oblique, ou *slash* en anglais (`/`) ;
- la barre verticale, ou *pipe* en anglais (`|`).

Nous vous déconseillons d'utiliser les caractères suivants :

- le tiret bas, ou *underscore* en anglais (_) ;
- le signe dièse (#) ;
- l'esperluette (`&`) ;
- l'arobase (`@`) ;
- le point d'interrogation (`?`) ;
- le signe dollar (`$`) ;
- les caractères accentués ;
- l'espace.

Le tiret et la virgule sont les plus simples ; la barre oblique peut poser des problèmes de répertoires et la barre verticale n'est pas très connue des internautes. Nous avons donc défini deux formats d'URL pour notre rubrique d'affichage des articles. Essayons de les formaliser en supprimant les numéros d'articles, de rubriques ou de pages, et en les remplaçant par leur signification :

- *article-ARTICLE-RUBRIQUE.html*
- *article-ARTICLE-PAGE-RUBRIQUE.html*

Bien entendu, `ARTICLE`, `RUBRIQUE` et `PAGE` représentent ici des numéros.

Rédaction des règles de réécriture

Maintenant que nous avons déterminé les différents schémas d'URL, il reste à écrire les règles de réécriture qui vont indiquer au serveur comment interpréter chacun de ces schémas.

Passons directement à la solution que nous allons commenter. Voici le contenu du fichier `.htaccess` situé dans notre répertoire *http://www.notre-site.com/articles/* :

```
#-----------------------------------------------------------
# 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]
```

Note : il ne doit pas y avoir de retour chariot sur une ligne de règle de réécriture.

Les lignes commençant par le signe dièse (#) sont des commentaires. N'hésitez pas à en ajouter pour rendre vos fichiers plus compréhensibles : ces lignes sont totalement ignorées par le module de réécriture d'URL.

Chaque fichier `.htaccess` est spécifique à un répertoire ; nous avons pris l'habitude d'indiquer en haut de ce fichier l'emplacement du répertoire sur le site. Chaque répertoire de votre site devra donc proposer son propre fichier `.htaccess`.

Les deux premières instructions (`Options +FollowSymlinks` et `RewriteEngine on`) ne doivent être présentes qu'une seule fois par fichier, avant toute règle de réécriture.

- L'instruction `Options +FollowSymlinks` est facultative mais peut servir dans certaines configurations.
- L'instruction `RewriteEngine on` indique que nous souhaitons utiliser le module de réécriture d'URL. Si vous avez un problème avec une règle de réécriture que vous venez d'ajouter, vous pouvez désactiver en quelques secondes la réécriture d'URL le temps de comprendre le problème : il vous suffit d'écrire `RewriteEngine off` à la place de `RewriteEngine on`.

La suite du fichier est constituée d'une série de règles de réécriture. Chaque règle est écrite sur une seule ligne (sauf règles complexes) et respecte le format suivant :

```
RewriteRule URL_A_REECRIRE URL_REECRITE
```

- `RewriteRule` est un mot-clé spécifique au module `mod_rewrite` qui indique que la ligne définit une règle de réécriture ;
- ensuite vient l'URL à réécrire, c'est-à-dire l'URL propre sans existence physique sur le serveur ;

- enfin vient l'URL réécrite, c'est-à-dire l'URL telle qu'elle sera appelée en interne sur le serveur.

Le format de l'URL à réécrire est basé sur les expressions régulières, dont la base devra être acquise pour pouvoir définir des règles de réécriture. Ne vous inquiétez pas, dans la plupart des cas c'est très simple.

Cet exemple de règle de réécriture permet déjà de gérer notre rubrique d'articles, mais il existe d'autres règles plus complexes que nous n'étudierons pas ici.

Documentation officielle sur l'URL Rewriting

Vous trouverez de nombreuses informations sur l'URL Rewriting en consultant le site officiel d'Apache à l'adresse suivante : *http://httpd.apache.org/docs/1.3/mod/mod_rewrite.html*.

Modification de tous les liens internes

Maintenant que nous avons défini les schémas d'URL et créé les règles de réécriture, il reste à vérifier que dans tout le site, tous les liens utilisent le bon schéma d'URL.

En effet, les règles de réécriture du fichier `.htaccess` ne suffisent pas à ce que tout votre site soit au nouveau format, avec des URL propres ! C'est à vous de changer la façon d'écrire les liens, que ce soit dans des pages statiques ou dans des pages dynamiques.

Bien entendu, vous devez pouvoir sauter cette étape si vous incluez la gestion de la réécriture d'URL dès la création du site, puisque vous aurez pris soin de générer dès le début des liens aux bons formats.

Mise à jour de test

Il est temps de tester. Transférez tous les fichiers modifiés en ligne, y compris le fichier `.htaccess`, puis rendez-vous dans votre navigateur pour vous assurer que la réécriture fonctionne.

Pour reprendre notre exemple, comparez ce que vous obtenez en allant sur :

http://www.notre-site.com/articles/article-12-2-5.html

et sur :

http://www.notre-site.com/articles/article.php?id=12&page=2&rubrique=5

Vous devriez avoir exactement la même page…

En cas de blocage complet du site (avec une erreur de type 500, par exemple), n'oubliez pas qu'il suffit de supprimer le fichier `.htaccess` (ou d'annuler les dernières modifications) pour que tout revienne dans l'ordre.

Nous vous conseillons d'utiliser un logiciel de vérification des liens au sein de votre site (vous pouvez, par exemple, choisir Xenu's Link Sleuth, à télécharger à l'adresse suivante : *http://home.snafu.de/tilman/xenulink.html*). Ce type de logiciel agit comme Googlebot : il parcourt vos pages en suivant tous les liens qu'il trouve. S'il ne trouve aucun lien mort

(un lien menant à une page introuvable), alors vous n'avez fait aucune erreur ni dans vos règles de réécriture ni dans vos liens internes. Sinon, corrigez en conséquence.

Optimisation automatique de toutes les pages

Une fois que vous avez mis en place la réécriture d'URL sur tout votre site, deux étapes restent à effectuer pour terminer son optimisation (du point de vue du référencement) :

1. Créer des liens vers toutes les pages

   Les sites dynamiques comportent bien souvent un grand nombre de pages. La mise en place de la réécriture d'URL permet une bonne indexation, mais ce n'est pas une condition suffisante pour que toutes les pages de votre site soient indexées. En effet, il est nécessaire de créer les conditions pour que les robots des moteurs puissent accéder à vos pages en suivant les liens présents sur votre site.

   – Si vous avez une rubrique contenant des articles (actualité, par exemple), prévoyez une zone d'archives avec des liens vers tous les articles, hiérarchisés de manière chronologique.

   – Si vous avez un forum avec des milliers de discussions, vérifiez que tous les liens qui permettent de naviguer de page en page utilisent le bon format d'URL. Vous pouvez également prévoir là aussi une rubrique d'archives, avec des liens vers tous les forums et toutes les discussions des forums, le tout réparti sur autant de pages que nécessaire (limitez-vous à une centaine de liens par page environ, éventuellement un peu plus).

   – Si vous avez un catalogue de produits, vous avez certainement classé ces derniers en catégories, sur un ou plusieurs niveaux. Présentez ces produits sous la forme d'un annuaire qui permet de naviguer dans tout le catalogue avec des liens classiques `<a href="">`. Cet annuaire peut être complété par un moteur de recherche interne, souvent très apprécié des internautes, et compatible bien entendu avec votre catalogue de produits.

   – Une page « Plan du site » adaptée peut également être créée dans cette optique (voir fin de ce chapitre).

   En d'autres termes, tous les liens à l'intérieur de votre site devront maintenant apparaître sous leur forme optimisée grâce à l'URL Rewriting. Les spiders ne devront plus trouver dans vos pages l'ancienne version des URL.

2. Optimiser le code de chaque page dynamique

   Une page dynamique n'est rien d'autre qu'une page HTML créée sur mesure par un script. En général, une telle page repose sur un modèle de page, reprenant le design du reste du site, et comportant certaines zones dont le contenu est généré en effectuant des requêtes dans une base de données.

   Nous avons vu dans cet ouvrage (chapitre 4) comment optimiser le code d'une page HTML pour le référencement. Naturellement, vous pouvez faire la même chose avec des pages dynamiques. Si vous respectez ces consignes, vous disposerez rapidement

d'un site dont les milliers de pages seront indexées et toutes optimisées pour le référencement !

Pour conclure, on peut dire que la mise en place de la réécriture d'URL est un travail parfois long, complexe et technique, mais qui permet d'obtenir des résultats sans commune mesure avec les sites statiques. Une fois bien mise en place, la réécriture d'URL – associée à une optimisation dynamique des pages – permet bien souvent de positionner le site sur Google ou les moteurs de recherche du marché pour des milliers d'expressions plutôt que quelques dizaines comme c'est le cas habituellement avec les sites statiques.

Identifiants de session

Les identifiants de session permettent, notamment aux sites de commerce électronique, de garder des éléments en mémoire au travers d'une navigation unique. Un identifiant, sous la forme d'une suite de lettres et de chiffres, est alors indiqué dans l'URL des pages consultées. Exemple :

http://www.fnac.com/livres.asp?NID=-1&RNID=-1&SID=6dfbd5e4-e0d7-a61f-8a96-15a11e2f478f&Origin=FnacAff&OrderInSession=1&UID=14C2FAADA-AE12-C964-759B-7640FAEC5548&TTL=061020071056&bl=HGACong2[1pro]liv

Le paramètre représentant l'identifiant de session est ici `SID` (pour *Session ID*). Or ce paramètre, quel que soit le nom qui lui est alloué dans l'URL, pose problème aux moteurs de recherche car il change pour chaque visite. Ainsi, une même page, visitée chaque jour par un robot, se verra attribuer un identifiant de session différent à chaque fois, et donc une URL différente. Problème quasi insoluble. Le moteur choisit alors, le plus souvent, de ne pas indexer la page.

On trouve cependant, dans les index des moteurs, quelques-unes de ces pages. Tapez des requêtes comme « inurl:session_id » ou « inurl:sessionid » sur Google. Vous obtiendrez alors quelques milliers, voire dizaines de milliers de pages. Ceci dit, il est clair que l'identifiant de session est un problème assez important et bloquant pour les moteurs. Il s'agit certainement de l'un des plus bloquants à l'heure actuelle.

Il est d'ailleurs conseillé :

- d'appliquer un numéro de session le plus tard possible dans la navigation (donc en évitant ce type de système sur la page d'accueil, la page de présentation des produits et en ne l'appliquant – par exemple – qu'à partir du moment où une réelle vente est en cours) ;
- d'utiliser de préférence les cookies (voir plus loin), qui permettent également ce type d'action et posent moins de problèmes aux moteurs. Mais cela n'est pas une solution technique simple si le site a été réalisé, au départ, en tenant compte des identifiants de session. Il n'est pas toujours facile de revenir en arrière sur ce point ;
- de passer à un système d'URL Rewriting qui peut, dans certains cas, résoudre quelques problèmes (voir paragraphe précédent) ;

- de créer des pages statiques de présentation des produits principaux, pour les moteurs de recherche et ne contenant pas d'identifiant de session dans l'intitulé de leur URL ;
- d'éviter tout nom de paramètre qui contienne l'appellation « `id` » (`sid`, `s-id`, `mid`, `r_id`, etc.) pour éviter que cet intitulé soit pris pour un identifiant de session par le moteur, même si ça n'est pas le cas…

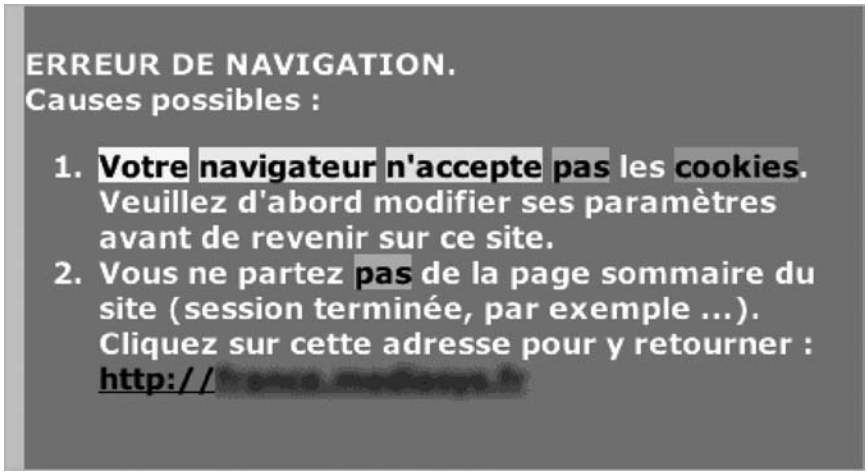
Cookies

Les cookies sont une autre façon de récupérer des paramètres et des données lors de la navigation d'un internaute d'une page à l'autre. Ces informations s'inscrivent dans un fichier, présent sur le disque dur de l'utilisateur, et le site va y puiser les indications dont il a besoin pour vous identifier.

En soi, les cookies ne posent pas de problèmes aux robots des moteurs, sauf que ce ne sont pas des navigateurs et qu'ils ne peuvent pas les « accepter ». Malheureusement, certains sites ne prévoient pas ce cas et le spider se voit afficher une page comme celle représentée à la figure 7-15 en lieu et place du document recherché.

Figure 7-15

Il faut penser aux robots des moteurs qui ne prennent pas en compte les cookies.



Pas très informatif, n'est-ce pas ? Il suffit donc que votre site web délivre une information (page « Plan du site », par exemple) au cas où le visiteur qui arrive ne soit pas un navigateur web du style Internet Explorer ou Firefox. Un visiteur qui n'accepte pas les cookies doit quand même pouvoir lire vos pages, sous une forme ou une autre et ne pas bloquer sur un message d'erreur.

Autre façon de pallier cette contrainte de façon partielle : utiliser les cookies le plus tard possible, comme pour les identifiants de session. Mais cela ne résout pas vraiment le problème puisque vos présentations de produits risquent de ne pas être indexées, ce qui serait dommage…

Accès par mot de passe

Ce problème sera assez rapidement traité : si votre site est accessible uniquement par mot de passe, vos pages ne seront donc pas indexées car les robots des moteurs n'en disposent pas. Si vous désirez obtenir une certaine visibilité sur les moteurs de recherche, vous devrez donc créer une zone accessible librement sur votre site à l'attention des internautes « non abonnés » mais aussi des robots. Il n'y a pas d'autre voie possible.

Tests en entrée de site

Beaucoup de sites web effectuent des tests en entrée de site en fonction de certaines données : type et langue du navigateur, adresse IP, résolution d'écran utilisée, etc. Cela ne pose pas, *a priori*, de problème aux robots des moteurs si leur cas n'est pas oublié, tout comme pour les cookies vus auparavant. Si vous faites un test sur le type de navigateur (« si Explorer alors…, si Firefox alors… »), par exemple, n'oubliez pas une condition « sinon… » pour traiter tous les autres cas, dont celui des robots. Sinon, ceux-ci seront bloqués à l'entrée de vos pages et ne pourront pas aller plus loin. Cela semble évident, mais on a déjà vu des cas bien pires sur le Web ces dernières années. Exemple : un site web qui effectue un test linguistique (basé sur une géolocalisation de l'adresse IP) et renvoie l'internaute vers le site anglais ou français selon le cas. Malheureusement, les adresses IP des spiders étant localisées aux États-Unis, le site français n'était jamais référencé... Voir à la fin de ce chapitre le cas des sites multilingues.

Redirections

Il arrive souvent que, suite au remaniement d'un site ou pour toute autre raison technique ou organisationnelle, une page, un répertoire ou un site entier change d'adresse. Dans ce cas, si votre site est déjà référencé sur les moteurs de recherche du Web, comment leur faire savoir que votre adresse a changé, et ce sans souci technique bien sûr ? Existe-t-il une solution fiable et transparente à la fois pour l'internaute et les robots des moteurs pour mettre en place cette redirection de l'ancienne adresse vers la nouvelle ?

Il existe en fait plusieurs façons de mettre en place une redirection d'une page web vers une autre : JavaScript, balise meta `refresh`, redirection 301 et 302, etc.

Aujourd'hui, la meilleure façon de faire est d'utiliser une redirection 301 (indication de redirection définitive). Si vous pouvez la mettre en place, n'hésitez pas, c'est celle qui est la mieux prise en compte par les moteurs. Elle est pérenne et fonctionne parfaitement.

En effet :

- La redirection JavaScript peut être assimilée à du spam (les pages satellites sont souvent basées sur cette technique). Elle est donc déconseillée. Elle n'est, de plus, pas lue par les moteurs dans la majeure partie des cas.
- La balise meta `refresh` fonctionne pour l'internaute mais elle ne permet pas, notamment, de transmettre le PageRank. Ainsi, si la page A, de PageRank 6, redirige vers B à l'aide

d'une balise meta `refresh`, B gardera son propre PageRank, celui de A ne lui sera pas transmis. Si tous les liens du Web pointent vers A, alors B aura un très faible PageRank.

- La redirection 301, quant à elle, transmet le PageRank d'une page à l'autre, d'où son indéniable avantage (la redirection 302, qui est une redirection temporaire, ne transmet pas ce PageRank).

Redirections en cascade et sur la page d'accueil

Il est important de rappeler ici deux points capitaux en ce qui concerne les interactions des redirections sur le référencement de votre site, et ce même si vous utilisez des redirections 301 :

– Nous vous conseillons de ne pas faire de redirection (même en 301) sur votre page d'accueil. Par exemple, si vous faites une redirection 301 de l'adresse *www.votresite.fr* vers *www.votresite.fr/fr/home/index.html*, cela peut poser des problèmes à certains moteurs. En théorie, cela devrait bien se passer, mais en pratique, nous avons étudié plusieurs cas où la redirection était mal analysée par Google. À éviter le plus possible donc.

– Ne jamais faire plusieurs redirections en cascade sur une page web. Par exemple : *www.votresite.fr/fr/* qui redirige (301) vers *www.votresite.fr/fr/home/*, qui elle-même redirige (301) vers *www.votresite.fr/fr/home/index.html*. Il y a de fortes chances pour que les moteurs ne suivent pas la deuxième redirection...

Nous allons donc nous pencher sur cette fameuse redirection 301 pour indiquer aux moteurs qu'une page a changé d'emplacement. De quoi s'agit-il et pourquoi cette redirection porte-t-elle ce nom ? En fait, à chaque fois qu'un navigateur accède à une page web, le serveur renvoie l'en-tête HTTP de la page avant la page elle-même. Cet en-tête contient quelques informations à propos du document, dont son statut sous la forme d'un code.

Vous connaissez certainement l'erreur, ou plutôt le code 404, qui s'affiche sur l'écran de votre navigateur lorsqu'une page que vous désirez afficher a disparu. Mais ce code n'est pas le seul existant, loin de là. Ces derniers sont classés par familles dont voici un résumé très sommaire :

- 100 : information ;
- 200 : OK, tout va bien ;
- 300 : redirection ;
- 400 : erreur au niveau du document demandé ;
- 500 : erreur sur le serveur.

Liste des codes HTTP

Vous trouverez à l'adresse suivante le document officiel de description des codes de statut (dont 404 et 301) du protocole HTTP :

http://www.w3.org/Protocols/rfc2616/rfc2616-sec10.html

Ainsi qu'une liste de ces codes à cette adresse :

http://www.indexa.fr/service/codes-http/index.phc?lang=uk&proc=1xx

Le code 301 signifie *Moved Permanently*. Cela veut dire que la page que vous vouliez atteindre a été déplacée de façon permanente à une autre adresse. Lorsqu'un robot va venir sur votre site pour crawler vos pages, il recevra et lira ce code si certains documents ont été déplacés. Encore faut-il bien configurer votre serveur pour ce faire…

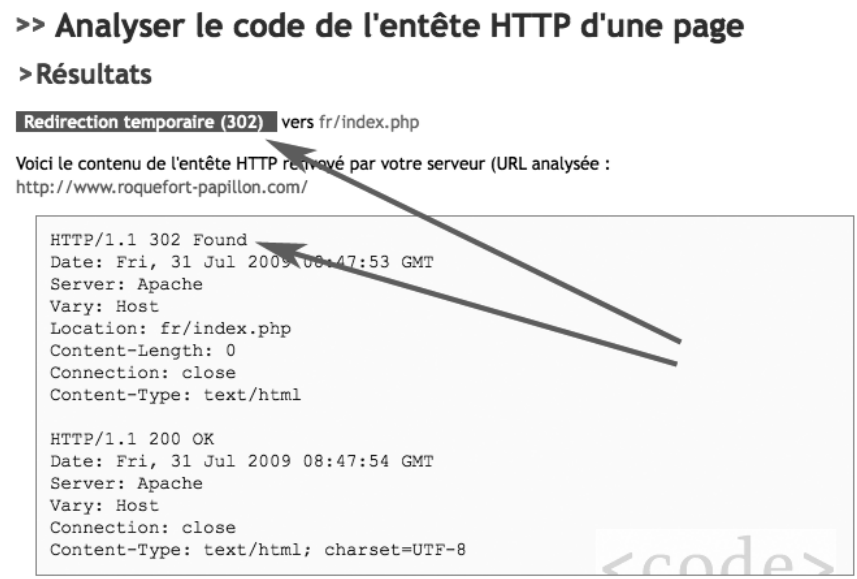


Figure 7-16

Analyseur d'en-tête http : ici, une redirection 302 a été mise en place sur le site testé.

> **Analyseur de code HTTP**
>
> Pour savoir si une redirection est faite sur une page et connaître son code de redirection, vous pouvez utiliser un analyseur comme celui du site WebRankInfo (*http://www.webrankinfo.com/outils/header.php*) : vous tapez une adresse URL et l'outil vous indique quel code le serveur renvoie (200, 302, 301, etc.). Très intéressant.

Si, par exemple, vous désirez rediriger tout accès à un fichier – quel qu'il soit – d'un répertoire donné vers une autre page ou un nouveau site, vous devez créer, dans l'ancien répertoire ou à la racine de votre site, un fichier nommé `.htaccess` et contenant la ligne suivante :

```
RedirectPermanent ancienne-adresse nouvelle-adresse
```

Exemple :

```
RedirectPermanent /ancien-repertoire http://www.nouveausite.com/
```

Les robots de Google et des autres moteurs lisent sans problème ces données, suivent ce type de redirection et remplacent sans souci l'ancienne page par la nouvelle. Ils transfèrent également le PageRank de l'ancienne page vers la nouvelle si la redirection et de type 301 (mais pas 302, qui correspond à une redirection temporaire).

Google en parle d'ailleurs dans son aide en ligne (*http://www.google.com/support/webmasters/bin/answer.py?hl=fr&answer=93633*) :

« Lorsque votre nouveau site est prêt, nous vous conseillons d'insérer le code de redirection permanente « 301 » dans les en-têtes HTTP de votre ancien site pour indiquer aux visiteurs et aux moteurs de recherche que votre site a changé d'adresse. »

Autre façon d'effectuer cette redirection si votre site accepte le langage de programmation PHP : ajouter dans chaque ancienne page l'en-tête suivant :

```
<?php
  header("Status: 301 Moved Permanently");
  header("Location: http://www.votrenouveausite.com/");
  exit();
?>
```

En langage ASP, une fonction similaire existe :

```
<%
  response.status = "301 moved permanently"
  response.addheader "location", "http://www.votrenouveausite.com/"
  response.end
%>
```

Il est également possible de créer une redirection 301 dans un fichier `.htaccess` au travers d'une règle de réécriture (URL Rewriting) :

```
RewriteEngine on
RewriteRule ancien_fichierl.htm  http://www.nouveau-site.com/nouveau-fichier.htm [R=301]
```

L'avantage du fichier `.htaccess` est qu'avec une seule commande, vous pouvez rediriger tout accès à un fichier, quel qu'il soit, vers une nouvelle adresse alors que les commandes PHP ou ASP doivent être insérées dans chaque page et induisent donc la présence effective d'un document à chaque ancienne adresse, ce qui est obligatoirement plus lourd à mettre en place.

Plus d'informations sur les redirections 301

Voici quelques liens qui vous en diront davantage à ce sujet :

– http://www.webrankinfo.com/dossiers/debutants/initiation-aux-redirections
– http://www.webrankinfo.com/dossiers/techniques/redirections-sauvages
– *http://www.pandia.com/sw-2004/40-hijack.html*

Hébergement sécurisé

De nombreux sites proposent un espace sécurisé, soit pour leurs clients, soit pour saisir un numéro de carte bancaire, soit pour d'autres raisons. S'il est logique que des pages qui contiennent des informations personnelles (accessibles, par exemple, sur saisie d'un mot de passe) ne soient pas disponibles pour les spiders des moteurs, on peut, en revanche, se poser la question de pages d'informations comme celles qui présentent la solution AdWords de Google, disponibles à l'adresse suivante : *https://adwords.google.fr/select/*.

Toute la partie description et FAQ de l'offre commerciale se trouve, par exemple, dans un hébergement sécurisé à l'adresse suivante : *https://adwords.google.fr/support/?hl=fr*.

Il serait logique que ces pages, qui n'ont rien de personnel par rapport à l'internaute, se retrouvent dans les index des moteurs. Mais ceux-ci les indexent-ils ? Le fait que les adresses de ces pages soient en *https* est-il un frein ou un blocage à leur indexation ou leur positionnement ?

À la mi-2009, les pages présentes en zone sécurisée (commençant par une adresse de type *https*) ne semblaient présenter de problèmes ni pour Google ni pour Yahoo!. En revanche, Bing semblait les ignorer. Est-ce partie remise uniquement pour ce dernier, au vu de la relative jeunesse de cette technologie ? Cela reste à voir.

Attention, cependant, de ne pas tirer de conclusions trop hâtives pour Google et Yahoo! : si, selon nos tests, ce type de page ne semble pas leur poser de problèmes, certains référenceurs nous ont fait part de difficultés, dans le passé, pour indexer et positionner les pages de certains sites présentant cette particularité. Prudence donc, un test préalable peut s'avérer indispensable selon le site pris en compte. À ce sujet, plusieurs points plus précis nous ont été fournis par ces sociétés de référencement. Nous vous les livrons « tels quels », car ils sont parfois complexes à vérifier techniquement.

- Il semblerait que les moteurs comme Google et Yahoo! aient pris comme stratégie d'indexer les pages web en *https*, uniquement si le site propose également des pages en *http* par ailleurs. En d'autres termes, si un site est accessible uniquement et exclusivement en *https*, sans aucune page disponible en version non sécurisée, cela peut poser des problèmes d'indexation. En tout état de cause, nous vous conseillons de créer des versions non sécurisées de vos pages d'informations en *https* afin de faciliter l'indexation et le positionnement de votre site sur tous les moteurs, notamment sur Bing.
- Le fait qu'une page soit en *https* peut poser des problèmes au niveau du calcul du PageRank de Google. En effet, le « s » supplémentaire est parfois oublié dans les liens qui pointent vers le site. Ainsi, si la version non sécurisée du site existe, c'est elle qui profite du lien, et non pas la version sécurisée. La popularité d'une page en *https* risque donc d'être moins importante qu'une page en *http*… Le fait de créer une version non sécurisée (*http*) évite également une erreur 404 sur un lien au sein duquel le « s » aurait été oublié.
- Certains hébergeurs peuvent bloquer les robots, par exemple à l'aide du fichier `htppd` (Apache), pour l'accès des pages sécurisées, et ce pour plusieurs raisons (sécurisation, charge serveur, etc.). Vérifiez donc auprès de votre hébergeur si ce n'est pas votre cas.

- Certains moteurs peuvent également vérifier le certificat du site : Est-il expiré ? Qui l'a émis ? On peut imaginer que le moteur refuse une page correspondant à un certificat dont la date de validité a expiré. De la même façon, il se peut que le robot refuse d'indexer une page sécurisée dont le certificat n'aurait pas été émis par une autorité de confiance reconnue. Exemple, si c'est Verisign qui l'a émis, l'indexation est effectuée, mais si l'émetteur est inconnu, cela peut poser problème.

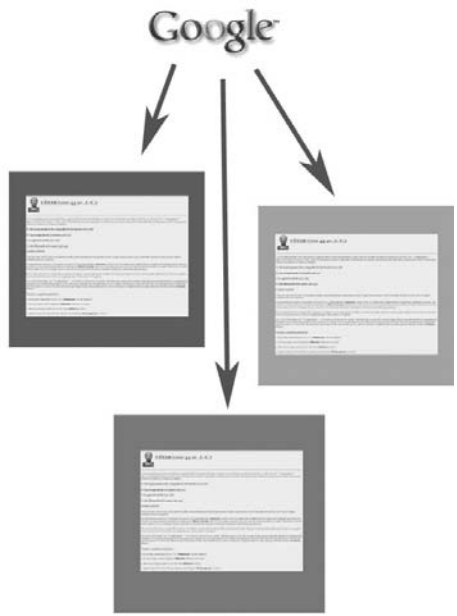
Duplicate content : un mal récurrent...

Depuis de nombreuses années, on entend parler, dans le monde du référencement, de problèmes dus au concept de duplicate content. De quoi s'agit-il exactement ? Quels types de problèmes ce phénomène pose-t-il et quels sont les remèdes possibles ? Nous allons essayer, dans ce chapitre, de répondre à toutes ces questions et de voir comment faire en sorte que le duplicate content ne soit plus qu'un mauvais souvenir pour vous si vous souffrez actuellement de ce type de problème (comme c'est le cas de très nombreux sites de la Toile)...

Mais tout d'abord, qu'est-ce que le duplicate content ? En fait, il s'agit d'une situation assez simple en soi : imaginons que Google (et les autres moteurs, bien sûr) ait, à un moment donné, indexé une ou plusieurs pages (sur le même site ou sur des sites différents) qui, selon lui, proposent un contenu identique ou, tout du moins, très proche, très similaire, comme le montre la figure 7-17.

Figure 7-17

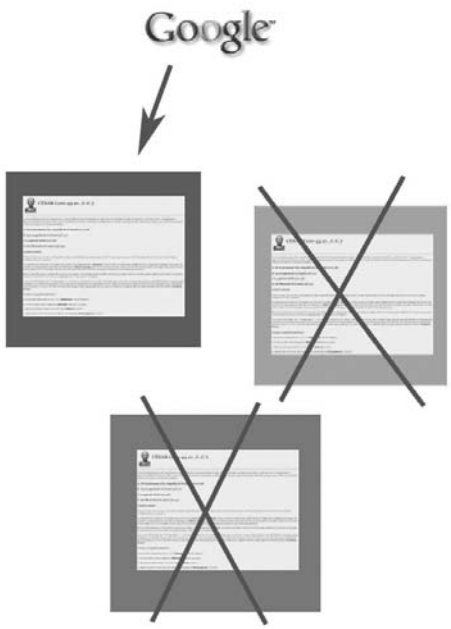
Google trouve sur le Web trois pages aux contenus éditoriaux très similaires, même si la mise en page/charte graphique est différente dans chacun des trois cas.



Il ne désire pas garder dans son index toutes ces pages trop proches les unes des autres et il décide donc de n'en garder qu'une seule. Ce sera celle qui, selon lui, propose le contenu « original », qui a donc été « copié » par les autres documents. Il prend en compte ce contenu original, qu'il appelle « canonique » et délaisse les autres pages qui deviennent pour lui « dupliquées » (figure 7-18).

Figure 7-18

Google choisit le contenu canonique pour ses pages de résultats.



Notons bien qu'il ne supprime pas les pages contenant le contenu dupliqué, mais qu'il les met dans son index secondaire (voir chapitre 2) et y donne accès au moyen d'un lien en bas de page s'il détecte un phénomène de duplicate content, comme indiqué sur la figure 7-19.

Pour limiter les résultats aux pages les plus pertinentes (total : 5), Google a ignoré certaines pages à contenu similaire. Si vous le souhaitez, vous pouvez relancer la recherche en incluant les pages ignorées.

Figure 7-19

Google donne accès aux pages considérées comme dupliquées au travers d'un lien, en bas de sa page de résultats.

Ce traitement, finalement assez logique, permet à Google de ne pas avoir trop de « doublons » dans son index et de fournir à ses utilisateurs des résultats plus pertinents. Cependant, il existe bon nombre de cas où cette notion de duplicate content peut poser des problèmes aux éditeurs de sites web. C'est ce que nous allons étudier maintenant...

Un détecteur de duplicate content

Vous pouvez utiliser le Similar Page Checker du site Webconfs (*http://www.webconfs.com/similar-page-checker.php*) pour estimer le taux de duplicate content entre deux pages. Vous saisissez les deux URL et l'outil vous indique le pourcentage de similarité entre les deux documents. Intéressant. On trouve des outils similaires (mais pas dupliqués...) chez WebRankInfo (*http://www.webrankinfo.com/outils/similarite.php*) et Iseom (*http://www.iseom.com/similar-page-checker1.html*) sans oublier un site dédié à l'adresse *http://www.duplicatecontent.net/*.

Problème 1 – Contenu dupliqué sur des sites partenaires

Le problème du duplicate content arrive très rapidement lorsqu'un même contenu se trouve sur des sites différents. Exemple type : une dépêche AFP qui va se trouver sur le site de l'agence de presse qui l'a conçue, mais également sur de nombreux sites web « officiels » qui la reprennent.

Autre exemple : un site web de contenu propose un article en ligne sur un sujet donné (mode, tourisme, sport, etc.) et cet article est repris par un site web partenaire, qui a signé un contrat pour avoir le droit de reprendre ce contenu.

Aujourd'hui, Google sait très bien « extraire » le contenu réel, éditorial, d'une page web et laisser « de côté » toute la partie « navigation/charte graphique » du code HTML. Quand il aura fait ce travail sur les deux pages contenant l'article en question, il sera en possession de deux textes strictement identiques. Dans ce cas, quelle version va-t-il prendre en compte ? La question n'est pas si simple et le choix risque d'être cornélien pour lui... Officiellement, Google indique qu'il prendra comme page canonique celle qui a le plus fort PageRank, celle qui est la plus populaire. Mais ce n'est peut-être pas le choix qui vous arrange le plus... Il faudra alors que vous l'aidiez pour le faire changer d'avis...

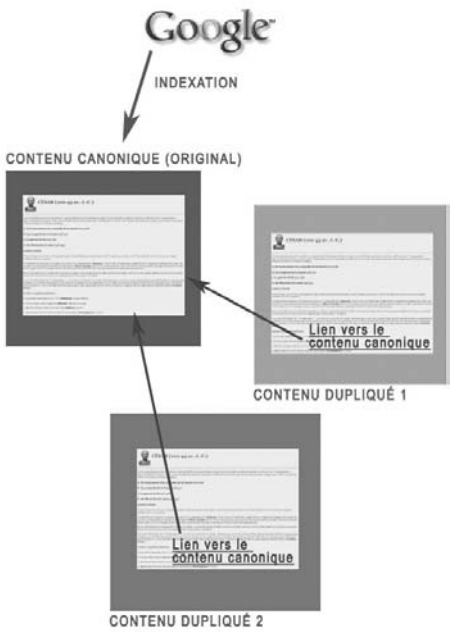
Figure 7-20

Exemple d'article repris à l'identique sur deux sites web différents

En effet, Google doit ici reconnaître quel est le contenu « canonique » (original) et quel est celui qui est dupliqué. Pour cela, il existe une façon de faire (recommandée d'ailleurs par Google) en demandant à vos partenaires – si vous êtes le propriétaire du contenu canonique – de mettre (si ce n'est déjà fait) un lien sur leur page dupliquée vers votre page canonique. Attention : pas un lien vers la page d'accueil du site canonique. Chaque page reprenant un de vos contenus doit « pointer » vers la page du site affichant le contenu original. Ceci est extrêmement important !

Figure 7-21

Lien depuis les pages dupliquées vers la page canonique



Ce lien vers le contenu canonique sera détecté par Google qui comprendra ainsi qu'un contenu est issu d'un autre et pourra « se faire son idée » sur la provenance originale du texte éditorial découvert. C'est également important pour Google News, outil sur lequel Google utilise fortement ses filtres de duplicate content car il est alors confronté quotidiennement à ce type de problème.

Sur cet outil, le lien « Trier par date et afficher les doubles » permet ainsi d'afficher les pages en duplicate content, triées et éliminées par défaut.

Figure 7-22

Détection du duplicate content dans Google News

Différences entre contenu éditorial et charte graphique

Notez que, même si vos pages ont 90 % de leur code HTML (représentant toute la partie « navigation, etc. ») identique, seul le contenu réel – éditorial – sera pris en compte par Google dans la détection du duplicate content (DC). Deux pages peuvent donc avoir un code HTML très différent mais un contenu éditorial identique. Cela ne leur évitera pas de tomber dans les filtres de « DC ». Ne l'oubliez pas !

On peut penser que le TrustRank, ou indice de confiance (voir chapitre 5), des différents sites entrant en ligne de compte ici joue également son rôle, Google octroyant plus de confiance au site web disposant du TrustRank le plus élevé. De même, la date de publication (dans le Sitemap spécifique à Google News – voir chapitre 8 – ou la date de découverte de l'article par le moteur) a bien entendu son importance dans la somme de critères qui lui permettent de définir le contenu canonique. D'autres points comme l'univers sémantique des liens de la page, peuvent entrer en ligne de compte.

Autre solution pour indiquer aux moteurs (Google, Yahoo! et Microsoft) qu'une page est dupliquée et une autre canonique : ajouter dans le code HTML de chaque version « dupliquée » l'indication suivante (dans la zone `<head>` du code HTML) :

```
<link rel="canonical" href="http://www.votresite.com/page-canonique.html" />
```

En remplaçant, bien sûr, l'adresse *http://www.votresite.com/page-canonique.html* par celle que vous désirez mettre en avant. Pour plus d'informations sur cette balise (proposée en janvier 2009 par les trois principaux moteurs), consultez la page suivante : *http://google.com/support/webmasters/bin/answer.py?answer=139394*.

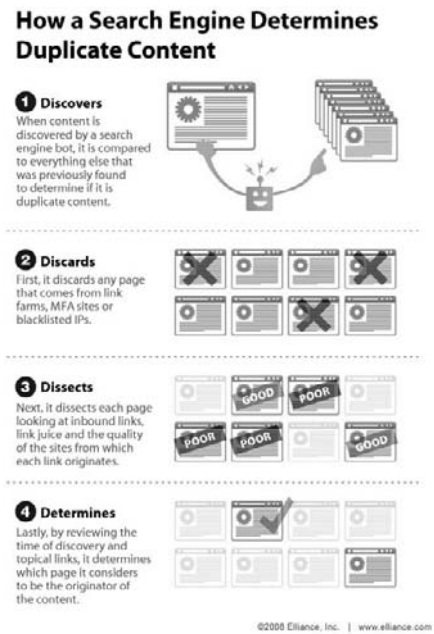
Autre solution pour éviter que votre contenu ne se trouve « dans la charrette » au profit de celui d'un de vos partenaires (qui aura mieux optimisé ses pages que vous...) : prévoir, dès le début du partenariat, que ses pages ne doivent pas être référencées. Par exemple, par l'ajout d'une balise meta `robots` avec valeur `noindex` ou *via* un fichier `robots.txt` adéquat (voir chapitre 9 pour ces deux notions). Le partenaire a ainsi le droit de reprendre votre contenu sur son site, mais il doit « barrer le passage » aux spiders des moteurs.

Bien entendu, cette solution est beaucoup plus facile à mettre en place avant négociation et signature du contrat qu'après... Cette situation est également valable pour le « rétrolien » vu auparavant. Obliger vos partenaires à insérer un lien vers votre contenu canonique doit être inclus dans le contrat que vous signerez avec lui. Il vous faudra ensuite bien vérifier que ce rétrolien est présent dans ses pages. De même, si vous mettez en ligne un blog ou, plus simplement, du contenu sous la forme d'articles, etc., proposez une « charte de reprise du contenu » dans laquelle vous indiquez l'obligation de ce lien vers la page canonique, et ce même si c'est le fil RSS (titre + résumé) qui est repris. On n'est jamais trop prudent...

L'illustration de la figure 7-23 explique bien comment le duplicate content est appliqué par les moteurs de recherche.

Figure 7-23

Comment les moteurs de recherche appliquent leurs filtres de duplicate content aux pages web. Source : http://searchengineland.com/080513-080033.php



Problème 2 – Contenu dupliqué sur des sites « pirates »

Le problème explicité précédemment risque également de se poser de façon plus accrue si votre contenu est repris... par des sites qui ne sont pas vos partenaires... Dans ce cas, il sera encore plus énervant de voir l'un de ces contenus s'afficher en bonne position sur votre moteur favori alors que le vôtre est passé dans les affres des filtres du duplicate content.

Bien entendu, il sera difficile de demander à un site web avec qui vous n'êtes pas « en affaires » de mettre en place un lien vers vous ou une balise indiquant que son contenu est dupliqué... S'il a envie de le faire, il le fera, mais s'il n'a pas envie, (et il y a de fortes chances pour que cela soit le cas)... il ne le fera pas.

Quelle est la solution dans ce cas ? Premièrement, il vous faudra privilégier l'approche « amiable » en trouvant l'adresse e-mail du responsable du site « pirate » (sur ses pages ou *via* une fonction de Whois qui vous indiquera à qui appartient le nom de domaine) pour lui indiquer que votre contenu est soumis à copyright et qu'il n'a pas le droit de le reprendre ainsi. Dans certains cas, l'éditeur du site distant sera de bonne foi et il stoppera ses activités illicites. Parions cependant que cette première approche ne donnera pas toujours des résultats positifs... Mais elle doit cependant être tentée.

Dans le cas où l'approche « en douceur » ne donne pas de résultats, il vous faudra alors durcir le ton, faire constater (avocat/huissier) la fraude, envoyer un courrier recommandé

avec accusé de réception et demander à un avocat ou à votre service juridique qu'il brandisse la menace d'un procès si ce type de pratique ne cesse pas immédiatement. Si le site copieur est situé en France, le problème peut être réglé assez rapidement. S'il se trouve à l'étranger, les problèmes risquent de s'accumuler assez rapidement pour vous car la situation sera complexe selon le pays d'hébergement du site...

Dans ce cas, il vous faudra certainement lâcher prise et tenter de recevoir « le plus de rétroliens possible » de la part des éditeurs de sites web reprenant vos contenus. Les moteurs vont trouver et identifier ces liens et comprendre ainsi que c'est le vôtre qui est canonique, pas celui des pirates qui, eux, n'auront pas de rétrolien à proposer... C'est alors la page qui aura la plus forte popularité (PageRank), notamment par l'analyse des liens émanant des pages dupliquées, qui sera retenue par le moteur.

Une solution peut également être d'insérer des liens internes (vers d'autres pages de votre site) dans votre contenu rédactionnel (par exemple, des tags sur certains mots). Si le site pirate reprend votre contenu, il reprendra (peut-être...) ces liens internes (qu'il faudra donc indiquer en absolu : *www.votresite.com/tags/mot.html* – et pas en relatif : *../tags/mot.html* – dans votre code HTML). Cela sera une indication intéressante pour le moteur lorsqu'il analysera les textes à filtrer : un lien non pas vers votre page canonique mais au moins vers votre site, c'est toujours ça de pris.

Duplicate content global et partiel

Il faut également noter que Google, si l'on en croit les brevets qu'il a déposé à ce sujet, arrive aujourd'hui non seulement à détecter les pages « globalement semblables », mais également à identifier des parties de contenus (*snippets*) qui seraient repris dans d'autres pages... Le fait de reprendre un contenu et, par exemple, de modifier l'ordre de ses paragraphes, ne sera peut-être pas suffisant, selon les cas, pour éviter les filtres de Google...

Il est donc important, lorsque vous mettez en place un projet de reprise de vos contenus par un site tiers, de suivre ces quelques conseils :

- Pensez à cette problématique au moment de la mise en place du partenariat pour éviter tout souci par la suite : déréférencement des pages du site partenaire, mise en place d'un rétrolien, etc. Tout est important et doit être prévu à l'avance.
- Multipliez les liens externes vers votre contenu canonique.
- Créez, éventuellement, deux versions de votre contenu : l'une destinée à votre site, l'autre, moins riche, pour vos partenaires (par exemple, pour un descriptif produit)...
- Affichez sur votre site une « charte de reprise du contenu » si vous autorisez ce type de pratique (notamment *via* des fils RSS).
- Insérez des liens internes (en adressage absolu) dans vos contenus.
- Faites une veille pour savoir qui reprend vos contenus, soit en saisissant comme requête sur un moteur de recherche une ou deux phrases de vos articles entre guillemets soit en utilisant des outils comme Copyscape (*http://www.copyscape.com/*),

Compilatio (*http://www.compilatio.net/fr/*), CopyTracker (*http://copytracker.org/*), Noplagia (*http://code.google.com/p/noplagia/*) ou TinEye (*http://tineye.com/*) pour les images.

Quelques liens sur la notion de duplicate content

Voici quelques liens qui nous ont semblé intéressants dans le cadre d'une stratégie de lutte contre le duplicate content (de nombreux articles émanent des blogs officiels de Google) :

- Detecting Duplicate and Near-Duplicate Files (brevet de Google)
  http://www.seoguide.org/google-patent-6658423.htm
- Detecting Query-Specific Duplicate Documents (brevet de Google)
  http://www.seoguide.org/google-patent-6615209.htm
- Understanding SEO Issues Related to Duplicate Content (SEO Guide)
  http://www.seoguide.org/seo201-duplicate-content.htm
- Contenu dupliqué (Google - Centre d'aide webmasters/propriétaires de sites web)
  http://www.google.com/support/webmasters/bin/answer.py?hlrm=fr&answer=66359
- Deftly Dealing with Duplicate Content (Google)
  http://googlewebmastercentral.blogspot.com/2006/12/deftly-dealing-with-duplicate-content.html
- Duplicate Content Due to Scrapers (Google)
  http://googlewebmastercentral.blogspot.com/2008/06/duplicate-content-due-to-scrapers.html
- Ranking As the Original Source for Content You Syndicate (Vanessa Fox)
  http://www.vanessafoxnude.com/2008/05/14/ranking-as-the-original-source-for-content-you-syndicate/
- Duplicate Content Summit at SMX Advanced (Google)
  http://googlewebmastercentral.blogspot.com/2007/06/duplicate-content-summit-at-smx.html
- The Illustrated Guide to Duplicate Content in the Search Engines (SEOMoz)
  http://www.seomoz.org/blog/the-illustrated-guide-to-duplicate-content-in-the-search-engines
- Rewriting the Beginner's Guide Part IV Continued - Canonical and Duplicate Versions of Content (SEOMoz)
  http://www.seomoz.org/blog/rewriting-the-beginners-guide-part-iv-continued-canonical-and-duplicate-versions-of-content
- Faut-il avoir peur du duplicate content ? (RankSpirit)
  http://www.rankspirit.com/duplicate-content.php
- L'URL canonique, selon Google (Annuaire Info)
  http://www.annuaire-info.com/google-url-canonique.html
- Compléments de Matt Cutts sur le duplicate content (WordPress Tuto)
  http://wordpress-tuto.fr/complements-de-matt-cutts-sur-le-duplicate-content-307
- Duplicate content : Google, Microsoft et Yahoo! s'entendent sur une balise commune (Abondance)
  http://actu.abondance.com/2009/02/duplicate-content-google-microsoft-et.html

Problème 3 – Même page accessible via des URL différentes

La première partie de ce chapitre consacré au duplicate content a exploré la problématique du contenu canonique dupliqué sur des pages d'autres sites, qu'ils soient partenaires ou non. Mais il ne s'agit pas là de la seule problématique gravitant autour du phénomène de duplicate content, loin de là...

En effet, les webmasters sont souvent confrontés au fait qu'une même page web, unique – proposant strictement le même code HTML – , soit accessible par des URL différentes sur un même site. Cette situation est dommageable pour un référencement et nous allons tenter d'expliquer pourquoi.

On le sait, les algorithmes de pertinence des moteurs de recherche majeurs sont fortement influencés par l'analyse des liens externes et internes qui pointent vers une page et les notions de popularité (quantité et qualité des liens entrants) et réputation (textes des liens pointant vers la page). Reportez-vous au chapitre 5 pour en savoir plus sur tous ces concepts. Pour prendre un premier exemple simple, chaque lien vers vos pages est une pierre de plus à ajouter à la qualité de votre référencement.

Si votre page d'accueil est très populaire, elle va, au travers de ses liens internes, transférer de la popularité (on parle de *Link Juice* ou « jus de lien », voir chapitre 5) aux pages internes vers lesquelles elles pointent. Ce jus de lien transmis par les liens est important pour les moteurs de recherche. Or, si une même page est accessible par plusieurs adresses différentes, elle sera considérée comme autant de documents différents par les moteurs de recherche. Donc un document unique A sera « vu » par Google et consorts comme plusieurs pages A', A'' et A''', par exemple, chacune de ces pages répondant à une URL spécifique et détenant une fraction de la popularité globale de A...

Exemple d'une même page accessible *via* plusieurs URL distinctes (pointant toutes vers le même document) :

- *http://www.votresite.com/*
- *http://www.votresite.com*
- *http://votresite.com/*
- *http://www.votresite.com/index.html*
- *http://www.votresite.com/index.html?source=plandusite.html*
- *http://www.votresite.com/index.html?sid=12457845124578*

Il sera donc important que, sur votre site, **chaque page soit accessible par une seule et unique adresse (URL)** afin que l'analyse des liens internes (popularité et réputation) soit la plus fine, juste et efficace possible.

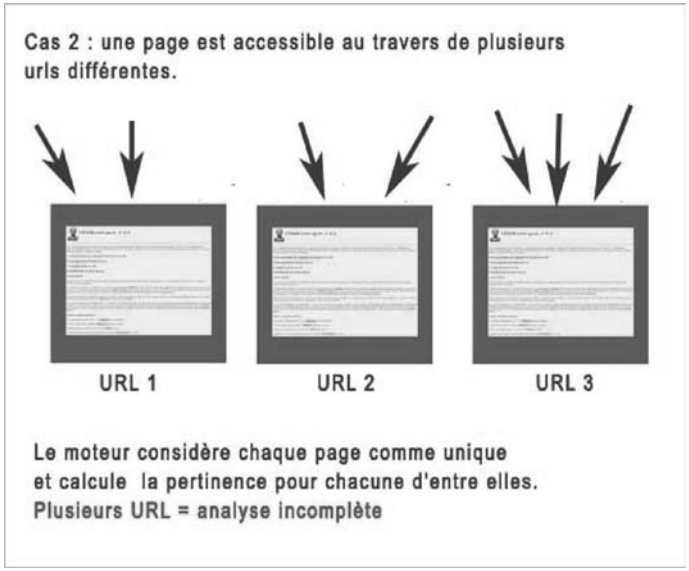
Nous allons voir, dans la suite de ce chapitre, quels sont les cas les plus fréquents de duplicate content de ce type et les solutions à y apporter.

Figure 7-24

Une page est accessible par une URL unique : l'analyse de sa popularité par les moteurs est complète.

Figure 7-25

Une page est accessible par plusieurs URL : l'analyse de sa popularité par les moteurs est parcellaire et morcelée.



Nom de domaine dupliqué

Le premier cas est assez fréquent : votre site est accessible par plusieurs noms de domaine, par exemple *www.abondance.com*, *www.abondance.net* et *www.abondance.fr*. Dans ce cas, la solution est simple : une redirection au niveau de votre DNS ou *via* un code 301 est à privilégier. Vous prenez, dans ce cas, en compte pour votre communication un seul nom de domaine (pour

nous : *abondance.com*) et vous redirigez tous les autres vers lui. Les redirections DNS et 301 sont bien interprétées aujourd'hui par les moteurs de recherche qui comprendront aisément que toutes ces adresses pointent vers un seul et unique site web. Pas de soucis majeurs.

Nom de site dupliqué

Le deuxième cas est également assez fréquent : votre page d'accueil est accessible sous des adresses de type *www.votresite.com* ET *votresite.com* (sans le préfixe *www*). C'est une bonne chose pour l'internaute car cela lui rend la saisie plus simple et plus rapide, mais cela peut aussi créer un phénomène de duplicate content pour les moteurs en rendant un seul site accessible sous deux adresses différentes...

Un fichier `.htaccess` bien conçu avec une règle de réécriture privilégiant l'une ou l'autre adresse (la plupart du temps celle avec *www*) résoudra le problème. Exemple :

```
RewriteEngine On
RewriteCond %{HTTP_HOST} !^www\.votresite\.com [NC]
RewriteRule (.*) http://www.votresite.com/$1 [QSA,R=301,L]
```

(Source : *http://www.webrankinfo.com/actualites/200510-contenus-dupliques.htm*)

Point important : Google vous propose également, dans ses Webmaster Tools (*http://www.google.fr/webmasters/*), dans la zone « Configuration du site>Paramètres », le choix « Domaine favori », qui permet d'indiquer au moteur quelle adresse « canonique » vous désirez que Google prenne en compte pour son indexation.

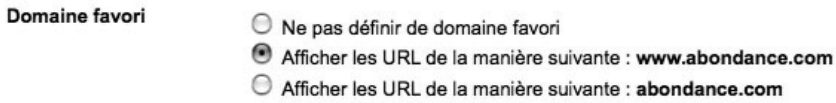


Figure 7-26

Les outils pour webmasters de Google vous donnent la possibilité de définir une adresse canonique pour votre site.

Ceci dit, cet outil n'étant disponible que pour Google et pas chez ses concurrents, cela ne vous dispensera pas de créer un fichier `.htaccess` idoine...

Adresse de la page d'accueil dupliquée

De la même façon, votre page d'accueil – toujours elle – est sûrement accessible à la fois au travers de l'adresse *http://www.votresite.com/* mais également d'une adresse de type :

- *http://www.votresite.com/index.html*
- *http://www.votresite.com/index.htm*
- *http://www.votresite.com/index.php*
- *http://www.votresite.com/accueil.php*
- etc.

Ces deux adresses (*www.votresite.com* et *www.votresite.com/****) risquent fort d'être considérées comme deux pages différentes également par les moteurs de recherche. Le problème est donc identique au cas précédent et une redirection 301 sera la bienvenue pour n'afficher qu'une seule adresse « canonique » (là aussi, la plupart du temps *www.votresite.com*).

De la même façon, évitez les liens vers des adresses comme *http://www.votresite.com* **et** *http://www.votresite.com/*. Le dernier slash (/) peut, là aussi, poser problème et on n'y pense pas forcément lorsqu'on crée les liens internes du site...

Choisissez donc l'adresse « canonique » que vous voulez pour votre page d'accueil, mais unifiez-la partout sur votre site dans les liens qui pointent vers elle. Google l'explique ici : *http://googlewebmastercentral.blogspot.com/2006/12/deftly-dealing-with-duplicate-content.html*.

Vérifiez bien également qu'au sein même de votre site, les liens, dans les pages internes, qui pointent vers votre page d'accueil pointent bien vers *www.votresite.com* et non pas vers un autre intitulé d'URL. On a souvent pas mal de surprises suite à cette vérification...

De la même façon, cette problématique peut se reproduire sur des pages internes et notamment des rubriques sommaires (*www.votresite.com/produits/* et *www.votresite.com/produits/index.html*) pour lesquelles le remède sera identique. Là encore, du travail de vérification en perspective...

Cas des sites dynamiques

Les sites dynamiques sont très intéressants de par les possibilités d'automatisation qu'ils proposent, mais ils sont aussi souvent source de soucis et de conflits dans les URL. On dénombre alors plusieurs problèmes qui peuvent subvenir en termes de duplicate content.

Cas 1 – Paramètres inversés dans l'URL

Par exemple, une même page pourra être accédée selon deux adresses :

http://www.votresite.com/catalogue?ref=123456&pays=fr&langue=fr

mais également :

http://www.votresite.com/catalogue?ref=123456&langue=fr&pays=fr

Ce sont deux pages exactement identiques, accessibles avec les mêmes paramètres mais pas dans le même ordre dans l'URL. Résultat : deux adresses distinctes et un cas classique de duplicate content. Là encore, il vous faudra vérifier, au sein de votre site, l'ordre dans lequel vous passez les paramètres dans vos URL et bien garder, à chaque fois, une stratégie cohérente sur l'ensemble du site à ce niveau.

Cas 2 – Pagination des listes

Ce cas est fréquent dans des pages qui listent des produits ou dans des fils de discussion de forums, par exemple. Une première page, listant un certain nombre d'« items « (discussions, produits, etc., le meilleur exemple restant encore une page de résultats de moteur de recherche...), sera accessible à l'adresse :

http://www.votresite.com/liste-produits?prod=telephones

Puis, si une deuxième page de produits est disponible (un exemple en est donné figure 7-27), celle-ci sera accessible (*via* le bouton Suivant, par exemple) à l'adresse :

http://www.votresite.com/liste-produits?prod=telephones&page=2

Figure 7-27

Exemple type d'une page listant des produits

Le problème surviendra si l'on revient (au travers du lien Précédent, par exemple) sur la page 1 et que l'on y accède *via* cette URL :

http://www.votresite.com/liste-produits?prod=telephones&page=1

Cette adresse est clairement différente de celle affichée en premier pour la même page... Attention donc à faire en sorte que tout accès à la première page se fasse sous la forme d'une URL unique. L'une ou l'autre (avec ou sans le paramètre indiquant le numéro de page) mais surtout unique !

Cas 3 – Réécriture d'URL

Si vous avez mis en place une réécriture d'URL (voir précédemment dans ce chapitre) sur votre site, vous devez avoir maintenant des adresses de type :

http://www.votresite.com/catalogue-telephone-nokia-810-kwt-fr-fr

au lieu de :

http://www.votresite.com/catalogue?ref=123456&pays=fr&langue=fr

C'est une bonne chose et votre référencement ne s'en portera que mieux. Mais n'oubliez pas, pour autant, de mettre en place, dans votre stratégie d'URL Rewriting, une

redirection 301 depuis l'ancien intitulé (avec les caractères `?` et `&`) vers le nouveau pour éviter tout souci. Ce serait trop bête d'œuvrer à créer des URL « propres » pour générer en même temps un phénomène de duplicate content...

Cas 4 – Plusieurs énoncés d'URL pour une même page

Ce cas se rapproche de celui qui concerne la duplication de la page d'accueil, déjà vu auparavant. Souvent, sur une page interne, lorsqu'on clique sur le logo générique ou sur un lien de type « Accueil », on est redirigé vers une adresse de type (ici, un exemple sous Lotus Notes) :

http://www.votresite.com/internet/webfr.nsf/0/08112EF3CAE171EBC12573930048A2C9?OpenDocument

au lieu de :

http://www.votresite.com/

Cela peut être dû à plusieurs raisons : vous désirez garder une indication sur la navigation de l'internaute et la page depuis laquelle il revient à l'accueil, votre système de navigation est tout simplement (*sic*) configuré ainsi, des identifiants de session sont automatiquement ajoutés dans l'URl, etc.

Là encore, les moteurs de recherche vont identifier votre page d'accueil au travers de plusieurs adresses distinctes, ce qui n'est pas une bonne chose. Le remède est toujours le même :

- soit vous simplifiez les URL pour toujours indiquer dans les liens leur intitulé « canonique ».
- soit vous mettez en place des redirections 301 depuis les intitulés « développés » (*http://www.votresite.com/internet/webfr.nsf/0/08112EF3CAE171EBC12573930048A2C9?OpenDocument*) vers les intitulés « canoniques » (*http://www.votresite.com/*). Encore une fois, ce conseil est valable pour n'importe quelle page du site et pas uniquement la page d'accueil...

Pour le cas des identifiants de session, toujours problématiques en termes de référencement, il faudra peut-être creuser une solution plutôt basée sur des cookies, qui laisses les URL « vierges » de toute indication de navigation et évite le phénomène de duplicate content, comme on l'a vu auparavant...

Quelques liens utiles...

Voici une nouvelle suite d'articles intéressants (pas si nombreux que cela, finalement, sur le Web), parlant des thèmes évoqués dans cette partie de chapitre, qui vous permettront certainement de creuser ces problématiques :

– How to Deal with Pagination & Duplicate Content Issues (SEOMoz)
http://www.seomoz.org/blog/how-to-deal-with-pagination-duplicate-content-issues

– Pagination and Duplicate Content Issues (Search Engine Journal)
http://www.searchenginejournal.com/pagination-and-duplicate-content-issues/7204/

– Lutter contre le duplicate content (Référencement, Design et Cie)
http://s.billard.free.fr/referencement/?2008/04/24/477-lutter-contre-le-duplicate-content

– Liste d'erreurs classiques de duplicate content (WebRankInfo)
http://www.webrankinfo.com/actualites/200703-erreurs-de-duplicate-content.htm

Problème 4 – Contenus proches sur un même site web

Pour terminer cette (longue) partie sur le duplicate content, nous allons envisager le cas de pages web proposant un contenu éditorial différent les unes des autres mais pouvant cependant tomber dans les affres des filtres de duplicate content, chez Google et ses concurrents.

Commençons par un exemple. La plupart du temps, on comprend vite les risques que l'on court en tapant sur Google (et les autres moteurs) la requête « site: » suivie du nom de domaine de son site. Si les résultats ressemblent à ce que l'on peut voir sur la figure 7-28, vous pouvez commencer à vous faire un peu de mouron.

Figure 7-28

Exemple d'un site dont de nombreuses pages ont la même balise <title>.

Dans cet exemple (« site:lepetitnicolas.net »), chaque page du site a le même contenu pour sa balise `<title>` (« Le site officiel du Petit Nicolas de Goscinny et Sempé »). Pire encore avec l'exemple de la figure 7-29 (« site:chateaudemontvillargenne.fr »).

Figure 7-29

Exemple d'un site dont de nombreuses pages ont la même balise <title> et la même balise meta description.

Dans ce cas, les balises `<title>` sont identiques sur de nombreuses pages, mais également les balises meta `description` (reprises dans le snippet ou résumé fourni par Google) en dessous. Faites le test sur votre site pour voir ce qu'il en est...

On pourrait multiplier ce type d'exemple à l'envi... Ils illustrent bien une problématique (souvent présente sur des forums non optimisés, par exemple) de duplicate content que l'on trouve finalement assez souvent : les moteurs de recherche, au moment de « filtrer » les contenus identifiés sur le Web, trouvent trop de similitudes dans le code HTML des pages et les classent en duplicate content même si leur contenu éditorial est différent.

La problématique peut également parfois être similaire mais légèrement différente : vous disposez d'un contenu (par exemple, des recettes de cuisine) que vous souhaitez proposer sur plusieurs de vos sites. Comment faire pour que la même recette ne soit pas considérée comme du duplicate content d'un site à l'autre et que chaque version soit prise en compte par Google ? On le voit, la problématique est ici l'inverse de celle étudiée dans le premier volet de ce chapitre...

Contenu éditorial différent, structures de page semblables : clairement différencier les codes HTML de chaque page

Prenons le premier cas : vos pages proposent un contenu textuel (éditorial) différent mais un code HTML trop proche. Comment faire ?

Tout d'abord, il sera essentiel de faire en sorte que le début du code HTML de chacune de vos pages (la partie `<head>`) soit clairement différent d'une page à l'autre. Proposez des

balises `<title>`, meta `description` et même meta `keywords` (pour Yahoo!) très descriptives et différentes d'une page à l'autre, que ce soit au niveau de la taille, de la structure et du contenu.

Évitez notamment, si possible, les structures trop redondantes, trop « reconnaissables » et « automatisées » comme sur la figure 7-30 (« site:placedestendances.com »).

Figure 7-30

Exemple d'un site pour lequel les pages web ont une balise <title> et des balises meta description dont la structure est trop répétitive.

Seule une faible portion du contenu des balises `<title>` et meta `description` est ici différente d'une page à l'autre (des paramètres en faible nombre sont en fait changés sur chaque page). N'hésitez pas à vous différencier plus que cela... Bref, faites en sorte que la partie `<head>` de vos pages soit très différente d'une page à l'autre et tout devrait bien se passer...

Ensuite, attaquez le corps de la page (partie `<body>` du code HTML). Toute la partie « header » (haut de page), « footer » (bas de page) et « liens de navigation interne » ne devraient pas poser de problème, Google sait différencier ce contenu de la partie plus strictement éditoriale (si Google classait en duplicate content toutes les pages qui ont le même header et le même footer, il n'y aurait plus beaucoup de documents dans son index...).

N'oubliez pas de donner du contenu en quantité suffisante (100 à 200 mots descriptifs au minimum) aux moteurs. Le titre (balise `<h1>`) et le chapô (premier paragraphe, trois à quatre premières phrases) doivent également être bien différenciés d'une page à l'autre.

Si vous suivez ces conseils, vous ne devriez pas avoir de trop gros problèmes du type de ceux évoqués au début de ce chapitre...

Cas du contenu identique à sauvegarder des filtres de duplicate content

Si vous désirez qu'un même contenu sur des pages différentes (ou des sites différents la plupart du temps) ne soit pas considéré comme du duplicate content (on a donc ici à envisager une problématique inverse : on désire que deux pages proposant le même contenu soient indexées de la même façon et que le duplicate content ne soit pas détecté), n'hésitez pas, sur ces pages, à modifier le plus possible ce qui existe « autour du texte » : images, vidéos, liens, encadrés, etc.

Certains changent, lorsque c'est possible, l'ordre des paragraphes (exemples de fiches produits où l'ordre n'est pas essentiel), mais Google a appris également à déceler le duplicate content lorsque l'ordre des informations change, ce n'est donc pas une stratégie gagnante à 100 %.

Changez également vos intitulés d'URL d'un site à l'autre et le plus possible, les footer, header et liens internes (si les contenus similaires se trouvent sur des sites distincts) pour proposer sur chaque site un « environnement » le plus différent possible...

Figure 7-31

Exemple de programme télé sur le site programme-tv.net

Essayez, le plus possible, de modifier les codes HTML proposés aux moteurs :

- Inversez l'ordre des balises `<title>`, des balises meta, etc.
- Ajoutez ou supprimez des balises meta peu importantes (`classification`, etc.).
- Ajoutez ou supprimez des commentaires (même si on sait que leur contenu n'est pas lu par les moteurs, ils peuvent changer les séquences linéaires de lecture du code).
- Codez vos pages en UTF-8 ou en ISO-8859-1 selon le cas...
- Proposez des attributs `alt` avec des contenus différents pour chaque image.

Figure 7-32

Le même programme sur tele-loisirs.fr. Comment faire pour que l'un ne « phagocyte » pas l'autre ?

- La structure des pages (en tableaux ou *via* des CSS) peut également être totalement différente d'un site à l'autre.
- etc.

Vous pouvez enfin ajouter du contenu à l'une ou l'autre page, sur une base commune (des encadrés différents, par exemple, ou des infos connexes sur le même sujet). Ce sera autant de travail qui différenciera chaque page... Variez également, si cela est techniquement possible pour vous, les hébergeurs et les adresses IP de vos serveurs d'un site à l'autre.

Attention cependant de faire en sorte que l'une des pages ne pointe pas vers l'autre, ce qui signifierait que l'article qui reçoit le backlink est considéré comme le contenu canonique pour Google (voir la première partie de ce chapitre).

Il existe en fait des dizaines de façons pour arriver à différencier le plus fortement possible vos contenus et vos pages. À vous de voir, en fonction de vos possibilités, notamment techniques, lesquelles vous pouvez mettre en œuvre...

Pour résumer, il est nécessaire de jouer sur plusieurs niveaux de réflexion pour lutter au mieux contre l'hypothèse du duplicate content dans le cas de contenus similaires sur des sites différents :

- Contenus éditoriaux différents : varier le pourcentage entre la partie « texte éditorial » proprement dite et les informations connexes (articles similaires, suggestions, photos, vidéos, fonctions communautaires) d'un site à l'autre.
- Structure des pages différentes : code HTML, emplacement des blocs de code, URL, etc.
- Chartes graphiques différentes : par exemple, proposer les mêmes liens, mais différemment présentés, etc.

Duplicate content : l'évangile selon saint Google...

Enfin, pour terminer ce chapitre, n'hésitez pas à lire ce que dit Google dans son centre d'aide pour webmasters (*http://www.google.fr/support/webmasters/bin/answer.py?answer=66359&query=dupliqu%C3%A9&topic=&type=*) au sujet du duplicate content. En voici les extraits qui nous ont semblé les plus importants et intéressants :

« Contenu en double :

Par contenu en double, on entend généralement des blocs de contenu importants, appartenant à un même domaine ou répartis sur plusieurs domaines, qui sont identiques ou sensiblement similaires. À l'origine, la plupart de ces contenus ne sont pas malveillants. Exemples de contenu non malveillant :

- forums de discussion pouvant générer à la fois des pages normales et des pages « raccourcies » associées aux mobiles ;
- articles en vente affichés ou liés *via* plusieurs URL distinctes ;
- versions imprimables uniquement de pages web.

Dans certains cas, cependant, le contenu est délibérément dupliqué entre les domaines afin de manipuler le classement du site par les moteurs de recherche ou d'augmenter le trafic. Ce type de pratique trompeuse peut affecter négativement la navigation de l'internaute qui voit quasiment le même contenu se répéter dans un ensemble de résultats de recherche.

Google s'efforce d'indexer et d'afficher des pages contenant des informations distinctes. [...]

Les mesures suivantes vous permettent de résoudre les problèmes de contenu en double de manière proactive et de vous assurer que les visiteurs accèdent au contenu que vous souhaitez leur présenter.

- Bloquez l'indexation des pages : plutôt que de laisser les algorithmes Google déterminer la « meilleure » version d'un document, vous pouvez nous indiquer votre version favorite. Par exemple, si vous ne souhaitez pas indexer les versions imprimables des articles de votre site, désactivez ces répertoires ou utilisez des expressions littérales dans votre fichier `robots.txt`.
- Utilisez des redirections 301 : si vous avez restructuré votre site, utilisez des redirections 301 (`RedirectPermanent`) dans votre fichier `.htaccess` pour rediriger efficacement les internautes, Googlebot et autres robots d'exploration. [...]
- Soyez cohérent : assurez la cohérence dans vos liens internes. Par exemple, n'établissez pas de lien vers *http://www.exemple.fr/page/*, *http://www.exemple.fr/page* et *http://www.exemple.fr/page/index.htm*.
- Utilisez des domaines de premier niveau : pour nous aider à présenter la version la plus appropriée d'un document, utilisez dans la mesure du possible des domaines de premier niveau pour gérer du contenu propre à un pays. Nous sommes plus enclins à

penser que le site *www.exemple.de* contient du contenu destiné à l'Allemagne, que *www.exemple.com/de* ou *de.exemple.com*.

- Diffusez du contenu avec prudence : si vous diffusez votre contenu sur d'autres sites, Google affichera systématiquement la version jugée la plus appropriée pour les internautes dans chaque recherche donnée, qui pourra être ou non celle que vous préférez. Cependant, il est utile de s'assurer que chaque site sur lequel votre contenu est diffusé inclut un lien renvoyant vers votre article original. Vous pouvez également demander à ceux qui utilisent votre contenu diffusé de bloquer la version sur leur site avec leur fichier `robots.txt`.
- Utilisez nos outils pour les webmasters afin de nous indiquer votre méthode d'indexation de site favorite : vous pouvez indiquer votre domaine favori à Google (par exemple, *www.exemple.fr* ou *http://exemple.fr*).
- Limitez les répétitions : par exemple, au lieu d'inclure une longue mention de copyright au bas de chaque page, insérez un récapitulatif très bref, puis établissez un lien vers une page plus détaillée.
- Évitez la publication de pages incomplètes : les internautes n'appréciant pas les pages « vides », évitez dans la mesure du possible les espaces réservés. [...]
- Apprenez à maîtriser votre système de gestion de contenu : vérifiez que vous maîtrisez l'affichage du contenu de votre site web. Les blogs, forums et systèmes associés affichent souvent le même contenu dans des formats divers. [...]
- Limitez les contenus similaires : si de nombreuses pages de votre site sont similaires, développez chacune d'entre elles afin de les rendre uniques ou consolidez-les en une seule. [...] »

On notera également, pour terminer, un article, sur le blog pour webmasters de Google (intitulé *Demystifying the « Duplicate Content Penalty »* et disponible à l'adresse suivante : *http://googlewebmastercentral.blogspot.com/2008/09/demystifying-duplicate-content-penalty.html*) qui explique, avec raison, que le duplicate content ne génère pas de « pénalités » au sens où on l'entend souvent sur le Web, même si cela peut être « pénalisant » (la nuance est importante...) pour votre visibilité sur les moteurs.

Si avec tout ça, vous vous laissez encore happer par les pièges du duplicate content sur les moteurs de recherche, c'est à désespérer de tout...

Le plan du site et les pages de contenu : deux armes pour le référencement

On oublie parfois, lors de la création d'un site web, deux stratégies importantes pour pallier les contraintes techniques qui peuvent freiner un référencement.

- La page « Plan du site » : on peut presque dire qu'elle est essentielle pour le référencement. Elle donne, sur un même document, des liens vers toutes les pages principales de votre site. Du caviar pour les robots ! Sachez que pour obtenir une bonne indexation de

votre site en termes quantitatifs, **chaque page doit être accessible aux robots depuis votre page d'accueil en trois clics au plus**. Donc dans ce cas :

– un clic pour la page d'accueil vers le plan du site ;

– un deuxième clic depuis le plan du site vers la page elle-même.

En deux clics, le tour est joué ! Vous pouvez éventuellement proposer une page intermédiaire si votre site contient énormément de documents : un premier « plan » propose les grands sommaires, les grandes zones de votre site. Les liens proposés pointent alors vers des sous-plans (un par zone) qui affichent des liens vers les pages finales. Ici, en trois clics le problème est réglé. C'est une garantie de bonne indexation quantitative par les moteurs.

- Les pages de contenu : de plus en plus utilisée, cette stratégie consiste à créer de vraies pages de contenu, mais plutôt optimisées pour les moteurs de recherche (bon titre, bon indice de densité, mots-clés mis en exergue, etc.). Elles peuvent ne pas s'insérer dans la navigation normale du site (les menus de navigation) mais, par exemple, être accessibles uniquement par l'intermédiaire de liens dans le plan du site. Pas de JavaScript, d'identifiants de session et autres obstacles techniques : du texte et rien que du texte, optimisé bien sûr. Encore une fois, ne cachez rien : ni le contenu lui-même ni les liens qui y mènent. Toute donnée cachée au moteur est dangereuse, ne l'oubliez pas !

Ne pas oublier la réputation et le Sitemap !

Une stratégie souvent oubliée également lorsqu'un site présente des problèmes d'indexation et de positionnement consiste à utiliser de façon plus accrue la réputation (voir chapitre 5) et les contenus textuels des liens qui pointent vers lui. En d'autres termes, si votre site est 100 % en Flash, vous pouvez obtenir des résultats très intéressants et à fort PageRank si des pages pointent vers lui, proposant des liens textuels contenant des mots importants pour votre activité… Quelques liens à contenu textuel optimisé depuis des pages à fort PageRank (supérieur ou égal à 6) peuvent énormément aider un site pourtant techniquement rédhibitoire au départ.

Enfin, n'oubliez pas les Sitemaps proposés par la plupart des moteurs de recherche majeurs et qui vous permettent de fournir au moteur un fichier contenant la liste des URL de votre site. Cela ne remplace pas l'indexation naturelle de vos pages par les robots, mais c'est très complémentaire. Et les premiers résultats que l'on voit apparaître semblent appréciables. Nous y reviendrons au chapitre suivant.

Cas spécifique des sites multilingues

Si vous avez décidé de créer un site web multilingue, vous devrez faire attention à plusieurs points afin d'améliorer et d'optimiser sa prise en compte par les moteurs de recherche. Nous allons les passer en revue, en commençant par la meilleure solution pour terminer avec la moins bonne.

Solution 1 – Un nom de domaine par langue

Il s'agit ici de l'option idéale : chaque site dispose de sa langue et de son nom de domaine qui lui est propre. Exemple :

- *www.votresite.com* : anglais (cible : États-Unis) ;
- *www.votresite.co.uk* : anglais (cible : Grande-Bretagne) ;
- *www.votresite.fr* : français (cible : France) ;
- *www.votresite.de* : allemand (cible : Allemagne).

Aucun des sites n'interfère avec les autres, chacun dispose de sa propre langue. Ainsi, rien ne peut gêner les robots des moteurs : c'est parfait. L'inconvénient majeur est que :

- Vous aurez à gérer de nombreux noms de domaine dans plusieurs pays différents.
- L'achat de ces noms de domaine peut s'avérer complexe dans certains pays, notamment dans les contrées qui demandent à disposer d'une structure professionnelle sur place.

Ceci dit, si vous désirez jouir d'une situation optimale par rapport aux moteurs de recherche, il faudra en passer par là.

Solution 2 – Un sous-domaine par langue

Autre solution intéressante et qui ne demande l'achat que d'un seul nom de domaine : la création d'un sous-domaine par langue. Exemple :

- *www.votresite.com* : anglais (cible : États-Unis) ;
- *uk.votresite.com* : anglais (cible : Grande-Bretagne) ;
- *fr.votresite.com* : français (cible : France) ;
- *de.votresite.com* : allemand (cible : Allemagne).

Avec un seul nom de domaine (*votresite.com*), vous créez autant de sous-domaines que vous le désirez, presque instantanément, un par pays et/ou langue. Chaque sous-domaine étant considéré comme un site web à part entière par les moteurs, cette situation est quasi identique au point vu précédemment. Seul petit inconvénient : l'utilisation et la compréhension des sous-domaines ne sont pas très répandues parmi les internautes, notamment dans le grand public, plus habitué à des adresses commençant par *www.*

Solution 3 – Un répertoire par langue

Troisième solution : créer sur votre site à nom de domaine unique un répertoire par langue. Exemple :

- *www.votresite.com/* : anglais (cible : États-Unis) ;
- *www.votresite.com/uk/* : anglais (cible : Grande-Bretagne) ;

- *wwwvotresite.com/fr/* : français (cible : France) ;
- *www.votresite.com/de/* : allemand (cible : Allemagne).

Le principal inconvénient de cette méthode est que dans ce cas, tous les sites ci-dessus sont considérés comme faisant partie d'un seul site (*www.votresite.com*) par les moteurs. Votre visibilité pourrait donc décroître fortement si les moteurs font du clustering (voir les parties du chapitre 4 traitant des noms de domaine et des sous-domaines).

Enfin, si vous optez pour cette solution, nous vous conseillons d'afficher directement la page en anglais, en affichant des liens, sous forme de menus déroulants ou de drapeaux, dans une zone de navigation comme le montre l'exemple de la figure 7-33.

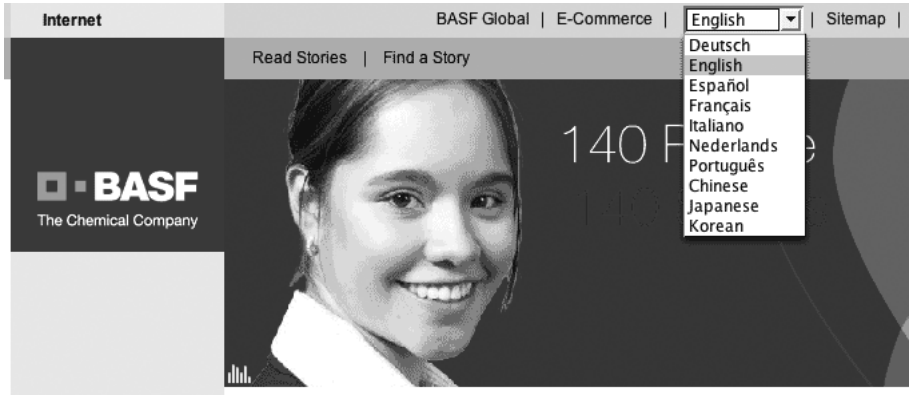


Figure 7-33

Exemple du site BASF (http://www.basf.com/) : la page d'accueil par défaut est en anglais et un choix vers les autres versions linguistiques est proposé sous la forme d'un menu déroulant.

Ce choix est intéressant car il permet d'afficher tout de suite une information utile à l'internaute qui, s'il désire l'obtenir dans une autre langue, utilisera le moyen proposé pour cela. Cette option est préférable à une « page de drapeaux » sans réel contenu comme celle présentée en figure 7-34.

Dans ce cas, la page n'est pas optimisée car :

- L'internaute doit cliquer pour obtenir une première information utile, cette page d'accueil n'étant absolument pas informative. Cela retarde donc son accès à l'information. Rappelons-nous que les pages importantes de votre site doivent se trouver à trois clics au maximum de la page d'accueil. Rajouter une page drapeaux « grillera » un clic...
- Le moteur aura « à se mettre sous la dent » une page d'accueil sans texte ou presque, ou en tout cas, sans texte descriptif de votre activité. La page d'accueil ayant la plupart du temps un PageRank élevé, vous sabotez votre référencement en n'y indiquant aucun contenu, aucun mot-clé important ce qui est fort dommage.

Figure 7-34

L'accès « par drapeaux » n'est pas conseillé. Ici, le site Roland-Garros (http://www.rolandgarros.com)

Solution 4 – Pages multilingues

Nous en avons déjà parlé au cours du chapitre précédent : évitez au maximum les pages affichant du contenu dans plusieurs langues distinctes. Les moteurs n'aiment pas cela car ils n'arrivent pas à distinguer la langue unique dans laquelle le document est écrit. Résultat : si votre page contient 50 % de français, elle risque de ne pas apparaître dans une recherche avec l'option « Pages francophones ». À éviter donc !

Conclusion

Comme on a pu le voir au cours de ce chapitre, il existe de nombreux critères freinant le référencement d'un site web par les moteurs de recherche. Pourtant, à la fin 2009, les critères réellement bloquants sont rares car il existe quasiment toujours des solutions aux problèmes éventuellement rencontrés. Encore faut-il, bien sûr, les mettre en œuvre… ce qui peut représenter du temps et/ou de l'argent si ces procédures sont sous-traitées.

Il s'agit encore ici d'un argument pour la prise en compte du référencement à la base, au départ du projet de création ou de refonte de site web : plus les options technologiques seront compatibles avec les moteurs de recherche, moins le travail à faire par la suite sera important… et onéreux.

8

Référencement, indexation et pénalités

Dans les chapitres précédents, vous avez pu appréhender la meilleure manière d'optimiser vos pages web afin qu'elles soient réactives par rapport aux critères de pertinence majeurs des moteurs de recherche. Mais pour qu'elles soient bien placées – bien positionnées – dans les résultats de ces moteurs, encore faut-il qu'elles soient présentes dans les bases de données (c'est-à-dire les index) de ces outils. Comment optimiser cette présence, en termes de rapidité et de nombre de pages indexées ? C'est ce que nous allons voir dans ce chapitre.

Comment « soumettre » son site aux moteurs de recherche ?

Voyons donc pour commencer les différentes voies qui vous sont proposées pour faire en sorte que votre site web soit rapidement présent dans les index des moteurs de recherche. Les objectifs de cette procédure seront les suivants :

- Indexer le maximum de pages. Votre site propose un certain nombre de pages web, la situation idéale sera celle où 100 % de ces documents seront indexés par les moteurs de recherche.
- Indexer le plus rapidement possible – si possible en 24 h, voire moins – un nouveau site ou les nouvelles pages, les nouveaux contenus d'un site déjà référencé.
- Faire en sorte que les robots des moteurs reviennent le plus souvent possible sur les pages web afin de prendre en compte les nouvelles versions mises en ligne.

Nous sommes ici dans le pur domaine du référencement : faire en sorte que vos pages – si possible optimisées au préalable (voir chapitres 4 et 5) – soient prises en compte par les moteurs de recherche afin d'obtenir, dans un deuxième temps, le meilleur positionnement qui soit.

Comme nous allons le voir, il existe plusieurs moyens de référencer son site. Certains sont très efficaces, d'autres moins. Ce sera à vous de faire votre choix en fonction de vos possibilités, de vos besoins et de vos attentes.

Le formulaire de soumission proposé par le moteur

Il s'agit ici de la voie officielle. Difficile de ne pas la signaler puisque la plupart des moteurs proposent un formulaire de soumission de site web vous permettant de leur signaler l'existence de votre source d'informations. Voici les adresses de ces formulaires pour les moteurs majeurs :

- Google – *http://www.google.fr/addurl/?hl=fr&continue=/addurl*
- Yahoo! (nécessite un compte Yahoo!) – *http://siteexplorer.search.yahoo.com/submit*
- Bing – *http://www.bing.com/docs/submit.aspx?FORM=WSDD2*
- Exalead – *http://www.exalead.fr/search/web/submit/*

Il faut noter qu'en 2009, seul le moteur Ask.com ne proposait pas ce type de formulaire parmi les outils les plus importants.

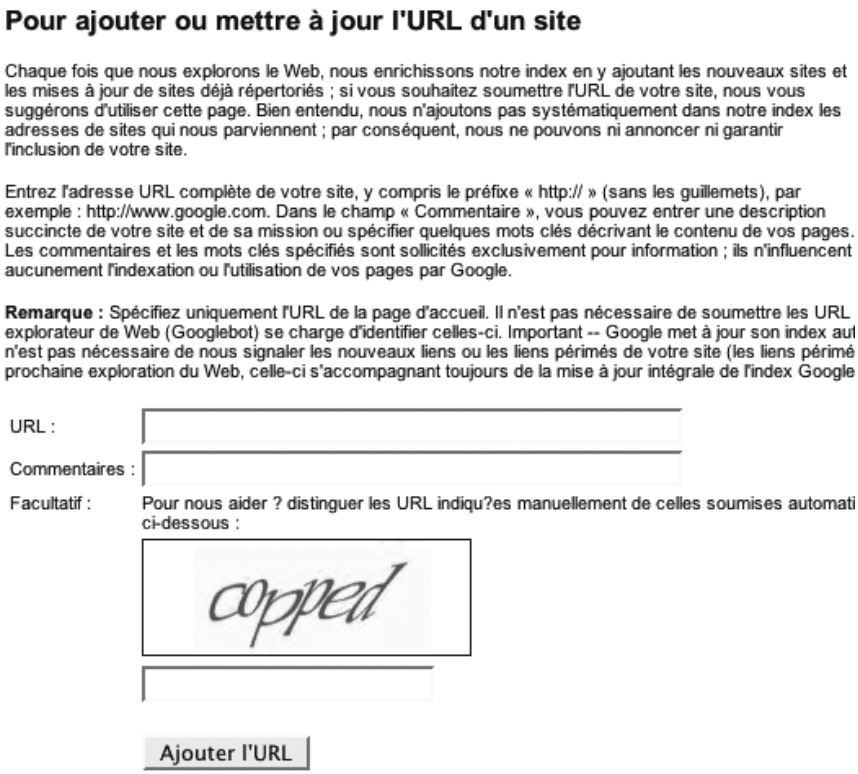
La procédure est extrêmement simple : vous remplissez le formulaire proposé en indiquant l'adresse de la page d'accueil de votre site et parfois quelques informations connexes (commentaires lettres et chiffres codés pour éviter que des robots n'effectuent cette opération, etc.). Vous envoyez le tout et c'est terminé. La figure 8-1 montre ce processus pour le moteur Google.

Quelques jours ou quelques semaines plus tard, les robots du moteur de recherche vont venir visiter la page se trouvant à l'adresse indiquée, prendre en compte dans un premier temps le document en question, puis les autres pages internes, dans un délai plus long, en suivant les liens affichés. Théoriquement, tout cela est donc parfait.

En pratique, cette voie n'est pas la plus efficace ni la plus rapide. Il existe de nombreux exemples de sites web jamais pris en compte bien qu'ils aient été plusieurs fois soumis par l'intermédiaire de ces formulaires. Les procédures officielles d'ajout de site des moteurs de recherche ne sont donc pas à privilégier dans le cadre d'une stratégie de référencement. Ceci dit, elles ont le mérite d'exister (ce qui est loin d'être négligeable) et elles peuvent éventuellement être prises en compte si d'autres voies avaient échoué pour signaler votre site aux moteurs… Mais il existe bien mieux, comme nous allons le voir tout de suite !

Figure 8-1

Formulaire de soumission de site de Google



Soumettez uniquement la page d'accueil de votre site

Il n'est pas nécessaire de soumettre les adresses de toutes vos pages au travers de ces formulaires de soumission. Seule la page d'accueil suffit. Le moteur trouvera ensuite vos autres pages en suivant les liens internes de votre site.

Seule la soumission manuelle est efficace

En règle générale, nous vous déconseillons l'emploi de logiciel ou de site web spécialisés dans la soumission automatique de sites web aux moteurs de recherche. L'emploi de tels outils peut même se révéler dangereux pour votre référencement, les moteurs n'appréciant que modérément ce type de méthode. Ils sont d'ailleurs tombés en désuétude avec le temps (les outils, pas les moteurs...). Suivez plutôt les conseils de ce chapitre et tout se passera bien.

Le lien depuis une page populaire

La possibilité que nous allons vous décrire ici est celle que vous allez devoir privilégier pour référencer votre site car c'est de loin la plus rapide et la plus fiable. Cette technique

est quasi infaillible, et nous l'utilisons fréquemment pour indexer en 24 h seulement – voire moins – la plupart des nouveaux sites que nous créons depuis plusieurs années. Jusqu'à maintenant, elle a toujours parfaitement fonctionné. Il n'y a donc aucune raison pour que cela ne soit pas le cas avec les vôtres…

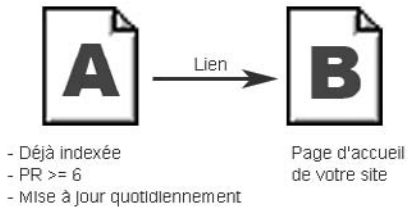
Comme pour toute recette, celle-ci nécessite des ingrédients. Imaginons que B soit le nom de la page d'accueil du site que vous désirez référencer. Pour ce faire, vous aurez besoin d'une autre page, que nous nommerons A, et qui répondra aux caractéristiques suivantes :

- La page A doit déjà se trouver dans les index des moteurs. Elle est donc déjà référencée.
- La page A peut se trouver sur votre site ou sur un autre sans différence (si B est la page d'accueil d'un nouveau site, il y a fort à parier que A appartienne à un autre site ; si B est une nouvelle page sur un site existant, A peut être la page d'accueil de ce même site).
- La page A doit être mise à jour quotidiennement.
- La page A doit être populaire et bénéficier d'un PageRank d'au moins 6, sachant qu'une valeur de 7 sera préférable. Bien sûr, un PageRank supérieur sera encore meilleur. Au pire, une valeur de 5 conviendra quand même.

Ensuite, il ne vous reste plus qu'à faire en sorte que la page A établisse un lien vers votre page B et cette dernière sera indexée dans les 24 h ou, au pire, 48 h par tous les moteurs majeurs. Ce mécanisme est illustré en figure 8-2.

Figure 8-2

Mécanisme d'indexation par les liens



Pourquoi cela fonctionne-t-il ? Explications…

- La page A est déjà indexée. Elle est donc connue des moteurs.
- La page A est mise à jour quotidiennement. Les robots des moteurs ont donc calqué leur délai d'indexation sur ceux de la page (voir chapitre 2) et viennent donc, logiquement, tous les jours au moins « aspirer » sa nouvelle version.
- Que se passe-t-il lorsque vous rajoutez le lien de A vers B ? Les robots le détectent immédiatement et se disent qu'un nouveau lien émanant d'une page populaire pointe certainement vers un site intéressant. Ils vont donc suivre ce lien et indexer quasi immédiatement la page B. Le tour est joué…

Essayez cette procédure, vous verrez qu'elle fonctionne parfaitement bien. Bien sûr, il vous faudra trouver, pour cela, une page A qui réponde aux critères énoncés ci-dessus.

Pas si facile, mais en cherchant bien, c'est tout à fait possible… Et le jeu en vaut vraiment la chandelle !

Les fichiers Sitemaps

Le format et le protocole Sitemap (*http://www.sitemaps.org/*) sont assez récents puisqu'ils ont été initiés par Google en juin 2005 (*http://actu.abondance.com/2005-23/google-sitemaps.php*). Il s'agit d'une solution permettant de fournir aux robots des moteurs de recherche (Google, Yahoo!, Bing, Exalead et Ask prenaient en compte ce format en 2009) un plan de votre site au format XML. Ces robots peuvent alors identifier et aller chercher toutes les pages disponibles, selon les indications fournies dans le fichier.

Figure 8-3

Le logo de l'offre Sitemaps de Google

Dans un premier temps, il est nécessaire de bien comprendre comment fonctionne le système, assez complet et parfois méconnu dans ses fonctionnalités avancées.

Le concept des Sitemaps

Le concept de ce format est extrêmement simple : vous créez un fichier XML qui contient la liste des pages de votre site, ainsi que certaines informations les concernant (fréquence de mise à jour, priorité de crawl, etc.). Vous chargez (*upload*) ce fichier sur votre serveur. Vous signalez au moteur sa présence grâce à une interface d'administration mise à votre disposition par l'outil ou à un fichier `robots.txt` adéquat (voir chapitre 9). Les robots de ce dernier viennent alors le lire et tiennent compte des données qui y sont proposées pour mieux indexer votre site, de façon plus approfondie et plus exhaustive. Simple, non ? Mais encore faut-il que votre fichier soit bien créé, bien soumis et bien placé sur votre site.

Cependant, notez bien que :

- L'utilisation d'un Sitemap n'est en rien une garantie que le moteur indexera toutes les pages qui y sont décrites. Il reste maître de la façon dont il indexe les sites. Mais l'utilisation d'un tel fichier facilite, logiquement, ce processus.
- De même, le fichier Sitemap n'est en rien une garantie que votre site sera mieux positionné. Cet outil n'est qu'un outil d'indexation et non pas de positionnement (*ranking*).
- Enfin, l'utilisation d'un Sitemap ne remplace pas le *crawling* classique de votre site par les robots, suivant les liens des pages web de façon traditionnelle. Les deux méthodes restent tout à fait complémentaires.

Formats du fichier à fournir à l'applicatif

Le protocole Sitemap reconnaît un certain nombre de formats.

- Les fichiers au format OAI-PMH (*Open Archives Initiative Protocol for Metadata Harvesting*). Ce format est cependant inutilisable pour les sites optimisés pour les mobiles. Peu fréquent, ce format (*http://www.openarchives.org/OAI/openarchivesprotocol.html*) est proposé uniquement pour les sites utilisant déjà ce standard. Nous n'en parlerons pas ici.
- Les fichiers de syndications RSS et Atom, ce qui peut-être très intéressant si votre site utilise déjà ce type de format et s'il propose des fils RSS à ses visiteurs. Il est tout à fait possible de signaler au moteur vos fichiers de syndication par ce biais, qui ne sera pas exhaustif, loin de là, mais qui présentera l'avantage d'être très rapide en indiquant aux moteurs les derniers articles parus sur votre site.
- Les fichiers texte (par exemple, *www.votresite.com/sitemap.txt*) contenant une adresse de page (URL) par ligne. Le fichier ne doit pas contenir plus de 50 000 lignes mais il est possible de créer plusieurs fichiers.

Le format texte pour une simple liste d'URL

Nous recommandons l'utilisation du fichier texte si vous désirez uniquement fournir aux moteurs de recherche une liste d'URL sans indiquer d'informations connexes (date de dernière modification, priorité d'indexation, fréquence de mise à jour) pour ces pages.

Les trois solutions ci-dessus sont intéressantes mais elles souffrent toutes d'un handicap majeur : elles permettent uniquement de donner une liste d'adresses, sans informations complémentaires à leur sujet (date de dernière modification, fréquence de mise à jour, etc.). C'est pour cela qu'il sera plus intéressant (mais également plus long et fastidieux) d'utiliser le format XML, dont il est important de signaler qu'il est fourni sous la coupe d'une licence Creative Commons (*http://creativecommons.org/licenses/by-sa/2.0/*), ce qui signifie que d'autres moteurs peuvent l'utiliser (la plupart des moteurs majeurs se sont ralliés au projet depuis sa création par Google en 2005).

Format des fichiers Sitemaps

Le format « Sitemap Protocol » décrit un fichier XML qui va fournir des indications pour chaque page de votre site.

Le fichier créé sera de la forme :

```
<?xml version="1.0" encoding="UTF-8"?>
<urlset xmlns="http://www.google.com/schemas/Sitemap/0.84">
  <url>
    <loc>url</loc>
    <lastmod>date</lastmod>
    <changefreq>fréquence de mise à jour</changefreq>
    <priority>priorité</priority>
  </url>
</urlset>
```

Ce fichier contiendra les indications suivantes :

- `urlset` (obligatoire) commence et termine (`/urlset`) le fichier en question ;
- `url` (obligatoire) décrit chaque page et contient les champs suivants :
  - `loc` représente l'adresse de la page (*http://www.votresite.com/page1.html*). Ce champ commence par *http://* et se termine éventuellement par un /. Ce champ ne peut contenir plus de 2 048 caractères.

Un champ qui doit être très précis

Le champ `loc` est important et assez strict dans son utilisation. Attention donc à bien suivre les indications suivantes :

– chaque URL doit être indiquée de façon absolue (pas d'affichage relatif du type *../directory/page.html*), donc toujours commencer par la mention *http://*.

– chaque page indiquée dans le fichier doit être située dans le répertoire où se trouve le fichier Sitemap ou dans un répertoire de niveau inférieur.

Exemple : vous créez le fichier *http://www.votresite.com/produits/Sitemap.xml*

Ce fichier peut décrire les pages suivantes :

– *http://www.votresite.com/produits/index.html*

– *http://www.votresite.com/produits/gamme.html*

– *http://www.votresite.com/produits/electricite/rupteur.html*

Mais il ne pourra pas décrire les pages suivantes :

– *http://www.votresite.com/contact.html*

– *http://votresite.com/produits/contact.html*

– *https://www.votresite.com/produits/index.html* (notez le *https* pour un accès sécurisé)

– *http://www.votresite.com/clients/reference.html.*

Ces pages seront refusées par le moteur lors de la lecture du fichier.

Pour cette raison, l'emplacement le plus logique pour un fichier Sitemap sera le niveau le plus haut de l'arborescence (*http://www.votresite.com/Sitemap.xml*). Ceci dit, rien ne vous empêche :

– de mettre le fichier Sitemap dans un autre répertoire (en tenant compte, dans ce cas, des restrictions évoquées ci-dessus) ;

– de créer plusieurs fichiers Sitemaps pour un même site.

  - `lastmod` est la date de dernière modification du fichier. Cette date doit répondre au format ISO 8601 (*http://www.w3.org/TR/NOTE-datetime*), le plus souvent sous la forme YYYY-MM-DD soit 2009-09-15 pour le 15 septembre 2009.
  - `changefreq` représente la fréquence de mise à jour de la page, à choisir parmi les possibilités suivantes : `always`, `hourly`, `daily`, `weekly`, `monthly`, `yearly` et `never`. Bien entendu, dans ce cas, il faudra faire des choix en optant pour la fréquence la plus vraisemblable si celle-ci n'est pas constante.

Fournissez des indications logiques et homogènes

Nous vous conseillons de ne pas tricher sur ce champ. Rien ne servira d'indiquer `hourly` pour toutes les pages de votre site, si la majorité de vos documents n'est jamais mise à jour. Le moteur a appris à reconnaître, par d'autres voies, la fréquence de mise à jour des documents qu'il indexe. Il semble évident qu'il n'appréciera que modérément si les données que vous lui fournissez sont dépassées par rapport à ses propres constatations sur votre site. Cela peut même vous desservir. Soyez donc le plus loyal possible sur cette indication, vous ne vous en porterez que mieux (et votre site également) à l'avenir.

Par ailleurs, cette indication n'est pas obligatoirement suivie à la lettre par les *crawlers*. Le fait d'avoir indiqué `never` ne signifie pas que les robots du moteur ne viendront qu'une seule et unique fois indexer la page et l'ignoreront par la suite. Ils reviendront quand même, ne serait-ce que pour être sûr qu'elle existe encore…

– `priority` indique l'importance que vous donnez à la page à l'intérieur de votre site. Sa valeur va de 0 à 1 et peut être, bien entendu, décimale (0.5, 0.7, etc.). Attention : pas de virgule, c'est le point qui marquera ici la décimale. Par exemple, la page d'accueil de votre site aura, vraisemblablement, une priorité de 1. Si rien n'est indiqué, la priorité par défaut est fixée à 0.5.

Fournissez des indications logiques et homogènes (suite)

Là encore, jouez le jeu et indiquez des niveaux de priorité réels. Évitez de positionner ce champ à la valeur 1 pour toutes vos pages. Point important également : ce niveau de priorité ne joue aucunement sur le ranking de vos pages. Il s'agit uniquement de données fournies aux robots pour crawler de façon plus ou moins prioritaire vos documents (si ces robots se servent de cette indication, ce qui n'est pas prouvé..).

Notez que le fichier doit obligatoirement être encodé au format UTF-8. Certains caractères doivent donc être encodés différemment, notamment dans les URL où l'esperluette (`&`) doit apparaître ainsi : `&`, etc.

Notez également que les champs `lastmod`, `changefreq` et `priority` sont optionnels.

Enfin, dans le domaine des restrictions, sachez que votre fichier non compressé (vous pouvez également fournir des fichiers compressés en GZip) doit avoir une taille inférieure à 10 Mo et contenir des informations sur 50 000 pages (URL) au maximum, ce qui laisse un peu de marge (d'autant plus que vous pouvez travailler sur plusieurs fichiers comme nous le verrons ci-après).

Exemples de fichiers

Ainsi, un fichier extrêmement simple, minimaliste, mais fonctionnel décrivant un site de trois pages, sera le suivant :

```
<?xml version="1.0" encoding="UTF-8"?>
<urlset xmlns="http://www.google.com/schemas/Sitemap/0.84">
  <url>
    <loc>http://www.votresite.com/</loc>
```

```
  </url>
  <url>
    <loc>http://www.votresite.com/produits.html</loc>
  </url>
  <url>
    <loc>http://www.votresite.com/apropos.html</loc>
  </url>
</urlset>
```

Ce fichier est très simple et n'aura qu'une fonction : signaler au moteur la présence des trois pages. Cependant, il peut paraître plus rapide, dans ce cas, comme nous l'avons indiqué auparavant, d'utiliser le format texte (`.txt`). Ce fichier (par exemple : `Sitemap.txt`) contiendra alors uniquement les lignes suivantes :

```
http://www.votresite.com/
http://www.votresite.com/produits.html
http://www.votresite.com/apropos.html
```

Un fichier plus complet au format XML, contenant plus d'informations, sera le suivant :

```
<?xml version="1.0" encoding="UTF-8"?>
<urlset xmlns="http://www.google.com/schemas/Sitemap/0.84">
  <url>
    <loc>http://www.votresite.com/</loc>
    <lastmod>2009-09-01</lastmod>
    <changefreq>daily</changefreq>
    <priority>1</priority>
  </url>
  <url>
    <loc>http://www.votresite.com/produits.html</loc>
    <lastmod>2009-08-12</lastmod>
    <changefreq>weekly</changefreq>
    <priority>0.8</priority>
  </url>
  <url>
    <loc>http://www.votresite.com/apropos.html</loc>
    <lastmod>2009-09-15</lastmod>
    <changefreq>monthly</changefreq>
    <priority>0.5</priority>
  </url>
</urlset>
```

Chaque page se voit alors décrite avec ses quatre champs spécifiques : URL, date de dernière modification, fréquence de mise à jour et priorité d'indexation. Autant de données que le moteur va utiliser pour mieux les découvrir.

Travail sur plusieurs fichiers

Le protocole Sitemap permet de travailler sur plusieurs fichiers XML. Il faudra dans ce cas créer un nouveau fichier descriptif (fichier mère), nommé `sitemap_index.xml`, qui va contenir les indications sur les sous-fichiers (fichiers filles) utilisés.

Sa structure est similaire à celle d'un fichier fille. Voici le format d'un tel fichier `sitemap_index.xml` :

```
<?xml version="1.0" encoding="UTF-8"?>
<Sitemapindex xmlns="http://www.google.com/schemas/Sitemap/0.84">
  <Sitemap>
    <loc>http://www.votresite.com/Sitemap1.xml</loc>
    <lastmod>2009-09-01</lastmod>
  </Sitemap>
  <Sitemap>
    <loc>http://www.example.com/Sitemap2.xml</loc>
    <lastmod>2009-09-15</lastmod>
  </Sitemap>
</Sitemapindex>
```

L'option `lastmod` indique ici la date de dernière modification du fichier Sitemap, et non pas des pages dont il détient la description.

Cas particulier des sous-domaines

Votre site, comme le site Abondance, utilise peut-être des sous-domaines tels que :

- *www.abondance.com*
- *actu.abondance.com*
- *offre.abondance.com*
- *outils.abondance.com*

Dans ce cas, chaque sous-domaine est considéré par le moteur comme un site à part entière. Le mieux est donc de créer un fichier Sitemap pour chacun des sous-domaines, décrivant les pages que chacun contient. Par exemple :

- *www.abondance.com/Sitemap-top.xml*
- *actu.abondance.com/Sitemap-actu.xml*
- *offre.abondance.com/Sitemap-offre.xml*
- *outils.abondance.com/Sitemap-outils.xml*

Chaque sous-domaine étant indépendant pour les moteurs, un fichier de type `sitemap_index.xml` (voir ci-dessus) n'est donc pas nécessaire dans ce cas. En revanche, vous devrez déclarer chacun de ces fichiers. Nous y reviendrons…

Attention aux intitulés avec ou sans la mention « www »

Attention, n'oubliez pas que le site *www.votresite.com* est considéré par Google comme étant différent de *votresite.com*.

La mise en place d'un fichier Sitemap s'effectue donc en quatre étapes chronologiques.

Étape 1 – Création du fichier

Vous pouvez créer un fichier Sitemap de plusieurs manières :

- En le créant manuellement à l'aide d'un éditeur de texte. Cette solution sera peut-être la plus simple, voire la plus rapide pour un (tout) petit site. Elle deviendra rapidement fastidieuse, voire impossible à gérer pour des moyens et gros sites.
- En utilisant un script, un logiciel ou un site web en ligne, qui effectueront automatiquement cette manipulation, tout en vous donnant la possibilité – ou non – de modifier les résultats créés.

Le choix de l'outil – Google en propose un également – est important car tous ne sont pas équivalents, loin de là, au niveau des fonctionnalités. En effet, selon la taille de votre site, il sera vite fastidieux de créer manuellement un tel fichier au format XML. Très rapidement, l'emploi d'outils automatisés s'avérera indispensable.

Les différentes solutions (scripts, solution en ligne, logiciels) s'adaptent en fait aux besoins des éditeurs de sites web.

- Les solutions en ligne comme Google Site Map Generator and Editor (*http://www.sitemapdoc.com/*) ou XML-Sitemaps (*http://www.xml-sitemaps.com/*), sont parfaites pour des petits sites, de quelques centaines de pages au maximum, ayant un besoin très ponctuel de création de fichier Sitemap.
- Les logiciels comme GSiteCrawler (*http://gsitecrawler.com/*) ou SiteMapBuilder (*http://www.sitemapbuilder.net/*) répondront aux attentes des éditeurs de sites web plus importants, en termes de nombre de pages, plus mouvants (nombreuses pages modifiées chaque jour ou chaque semaine), c'est-à-dire à un besoin plus professionnel de création de tels fichiers.
- Enfin, les scripts seront indispensables pour automatiser la mise à jour du fichier Sitemap (la modification d'une page entraînant automatiquement celle des données la décrivant dans le fichier Sitemap). Ces scripts répondent donc à des besoins très pointus d'intégration d'informations, de façon rapide, fiable et automatisée dans les fichiers Sitemaps. Ils seront en revanche réservés aux programmeurs et développeurs web.

Quoi qu'il en soit, il existe aujourd'hui, quelle que soit la taille de votre site web et vos connaissances techniques, une solution pour créer un fichier Sitemap. N'hésitez pas, les quelques minutes consacrées à cette tâche pourraient grandement aider à une meilleure intégration de votre site dans l'index du moteur.

Sachez également que la plupart des outils et plates-formes de création de sites web ont intégré, depuis la création du standard en 2005, une brique « Sitemap » à leurs options. Souvent, les webmasters ont à leur disposition cette fonction mais ne l'ont pas activée. Vérifiez bien que vous avez ce type de possibilité sous la main (ou la souris...).

Étape 2 – Validation du fichier

Pour être sûr que votre fichier est bien conforme au format XML, vous pouvez utiliser un certain nombre de programmes de validation dont vous trouverez la liste aux adresses suivantes :

- *http://www.w3.org/XML/Schema#Tools*
- *http://www.xml.com/pub/a/2000/12/13/schematools.html*

Ceci dit, au vu de la relative simplicité des fichiers Sitemaps et du fait que vous allez rapidement utiliser un applicatif qui automatise sa création, vous n'aurez rapidement plus à valider vos fichiers puisqu'on peut imaginer que les documents fournis par les logiciels sont propres.

> **Lisez bien le site officiel**
>
> Un site officiel sur le format Sitemap a été mis en ligne à l'adresse *http://www.sitemaps.org/* (en français et en anglais). N hésitez pas à le consulter pour en savoir davantage sur ce point, il regorge d'informations intéressantes.

Par ailleurs, sachez que l'interface d'administration de vos Sitemaps dans les Google Webmaster Tools (*http://www.google.com/webmasters/tools/?hl=fr*), notamment, vous donne aussi la possibilité de tester leur validité et notamment l'exactitude des URL fournies (voir ci-dessous).

Étape 3 – Déclaration du fichier

Placer votre fichier sur votre site ne suffit pas. Il faut signaler aux moteurs qu'il existe pour que ceux-ci viennent le prendre en compte. Deux solutions pour cela :

- Utiliser l'interface d'administration proposée par Google, Yahoo! et Bing. Vous devez disposer d'un compte Google, Yahoo! ou Microsoft, qu'il est possible de créer gratuitement sans aucun problème. Prenons l'exemple de Google (l'interface se trouve dans la zone Webmaster Tools du moteur de recherche : *http://www.google.com/webmasters/tools/?hl=fr*).

  Une fois identifié sur Google, vous avez accès à une interface d'administration très simple, telle que représentée sur la figure 8-4.

Figure 8-4

Interface d'administration des fichiers Sitemaps

Le lien « Envoyer un Sitemap » permet de signaler, sur une page de soumission spécifique, l'adresse de votre (ou de vos) fichier(s).

Devant chaque fichier enregistré, plusieurs liens sont disponibles :

– « Nombre d'URL fournies » indique le nombre d'adresses que Google a trouvées dans le fichier.
– « URL indexées » indique combien, sur les URL soumises, ont été réellement indexées par Google.
– la colonne « Téléchargé » vous indique quand votre fichier a été lu la dernière fois par Google ce qui peut constituer une information très intéressante.

Les autres indications sont assez classiques, nous ne reviendrons pas dessus. Des données statistiques sont également fournies.

- L'autre façon de déclarer votre fichier Sitemap (notamment auprès d'Exalead ou Ask.com qui ne proposent pas de telles interfaces web pour les gérer) est de notifier l'adresse du fichier Sitemap grâce à une fonction nommée `autodiscovery`. Cette dernière permet au moteur de découvrir le fichier Sitemap de façon très simple en indiquant son emplacement physique dans un fichier `robots.txt` en ajoutant simplement cette ligne :

```
Sitemap: <sitemap_location>
```

Exemple :

```
Sitemap: http://www.votresite.com/sitemaps.xml
```

Le robot, lorsqu'il lira le fichier `robots.txt` (voir chapitre 9), aura ainsi immédiatement l'indication de la localisation de votre fichier Sitemap et pourra le lire sans souci. Vous trouverez plus d'informations sur ce sujet à l'adresse suivante : *http://www.sitemaps.org/fr/protocol.php#informing*.

Étape 4 – Mise à jour du fichier

Enfin, il se peut que le contenu de votre site change dans le temps : nouvelles pages qui apparaissent, anciennes qui disparaissent, etc. Vous devez donc mettre à jour en conséquence votre fichier Sitemap, au fur et à mesure des changements. Google viendra l'indexer fréquemment.

Plus d'informations sur le format Sitemap

Voici quelques liens importants au sujet de l'offre Google Sitemaps :
– Documentation (en français)
http://www.google.com/support/webmasters/bin/topic.py?topic=8476
– Le blog de Google dédié à l'offre Sitemaps
http://googlewebmastercentral.blogspot.com/
– Le fichier Sitemap du site Google (on n'est jamais si bien servi...)
http://www.google.com/sitemap.xml
– Le site officiel
http://www.sitemaps.org/

À noter également le changement de nom du système Google Sitemaps qui a été inclus en août 2006 dans une offre plus globale baptisée Google Webmaster Tools (*http://www.google.com/webmasters/tools/?hl=fr*). Ce changement de nom n'affecte pas l'offre en elle-même. Notez également que cette offre évolue quasi quotidiennement et que certaines informations et copies d'écran présentées dans ces pages seront peut-être obsolètes lorsque vous lirez ces lignes. Consultez attentivement le blog dédié à cet applicatif (voir encadré « Plus d'informations sur le format Sitemap ») pour vous tenir au courant des dernières nouveautés proposées !

Des espaces pour webmasters de la part des moteurs de recherche

Bonne nouvelle : avec le temps, les moteurs de recherche se rapprochent des webmasters en général et des référenceurs en particulier, en leur proposant des services et des outils leur permettant de mieux suivre la façon dont leur site est indexé. Google a été le pionnier en la matière avec son Webmaster Tools : *http://www.google.com/webmasters/*.

Yahoo! a rapidement suivi avec l'outil Site Explorer : *http://siteexplorer.search.yahoo.com/*.

Microsoft propose également son propre site, similaire à celui de ses deux concurrents :

http://www.bing.com/webmaster

Ces outils sont des espaces absolument indispensables pour toute personne s'intéressant aux moteurs de recherche et au référencement. N hésitez donc pas à y inscrire votre site au plus vite !

Différents types de Sitemaps

Sachez enfin qu'il n'existe pas que des Sitemaps pour le moteur de recherche web. Le protocole, au fil des années, s'est perfectionné et il est maintenant possible de créer :

- Des Sitemaps pour les vidéos (*http://www.google.com/support/webmasters/bin/answer.py?hl=fr&answer=80472*). Exemple :

```
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9"
        xmlns:video="http://www.google.com/schemas/sitemap-video/1.1">
<url>
  <loc>http://www.example.com/videos/some_video_landing_page.html</loc>
    <video:video>
<video:content_loc>http://www.site.com/video123.flv<video:content_loc>
      <video:player_loc allow_embed="yes">http://www.site.com/videoplayer.swf?video
      ➥=123</video:player_loc>
<video:thumbnail_loc>http://www.example.com/miniatures/123.jpg<video:thumbnail_loc>
      <video:title>Barbecue en été<video:title>
      <video:description>Pour des grillades réussies<video:description>
      <video:rating>4.2</Video:rating>
      <video:view_count>12345</video:view_count>
      <video:publication_date>2007-11-05T19:20:30+08:00.</video:publication_date>
      <video:expiration_date>2009-11-05T19:20:30+08:00.</video:expiration_date>
      <video:tag>steak</video:tag>
      <video:tag>viande</video:tag>
      <video:tag>été</video:tag>
      <video:category>Barbecue</video:category>
```

```
      <video:family_friendly>yes</video:family_friendly>
      <video:expiration_date>2009-11-05T19:20:30+08:00<video:expiration_date>
      <video:duration>600</video:duration>
    </video:video>
</url>
</urlset>
```

- Des Sitemaps pour les sites mobiles (*http://www.google.com/support/webmasters/bin/topic.py?hl=fr&topic=8493*). Exemple :

```
<?xml version="1.0" encoding="UTF-8" ?>
 <urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9"
        xmlns:mobile="http://www.google.com/schemas/sitemap-mobile/1.0">
    <url>
        <loc>http://mobile.example.com/article100.html</loc>
        <mobile:mobile/>
    </url>
</urlset>
```

- Des Sitemaps pour Google News (*http://www.google.com/support/webmasters/bin/topic.py?hl=fr&topic=10078*). Exemple :

```
<?xml version="1.0" encoding="UTF-8"?>
<urlset xmlns="http://www.google.com/schemas/sitemap/0.9"
 xmlns:news="http://www.google.com/schemas/sitemap-news/0.9">
   <url>
      <loc>http://example.com/article123.html</loc>
      <news:news>
         <news:publication_date> 2006-08-14T03:30:00Z </news:publication_date>
         <news:keywords>Business, Mergers, Acquisitions</news:keywords>
      </news:news>
   </url>
</urlset>
```

- Des Sitemaps pour Google Maps (*http://www.google.com/support/webmasters/bin/topic.py?hl=fr&topic=14688*). Exemple :

```
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9"
        xmlns:geo="http://www.google.com/geo/schemas/sitemap/1.0">
<url>
   <loc>http://www.example.com/download?format=kml</loc>
   <geo:geo>
       <geo:format>kml</geo:format>
   </geo:geo>
</url>
<url>
   <loc>http://www.example.com/download?format=georss</loc>
   <geo:geo>
       <geo:format>georss</geo:format>
   </geo:geo>
</url>
</urlset>
```

- D'autres Sitemaps sont également disponibles, quoi que moins utiles au quotidien (Google recherche de code, par exemple). Plus d'informations à cette adresse : *http://www.google.com/support/webmasters/bin/topic.py?topic=20986*. En revanche, il n'existe pas de Sitemaps pour les images, en tout cas au moment où ces lignes sont écrites...

La prise en compte par d'autres robots que ceux crawlant le Web

Les porte-paroles techniques de Google ont également indiqué en 2006 que tous les robots du moteur œuvraient pour découvrir de nouveaux documents. Ainsi, si vos actualités sont prises en compte dans Google News, si vous affichez des liens sponsorisés AdSense sur vos pages, etc., celles-ci auront plus de chances d'être rapidement référencées dans l'index web du moteur car tous ses robots se transmettent les informations et les URL des nouvelles pages. Il en est de même, semble-t-il, pour tous les applicatifs de Google qui utilisent un robot : vidéo, blogs, images, AdSense, etc.

Ne l'oubliez pas, toutes ces voies peuvent être explorées pour voir votre site encore mieux « aspiré » par les moteurs de recherche.

Le référencement payant

Dernière solution : certains moteurs, comme Yahoo!, proposent une offre de référencement payant (ou *paid inclusion*, *trusted feed*, ou encore *XML Feed*). Vous obtenez ainsi une garantie de référencement (mais pas de positionnement) et de mise à jour fréquente dans l'index du moteur.

L'offre de Yahoo!, baptisée Yahoo! Search Submit Express, est décrite à cette adresse : *http://searchmarketing.yahoo.com/srchsb/sse.php?mkt=us*.

Tarifs 2009 : 49 $ la première URL, puis 29 $ jusqu'à la 10e et 10 $ l'URL au-delà. Un tarif est ensuite appliqué au clic (0,15 $ ou 0,30 $ en fonction de la catégorie dans laquelle votre site se trouve).

Rappelons qu'il ne s'agit aucunement ici de garanties de positionnement mais bien uniquement de référencement et d'indexation. Seules les garanties suivantes peuvent vous être apportées par ces offres :

- référencement garanti des pages web – pour lesquelles vous payez – dans l'index du moteur ;
- traitement rapide de la demande ;
- mise à jour fréquente du contenu indexé (en général dans les 48 h).

À vous de voir si le trafic généré par les outils de recherche qui les proposent est assez important pour mériter la dépense engendrée. En pratique, il faut bien avouer que c'est rarement le cas...

Notez enfin que Google ne propose aucune offre de ce type et qu'il est vraisemblable – bien que nous ayons appris à ne jamais dire « jamais » sur le Web – qu'il n'est pas près d'en proposer une…

Enfin, pour continuer dans ce chapitre, il n'est pas négligeable de passer en revue – au risque de faire une légère digression – un certain nombre de pénalités que les moteurs de recherche, et notamment Google, infligent aux sites qu'ils estiment comme fraudeurs, car cela a bien sûr un grande implication dans leur référencement. Et cela aura également une incidence dans l'optimisation du délai d'indexation que nous étudierons par la suite dans ce même chapitre.

Détection des méthodes de spamdexing

Une attitude intéressante à avoir, lorsqu'on désire améliorer son propre référencement, est parfois de se mettre à la place des moteurs de recherche afin de comprendre leur fonctionnement, ce qui permet de plus facilement s'adapter à leurs contraintes. Cela est certainement tout à fait vrai en ce qui concerne la détection du spam (ou spamdexing : fraude sur l'index des moteurs).

En effet, les moteurs de recherche font quotidiennement face à des milliers de webmasters de par le monde qui optimisent, voire « suroptimisent » leurs pages de façon plus ou moins *borderline*. Sachant qu'en fait toute optimisation peut être assimilable à du spam, selon le degré d'optimisation mis en place au regard des règles internes de chaque moteur, parfois bien difficiles à appréhender, il faut bien le dire...

Quelques pistes de réflexion

Voici donc quelques pistes que les moteurs de recherche pourraient explorer, ou explorent déjà, afin de détecter les sites suroptimisés et d'améliorer la qualité de leurs données. Il est important de noter que, bien entendu, tous les moteurs font face sans exception à ce « fléau » du spamdexing.

Pour un moteur de recherche, il n'y a pas pléthore de méthodes permettant de détecter un tricheur. La première chose à vérifier est le cache de l'index. En effet, un site web qui refusera d'afficher son cache pourra d'emblée être considéré comme suspect (ce qui n'est pas synonyme de coupable, notez-le bien) car il ne désire pas que les internautes voient la version de ses pages qu'il a fourni aux moteurs de recherche.

Les différentes méthodes de spam pourraient être regroupées en quatre familles : répétition abusive de mots, texte caché, développement artificiels de liens et *cloaking*. Voici quelques informations à leur sujet.

Détection des abus de densité de mots-clés

Il n'est pas naturel pour un site web de répéter énormément de fois un même mot-clé dans un document, ne serait-ce que par souci rédactionnel. Analyser une page d'accueil est la plupart du temps suffisant pour détecter un site trop optimisé.

La première étape consiste à découper dans un tableau les zones contiguës de texte pour obtenir tous les blocs de texte. En analysant les zones suffisamment denses, y trouver une répétition anormale de mots-clés sera assez aisé et parfois assimilable à du spam. Les moteurs se basent actuellement sur l'indice de densité (nombre d'occurrences du mot dans la

page divisé par le nombre total de mots). Allez consulter, par exemple, le cache des pages Google sur des mots-clés ultras concurrentiels pour en être convaincu... Répéter plus de dix fois une expression clé dans une page « classique » n'est en règle générale pas « naturel ».

Voici un exemple de méthode très simple pour détecter les mots-clés trop denses :

- ne pas prendre en compte la balise meta `keyword` (comme Google le fait déjà) ;
- détecter les titres non naturels (ceux qui se présentent sous la forme d'une succession de mots-clés) ;
- détecter les mots-clés ayant plus de dix occurrences dans la même page ou un indice de densité trop important (par exemple, supérieur à 5 %) ;
- prévoir *in fine* une intervention humaine afin de vérifier la page douteuse et prendre une éventuelle décision de sanction.

Détection des textes cachés

Pour les textes cachés, la tâche est plus compliquée car si ces derniers ne font pas l'objet d'abus de densité, il faudra alors déterminer s'ils sont ou non du spam, ce qui n'est pas aisé de façon automatique.

S'il est facile de détecter du texte ayant la même couleur ou une couleur proche de la couleur de fond d'un site web (méthode assez préhistorique, il faut bien le dire), il n'en est pas de même pour les autres façons de cacher du texte dans une page HTML.

En effet, du contenu peut être caché pour des raisons tout à fait valables notamment dans des *layers*. C'est par exemple le cas dans de nombreux sites qui se prévalent du label « Web 2.0 »... De même, les fonctionnalités W3C d'accessibilité peuvent permettre de cacher du texte de façon tout à fait « loyale ». Par défaut, le fait de cacher du texte n'est donc pas obligatoirement répréhensible. Comment séparer le bon grain de l'ivraie ?

Le moyen ultime, mais très lourd en ressources, serait de se baser sur le texte réellement affiché sur le navigateur en lançant un script automatique qui va récupérer, par l'intermédiaire d'un navigateur comme Firefox, le texte affiché de la page scannée et de le comparer avec le texte « aspiré » par le robot du moteur. S'il diffère beaucoup, une intervention humaine visuelle sera nécessaire pour comprendre pourquoi et vérifier qu'il s'agit ou non de spam.

Le moteur de recherche devra donc disposer de ressources importantes pour vérifier visuellement ces textes cachés en utilisant diverses adresses IP qui ne lui sont pas liées (pour ne pas se faire passer pour un spider « classique »).

Détection des barres de liens et des faux liens

Dans le cas où, comme pour Google, l'algorithme donnerait un poids important à l'analyse des liens internes et externes, la tentation est grande de mettre en place ce type de lien de façon quasi « industrielle ».

Solution relativement simple : détecter des enchaînements consécutifs de plus de trois liens avec un contenu textuel contenant des mots-clés (sans texte de présentation). Ce sera une bonne façon de détecter d'éventuels liens « non naturels ».

Exemples de liens non naturels (très souvent présents dans les footers des pages) :

référencement, achat appartement, immobilier var (ils sont légion sur Internet !)

Exemples de liens naturels :

- Lien au début d'une phrase.

  Exemple : Référencer votre site web en lisant Abondance

- Lien au milieu d'une phrase.

  Exemple : pour aller encore plus loin, contactez un SEO ou un spécialiste du référencement

- Lien sur une phrase entière.

  Exemple : Abondance dévoile un secret bien gardé, SCOOP !

Il convient de ne pas prendre en compte les liens non naturels et bien entendu, de ne pas pénaliser les sites cibles.

Méthodes « avancées » : détection de cloaking

Le cloaking consiste à détecter le type de visiteur qui se connecte sur un site web et à afficher du texte différent pour les visiteurs « humains » et pour les robots des moteurs de recherche. Il peut être diaboliquement efficace mais reste cependant assez facile à détecter pour un moteur. Il est aujourd'hui totalement interdit (blacklistage quasi assuré après détection). On peut être d'accord ou non avec cette vision du cloaking (qui serait pourtant un remède intéressant pour le référencement de sites web « à problèmes » comme le Flash), mais les moteurs de recherche ont aujourd'hui tranché, certainement au vu de certaines pratiques détectées dans le passé : l'utilisation du cloaking est considérée comme du spamdexing, point barre... La communication, notamment de la part de Google et de Yahoo!, est très claire sur ce point depuis de nombreux mois.

Il existe globalement trois types de cloaking (bien qu'on puisse en définir d'autres si on rentre plus dans les détails techniques) : le cloaking par agent (« Googlebot » pour Google, « Slurp » pour Yahoo!...), par adresse IP et par hôte.

Informations complémentaires sur le cloaking

Il existe sur le Web d'innombrables sources d'informations fournissant moultes données de ce type sur les spiders des moteurs de recherche, des plus connus aux plus obscurs... Exemples :

– Quelques adresses IP de Googlebot, le robot de Google :
http://www.robots.darkseoteam.com/adresses-ip-googlebot.php

– Quelques user-agents de Googlebot :
http://www.robots.darkseoteam.com/user-agent-googlebot.php

– Quelques hôtes de Googlebot :
http://www.robots.darkseoteam.com/hotes-googlebot.php

Le cloaking nécessite un développement avec un langage dynamique tel que PHP, ASP, JSP, PERL ou autres. Il consiste pour l'éditeur d'un site, à analyser en temps réel qui se connecte sur le serveur, *via* l'agent, son adresse IP ou son hôte et ainsi détecter un éventuel moteur de recherche et lui afficher dans ce cas un texte optimisé/spammé.

Tous les moteurs de recherche se sont fait « berner » par le passé par ce type d'action. Le premier à réagir a été Google à partir de 2002 en commençant à blacklister de façon manuelle après vérification les sites tricheurs, le plus souvent suite à une dénonciation. De très nombreux sites français ont alors été blacklistés au fur et à mesure des années.

Une méthode simple et infaillible à mettre en place pour les moteurs est donc d'utiliser pour leurs spiders une adresse IP n'étant pas reliée auxdits moteurs et qui va simuler de façon automatique une connexion naturelle (du cloaking inversé en quelque sorte !) à un site web en parallèle d'une connexion émanant « officiellement » du moteur de recherche. Le moteur se fait donc passer un internaute *lambda*. La comparaison des deux résultats indiquera ensuite si une procédure de cloaking a été mise en place par le webmaster.

Un autre type de spam qui n'est pas géré par les moteurs de recherches : les sites satellites n'ayant pas le même contenu de texte (pour éviter le phénomène de duplicate content). Légion sont aujourd'hui les sociétés qui créent de nouveaux sites web en présentant les mêmes activités mais de façons différentes et sur des noms de domaine différents.

Les moteurs pourraient les détecter et pénaliser les sites les plus récents. Il arrive souvent qu'un même site soit représenté quatre fois en page d'accueil alors que c'est la même société et les mêmes activités. Là encore, les moteurs tentent de pénaliser ce type de spamdexing en affinant leurs algorithmes de duplicate content.

Voici des pistes pour les détecter : vérifier les numéros de téléphone (comme Google le gère sur Google Maps), analyser les propriétaires des noms de domaine, etc. Travailler à détecter les webmasters *borderline* doit être passionnant (n'est-ce pas Matt Cutts ?).

Google profile-t-il les référenceurs ?

Selon le site Outspoken Media (*http://outspokenmedia.com/seo/google-profiles-seo-as-criminals/*), Google « profilerait » les sociétés de référencement en fonction du danger potentiel qu'elles représentent. Rien ne dit cependant que ce soit le cas dans la réalité... Lire également à ce sujet :

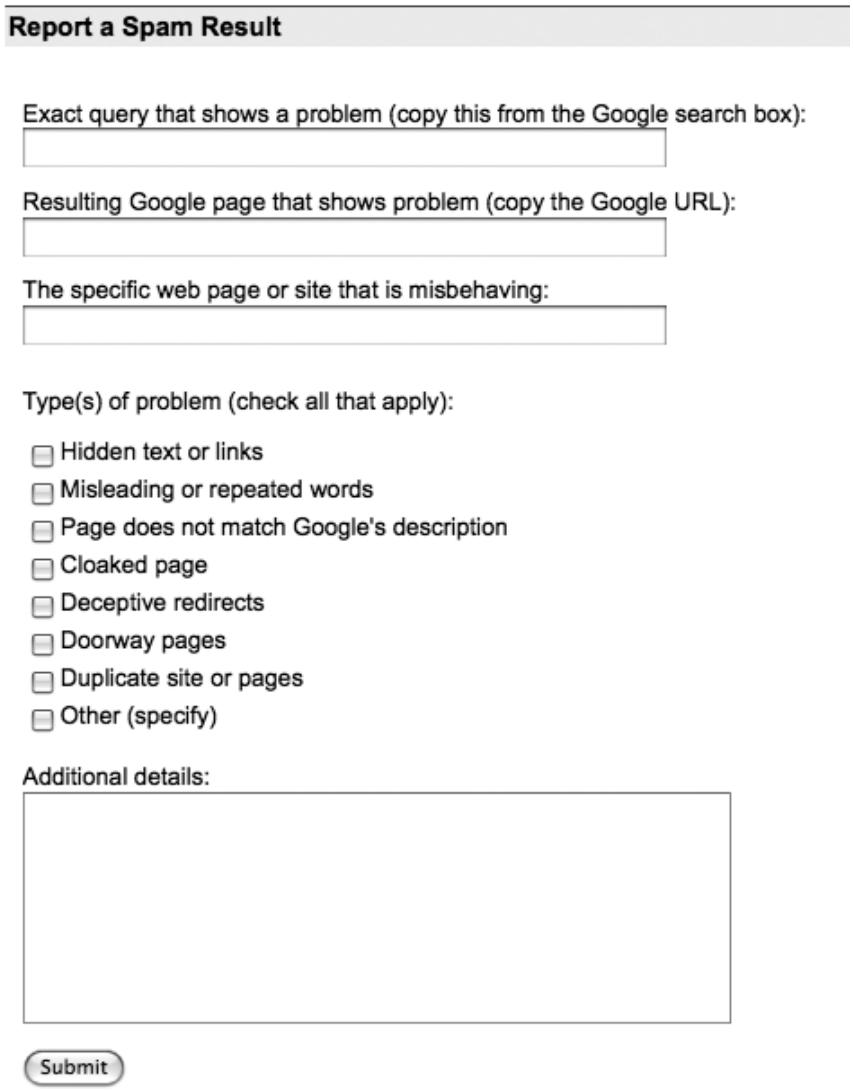
- *http://www.wolf-howl.com/google/google-profiles-seo/*
- *http://seogadget.co.uk/matt-cutts-youa-smx-advanced-day-1-roundup/*
- *http://www.seomoz.org/blog/googles-web-spam-team-deriving-value-from-profiling-seo-operators-of-interest*

La délation

Il ne faut pas s'y tromper : l'une des principales sources de détection du spamdexing par Google est certainement son formulaire de Spam Report, représenté figure 8-5, permettant de dénoncer au moteur toute pratique considérée comme frauduleuse (*http://www.google.com/contact/spamreport.html*).

Figure 8-5

Formulaire de Spam Report sur le site de Google



C'est bien pour cela qu'il est totalement vain de tenter de frauder et de contourner les moteurs de recherche aujourd'hui. Si vous arrivez à être plus malin que les algorithmes des moteurs (et on ne voit pas pourquoi vous n'y arriveriez pas, ce n'est pas si compliqué...), il y aura toujours quelqu'un sur le Web (et notamment vos concurrents) qui s'en apercevront (ce n'est pas beaucoup plus compliqué...) et qui vous dénonceront. Et ce jour-là, vous vous en mordrez les doigts… ou plutôt le clavier (ou la souris, au choix).

Il y a tellement de choses à faire de façon honnête, loyale et pérenne que ce serait quand même dommage de voir votre site pénalisé, parfois à long terme. Réfléchissez-y avant de vous lancer dans des stratégies douteuses...

Les pénalités infligées par Google

Depuis que Google existe, ce moteur de recherche pénalise, comme la plupart de ses confrères, les sites web qui tentent de le « spamindexer », ou en d'autres termes, de contourner ses algorithmes de pertinence pour tenter, grâce à moultes techniques prohibées, de mieux se positionner dans ses pages de résultats.

Au fil du temps, Google a mis en place une panoplie assez complète de pénalités, parfois *soft*, mais également parfois très dures (comme la liste noire) selon la gravité estimée de la faute commise.

Nous allons essayer, dans ce chapitre, de répertorier les différentes pénalités imaginées par les ingénieurs de Google, sachant qu'il est complexe d'en parler avec exactitude puisque Google n'a que très rarement communiqué sur ce point. La plupart des informations proposées ici sont donc issues de tests empiriques, de discussions dans des forums, d'avis d'experts, etc.

Sachez cependant que le contenu ci-après a été transmis à Google et lu par les équipes qui gèrent ces pénalités. Nous n'avons pas eu de retour détaillé de leur part sur son contenu, nous en conclurons donc qu'il n'y a pas d'erreur monumentale dans les indications qui suivent, car, dans ce cas, nos correspondants auraient certainement jugé utile de nous en faire part...

Il n'en reste pas moins vrai qu'il est intéressant de comprendre comment fonctionne la « Spam Brigade » de Google et de bien comprendre que la seule façon de ne jamais avoir à faire à elle est bien de concevoir un site web pour les internautes, optimisé pour les moteurs de recherche, mais sans chercher à aller trop loin. Toute velléité de « suroptimisation » est donc risquée... N'oubliez pas que, même si vos techniques un peu *borderline* échappent aux ingénieurs de Google (qui ont beaucoup de choses à faire et de sites à vérifier), vos concurrents risquent, eux de s'en apercevoir et se feront un plaisir de vous dénoncer auprès du moteur sur son formulaire de Spam Report (*https://www.google.com/webmasters/tools/spamreport?hl=fr*)... Un webmaster averti en vaut toujours deux...

Techniques à ne pas employer

Il est tout d'abord important de comprendre quelles techniques d'optimisation sont à éviter pour ne pas avoir à subir les foudres des moteurs de recherche. En voici quelques-unes :

- pages satellites (alias, fantômes, doorway pages, etc. : pages conçues spécialement pour les moteurs de recherche et contenant une redirection vers le site « réel ») ;

- cloaking (action de fournir des documents différents à un internaute et à un spider par détection automatique de ces derniers) ;
- *keyword stuffing* (répétition non naturelle de mots-clés à l'intérieur d'une page) ;
- contenu textuel et/ou liens cachés au sein du code HTML d'une page à des fins de référencement ;
- ajout dans une page, d'une façon ou d'une autre, de contenu n'ayant pas de rapport direct avec celui qui apparaît de façon visible dans la page ;
- redirection frauduleuse.

En règle générale, n'essayez pas de jouer « aux gendarmes et aux voleurs » avec les moteurs de recherche et relisez bien cette page intitulée « Conseils aux webmasters » sur l'aide en ligne de Google :

http://www.google.com/support/webmasters/bin/answer.py?answer=35769

Le moteur de recherche leader propose également un guide intitulé *Making the Most of Your Content - A Publisher's Guide to the Web* (*http://books.google.com/googlebooks/pdf/webmastertools.pdf*) que vous pouvez lire avec intérêt. À lire également : *Guide de démarrage Google - Optimisation pour les moteurs de recherche*, proposé par le même moteur à l'adresse suivante : *http://www.google.fr/intl/fr/webmasters/docs/search-engine-optimization-starter-guide-fr.pdf*. Indispensable pour obtenir les bases du référencement et apprendre les règles à ne pas transgresser.

Bref, il existe quelques « règles d'or » à suivre pour éviter tout problème de pénalité sur les moteurs de recherche. Les voici :

1. On ne cache rien (ce que l'internaute voit, le moteur le voit et vice versa).
2. Un site web est avant tout fait pour les internautes.
3. Une optimisation de qualité pour les moteurs (balise `<title>`, texte, liens, etc.) fournit rapidement une bonne visibilité à un contenu de qualité.
4. Il est important de veiller à la bonne indexabilité de son site par les spiders (navigation, liens *spider friendly*, fichier Sitemap, etc.).

Si vous suivez ces quelques conseils, vous ne devriez pas avoir de réels soucis avec les équipes de lutte contre le spam qui officient au sein des moteurs. Cependant, il se peut que, même sans le faire exprès (ceci dit, quand on est pénalisé, on sait la plupart du temps pourquoi...), vous passiez, à un moment ou à un autre, de l'autre côté de la frontière. Voici à quelle sauce vous risquez alors d'être puni...

Pénalité numéro 1 – Le mythe de la Sandbox

La « Sandbox » ou « bac à sable » est une pénalité dont on a beaucoup parlé il y a quelques années de cela, notamment à partir de 2004, mais qui semble moins d'actualité aujourd'hui. Il semblerait qu'elle frappe certains sites au moment où Google les découvre et estime que « quelque chose ne va pas » au niveau de l'analyse de ses liens entrants.

Par exemple, Google identifie, lors de sa première visite du site, le fait que ce dernier a déjà obtenu de très nombreux liens entrants (backlinks), ou de nombreux liens entrants depuis des sites distants, chaque fois avec le même « texte d'ancre » (notion de réputation), etc. Google mène d'importants travaux sur la « courbe de vie d'un site ». Il horodate toutes les informations qu'il acquiert et, lorsqu'il découvre une nouvelle source d'informations, il compare notamment sa structure et la situation constatée en termes de liens entrants par rapport à la moyenne de celle d'un site qui vient de sortir. Si les courbes et indices ne concordent pas, il peut y avoir manipulation.

Certains sites qui auraient trop « forcé la dose » dès leur lancement en termes de popularité et de réputation se seraient donc vu, par le passé, « mis en quarantaine dans la Sandbox ». Résultat : aucune possibilité de sortir en bonne position sur une quelconque requête pendant plusieurs semaines (*a priori* de 1 à 12 selon les chiffres le plus souvent constatés). Le site est bien là, il est bien « référencé », mais jamais positionné. Puis, un jour, sa pénalité est terminée, il sort de la Sandbox, sa période de quarantaine est terminée, et il se classe tout de suite mieux... Il semblerait que la Sandbox touche, dans ce cas, toutes les pages d'un même site.

Il semblerait cependant que l'on entende beaucoup moins parler de ce phénomène depuis quelques temps. Google l'a-t-il abandonné ou remplacé par un autre système de pénalité ? Nul ne le sait... Mais il est sûr que ce phénomène a traumatisé plus d'un webmaster qui se voyaient parfois « sandboxé » tous les matins, créant une réelle psychose sur le Web pendant de nombreux mois...

Pour en savoir plus sur la Sandbox...

Voici quelques articles qui devraient vous en dire plus, en français et en anglais sur le phénomène de « Sandbox » :

- *Google et l'effet « Sandbox »* de Olivier Duffez

  http://www.webrankinfo.com/dossiers/strategies-de-liens/sandbox

- *Analyse 2005 de la Sandbox* (traduction française)

  http://www.7-dragons.com/archives/sandbox.php

- *La Google SandBox* de Christophe Da Silva

  http://www.arkantos-consulting.com/articles-referencement/google/20060517-la-google-sandbox.php

Pénalité numéro 2 – Le déclassement

Parfois, pour un site donné mais le plus souvent pour une requête donnée, un site web perd, du jour au lendemain, plusieurs (et parfois de très nombreuses) places dans les résultats du moteur. Ces pénalités sont connues sous le nom de « minus 30 », « minus 60 » ou « Position 6 penalty » pour l'une d'entre elles, apparue en 2008 (*http://blog.abondance.com/2008/01/position-6-penalty-mythe-ou-ralit.html*).

Ainsi, une page web sera « déclassée » pour une requête spécifique, par exemple « britney spears », et disparaîtra en quelques heures des premières pages de résultats pour ces mots-clés alors qu'elle reste toujours bien placée pour des requêtes utilisant la syntaxe « allintext:birtney spears » ou « allintitle:britney spears ». Cela semble prouver qu'un site web, voire une page web, serait pénalisée par Google pour une requête bien particulière...

Selon la pénalité, la perte en termes de positions peut être minime (Position 6 penalty : le site passe de la 1re à la 6e place, ce qui le fait passer « en dessous de la ligne de flottaison », affectant donc de façon forte le trafic généré) ou plus importante (minus 30 : perte de 30 places ou plus...). Une pénalité « minus 950 » a même été évoquée sur certains forums...

Là encore, un silence de cathédrale nous revient de la part de Google lorsque ces pénalités sont évoquées (ce qui n'est pas illogique, notez-le bien). Difficile donc de faire la part des choses entre mythe et réalité... Mais il semblerait bien que ces pénalités punissent une suroptimisation des pages du site (*keyword stuffing*, texte et liens cachés, etc.) et ne touchent que certaines pages d'un site et pas les autres.

Matt Cutts et les pénalités Google...

Matt Cutts est le « porte-parole référencement » chez Google et son blog personnel est très lu. Voici quelques posts, sur ce blog, qui parlent des pénalités infligées par Google à certains sites. Ils sont hélas assez rares :

- *Alerting Site Owners to Problems*
  http://www.mattcutts.com/blog/webmaster-communication/
- *Confirming a Penalty*
  http://www.mattcutts.com/blog/confirming-a-penalty/
- *Notifying Webmasters of Penalties*
  http://www.mattcutts.com/blog/notifying-webmasters-of-penalties/

Pénalité numéro 3 – La baisse de PageRank dans la Google Toolbar

Google a également utilisé, lors d'une campagne de communication contre la vente de liens (*paid linking*), une pénalité consistant à faire baisser, dans la barre d'outils qu'il propose (Google Toolbar), la valeur du PageRank affiché, sans que cela affecte, *a priori*, le positionnement du site dans ses résultats sur les requêtes saisies par les internautes.

Cette pratique a fait couler beaucoup d'encre (virtuelle), aussi nous n'y reviendrons pas plus en détail, d'autant plus que les effets n'en sont pas réellement dévastateurs pour le trafic généré. Disons qu'il s'agit plus là d'un « avertissement » clairement destiné aux

webmasters s'intéressant au référencement, avertissement qui permet de véhiculer une communication institutionnelle sur les blogs et les forums spécialisés... et d'alimenter le buzz...

Pénalité numéro 4 – La liste noire

On rentre ici dans la pénalité la plus « dure » avec la « blacklist » ou « liste noire ». Pour savoir si vous y avez plongé (malheur à vous !), le plus simple est d'utiliser une requête avec la syntaxe « site:www.votresite.com ». Si Google ne renvoie aucun résultat alors qu'auparavant ce n'était pas le cas, il y a effectivement de fortes chances pour que votre site soit blacklisté et... que vous l'ayez bien cherché...

Pour en savoir plus sur les pénalités de Google...

Voici quelques articles qui devraient vous en dire plus, en français et en anglais sur les différentes pénalités infligées par Google dans le cadre de sa « recherche qualité » :

– *Les pénalités infligées par Google*

http://www.annuaire-info.com/google-penalites.html

– *Sandbox et Blacklist de Google*

http://www.go-referencement.org/google/sandbox-et-blacklist-de-google.html

– *Coffee Talk with Senior Google Engineer: Matt Cutts*

http://www.seroundtable.com/archives/002809.html

– *Google's Most Common Penalty*

http://www.threadwatch.org/node/11502

– *Google Ranking #6 Penalty/Filter*

http://www.seobook.comv/google-ranking-6-penalty-filter

Que faire si vous êtes pénalisé ?

Deux points sont importants en ce qui concerne les pénalités infligées par les moteurs de recherche en général et Google en particulier :

- La plupart des pénalités sont infligées par des êtres humains. Si des « alertes » sont certainement envoyées par des outils automatisés ou *via* le formulaire de Spam Report, les équipes de Search Quality de Google effectuent la plupart des pénalisations à la main, après vérification de la fraude éventuelle.
- Dans 9 cas sur 10, lorsqu'un site est déclassé, mis en Sandbox ou blacklisté, son webmaster sait pourquoi sans que Google ait à le lui expliquer... Reste un cas sur 10 qui est parfois assez incompréhensible...

Sachez que, depuis que l'interface Webmaster Tool existe, Google améliore souvent celle-ci et cela a notamment été le cas en juillet 2007 avec la mise en place d'un centre de

correspondance privée avec les webmasters (*http://actu.abondance.com/2007-29/google-message-center.php*).

Ainsi, dans l'interface de ses outils pour webmasters (*http://www.google.com/webmasters/tools/*), le moteur de recherche propose une zone de contact avec ses équipes, comme on le voit sur la figure 8-6.

Figure 8-6

Interface d'échange de messages entre Google et les webmasters

C'est par ce biais, ce Message Center, que, si votre site est déclassé ou pénalisé, vous risquez de recevoir un message vous signalant ce fait. Le blog officiel de Google pour les Webmaster Tools l'a clairement indiqué à l'époque (*http://googlewebmastercentral.blogspot.com/2007/07/message-center-let-us-communicate-with.html*).

C'est également *via* cette interface que vous pourrez vous expliquer et demander à ce que votre site soit revu par les équipes de Google après avoir corrigé votre site (*http://www.google.com/support/webmasters/bin/answer.py?answer=35843&hl=fr*). La situation devrait alors s'améliorer rapidement. Tout du moins, nous l'espérons pour vous... Notez que vous recevez maintenant une notification dans le Message Center lorsque la demande de reconsidération de votre site a été prise en compte (*http://googlewebmastercentral.blogspot.com/2009/06/reconsideration-requests-now-with.html*).

Mise en liste noire

C'est évidemment un cas radical et, surtout, très rare, quoi qu'en pensent certains. En effet, on voit souvent apparaître dans les forums de discussion de nombreuses questions sur la mise en liste noire de sites web. Des webmasters paniqués ne retrouvent plus leur

site dans les résultats des moteurs de recherche et se posent des tas de questions sur la procédure à suivre dans ce cas.

Nous allons essayer d'expliquer dans les pages suivantes ce qu'il faut faire lorsqu'une telle situation se produit. On s'apercevra rapidement que, dans de nombreux cas, la solution est évidente et que le terme « blacklistage » est le plus souvent bien exagéré et le fruit d'un vent de panique passager… Mais tentons tout d'abord de récapituler les étapes de façon chronologique.

Étape 1 – Respirer un grand coup…

Dans un premier temps, imaginons donc que vous testiez, un beau matin, la présence ou le positionnement de vos pages sur les moteurs de recherche et que vous ne le trouviez plus sur l'un d'entre eux (ou plusieurs). Que s'est-il passé pendant la nuit ? Votre site a-t-il disparu ? Pas de panique…

1. Dans un premier temps, respirez un grand coup et ne paniquez pas, la situation n'est peut-être pas si grave. Avant de vous précipiter sur les forums de discussion pour indiquer que votre site est blacklisté par les moteurs alors que vos concurrents, qui font bien pire, sont encore là, asseyez-vous, prenez un café (pas trop fort) et faites quelques vérifications. Par exemple, refaites les requêtes effectuées dans un premier temps (n'avez-vous pas fait une faute de frappe ?). Essayez également d'effectuer la même requête depuis un autre ordinateur utilisant un système d'exploitation (et un navigateur) différent et si possible pas localisé dans la même zone géographique (en passant, par exemple, par un système *anonymizer*, qui permet de naviguer sur le Web de façon anonyme, sans laisser de « traces » ou par un proxy afin de ne pas fournir la même adresse IP au moteur). Cela peut avoir son importance.
2. Vérifiez que votre site est bien encore présent dans l'index des moteurs. La syntaxe « site: » (« site:www.votresite.com ») fonctionne sur tous les moteurs majeurs (Google, Yahoo!, Bing) comme nous l'avons vu précédemment. Vérifiez déjà, dans un premier temps, que vos pages sont toujours là. Si c'est le cas, vos pages sont certainement déclassées, de façon temporaire ou définitive. Bonne nouvelle, vous n'êtes pas blacklisté. Allez au point 4. Si vos pages ont disparu, il y a certainement un problème, mais peut-être temporaire seulement. Allez au point 5.
3. Vérifiez vos logs pour identifier si les robots des moteurs passent bien sur votre site. Dans le cas où les pages de votre site ne seraient pas visitées, essayez d'en trouver la cause. Cela peut venir d'erreurs dans les liens (404), d'erreurs sur le fichier `robots.txt`, d'erreurs de programmation (cela peut arriver si les pages de votre site sont prévues pour s'afficher lorsque le user-agent IE ou Mozilla est détecté mais pas celui d'un robot !).

Étape 2 – Vos pages ont été déclassées (elles sont moins bien positionnées)

4. Vos pages ont été déclassées par rapport au positionnement de la veille. Elles sont toujours là mais elles sont moins bien classées. Il peut y avoir plusieurs explications :
   – Le moteur a changé son algorithme de pertinence. Cela arrive très régulièrement.

En règle générale, si votre site est fortement déclassé, c'est qu'il y a eu une modification majeure de la part des moteurs. C'est assez rare mais cela arrive (les Google Dance ou réorganisations techniques baptisées Florida, Bourbon, Jagger ou Big Daddy, par exemple, ont fait couler beaucoup d'encre par le passé…). Dans ce cas, ces changements sont certainement amplement discutés et commentés sur les forums de discussions spécialisés. Pour obtenir une liste de ces forums, rendez-vous à l'adresse *http://ressources.abondance.com/forums.html*. Vous pourrez ainsi consulter les archives récentes avant de poster tout message. Il y a de fortes chances pour que vous y trouviez des informations sur ce qui s'est passé.

– Vos conditions techniques d'hébergement ont changé, peut-être à votre insu. Si vous ne maîtrisez pas totalement les aspects techniques, notamment de votre hébergement, demandez aux techniciens qui s'en occupent si des modifications ne sont pas intervenues dans le mois qui vient de s'écouler : changement de serveur, d'adresse IP, mise en place de redirection, de filtres robots, etc. Tout changement technique peut éventuellement avoir une incidence sur votre présence et votre positionnement sur les moteurs. Vérifiez cela.

– Le contenu de vos pages a changé. Si c'est le cas, le moteur de recherche va prendre en compte cette modification, ce qui est logique. Parfois, cela peut améliorer votre note de pertinence, parfois cela peut la faire chuter. C'est le jeu… À vous de voir ce qu'il faut faire pour éventuellement revenir en arrière et les implications que cela peut entraîner.

– Vous avez un peu trop optimisé votre site et celui-ci est pénalisé par les moteurs. C'est possible… Dans ce cas, allez au point 5.

– Vos pages sont classées en pages similaires (*duplicate content)*. Le code source de chaque page est pour le moteur en grande partie identique, beaucoup de cas de pages similaires apparaissent, notamment sur les sites de commerce électronique, où seules deux ou trois informations changent dans chaque document (le nom du produit, le nom de la photo, le prix). Pensez à placer en haut du code source un contenu assez différent pour chaque page. Vérifiez aussi que d'autres sites n'aient pas utilisé vos contenus, cela arrive de plus en plus malheureusement. Voir au chapitre 7 le cas du duplicate content qui y est amplement décrit.

– Vous ne savez pas ce qui se passe car vous sous-traitez le référencement de votre site web. Cela déplace le problème mais ne le modifie pas outre mesure. Vous devrez alors faire un point de la situation avec votre prestataire.

En tout état de cause, ne faites rien dans un premier temps et attendez toujours au moins une semaine voire quinze jours avant de modifier quoi que ce soit ! (à moins que cela soit urgent, mais sachez que tout ce que vous allez faire peut également aggraver la situation, peut-être même de façon irréversible, alors qu'une simple attente peut tout résoudre en quelques heures…). Donc, si cela vous est possible, il est recommandé d'attendre. On a vu des dizaines de cas, dans le passé, où des modifications de positionnement sur Google ou d'autres moteurs n'ont été que

temporaires. Au bout de quelques jours, voire quelques heures, tout revenait dans l'ordre. Cela peut provenir d'un retour en arrière du moteur, qui a estimé ses changements trop abrupts (le cas s'est déjà produit) ou simplement d'un « dérèglement » dû à la synchronisation des *data centers* (les différents serveurs du moteur de recherche, disséminés à travers le monde, qui doivent continuellement synchroniser leurs données pour détenir les mêmes index).

Étape 3 – Votre site a disparu de l'index

5. Votre site n'apparaît plus dans les pages de résultats du moteur. C'est effectivement problématique… mais ce n'est peut-être pas si grave. Voici quelques raisons qui ont pu conduire à cette situation :

   – Problème technique sur votre site : avez-vous changé quelque chose récemment au niveau d'éventuelles redirections, du fichier `robots.txt`, des balises meta `robots` ? Si votre site est un site dynamique ayant de nombreuses pages, il se peut que lors du crawl par les moteurs de recherche, la charge machine du serveur soit mise à mal ce qui emmène parfois les hébergeurs à interdire le crawl de la partie dynamique. Cela peut paraître idiot, mais il y a des vérifications qu'il vaut mieux faire rapidement. Pistez donc tout changement survenu dans les semaines précédentes et agissez en conséquence.

   – Votre serveur a-t-il été disponible tout le temps dans le mois qui vient de s'écouler ou a-t-il fait l'objet d'une panne ? Il se peut que le robot du moteur, lorsqu'il a voulu lire vos pages, ait trouvé un serveur inaccessible pour raisons techniques. Même si les moteurs arrivent aujourd'hui à contourner ce type de problème (le spider programme une autre visite quelques minutes ou quelques heures plus tard), il peut s'agir d'une raison valable pour que votre site ait été provisoirement considéré comme ayant disparu du Web. Honnêtement, cette version de l'histoire est cependant de moins en moins probable, les moteurs ayant heureusement fait de gros progrès à ce sujet ces dernières années.

Faites surveiller votre site par un outil adéquat

Notre conseil : souscrivez à des services de surveillance de disponibilité comme *http://www.pingwy.com/* ou *http://www.netvigie.com/*.

   – Relisez le point 4 pour vérifier que certains points qui y sont listés ne correspondent pas à un événement qui se serait passé sur votre site. Par exemple, certains changements sur votre site (relookage, etc.) peuvent avoir influencé la façon dont votre site est pris en compte par les robots. En tout état de cause, vérifiez que, si certaines pages ont changé d'URL, les robots en trouvent facilement la nouvelle version (voir chapitre 7).

– Enfin, votre site a peut-être été mis en liste noire par le moteur. Sachez cependant que cela n'est pas monnaie courante de leur part et qu'une mise en liste noire est toujours manuelle et décidée par un être humain. En d'autres termes, un blacklistage, ça se mérite, comme on l'a vu auparavant… Et vous avez peut-être reçu un avertissement dans le Message Center des Webmasters Tools...

  De plus, on peut estimer qu'un site blacklisté est clairement allé trop loin, notamment dans ses techniques d'optimisation, pour mériter une telle sanction : contenu caché, liens en masse, etc. Toute optimisation « un peu trop poussée » comme des phrases entières en gras ou dans des balises `<h1>`, ne peut pas générer une mise en liste noire (elle peut cependant générer une pénalité comme un déclassement, de façon provisoire ou définitive, de certaines pages).

En cas de mise en liste noire :

– soit vous pensez que votre site n'a rien fait pour mériter une telle sanction et décidez d'en informer Google (voir la procédure ci-après) ;

– soit votre site était effectivement à revoir à ce niveau. Les techniques à éviter sont claires et nous les avons déjà passées en revue : texte, liens ou contenu cachés (balises `noscript` orphelines – sans balises `script` –, utilisation trop poussées de systèmes comme `display:none` ou `visibility:hidden`, etc.), balise remplie de mots-clés, cloaking, pages satellites, page visible dédiée aux moteurs et sans intérêt pour l'internaute, utilisation de systèmes outranciers pour capter de nombreux liens vers votre site, etc. Vous êtes allé trop loin et il vous faudra donc bien faire votre *mea culpa*…

Comment demander une réinsertion dans l'index de Google après une mise en liste noire ?

Dans tous les cas, sachez qu'il n'existe aucun moyen de savoir, de façon officielle et certaine, si un site web a été mis en liste noire par un moteur, quel qu'il soit. Il ne peut donc s'agir que de suppositions. Votre site peut avoir disparu de l'index d'un moteur sans nécessairement être blacklisté… C'est pour cela qu'il est important d'attendre quelques jours avant d'entreprendre toute action. Vous pourrez également recevoir un message de Google dans le Message Center des Webmaster Tools. S'il a été mis en liste noire, il y a de fortes chances pour que vous en soyez averti à cet endroit. En tout cas, vous trouverez dans cette zone toute information que Google juge intéressant de vous fournir à propos de votre site.

Tout n'est donc pas perdu. Il existe également une procédure de demande de réinsertion dans l'index de Google une fois qu'un site est blacklisté, voire pénalisé. Pour cela, vous devez aller dans la zone Google Webmaster Tools et cliquer sur le lien « Réexamen du site » (*http://www.google.com/support/webmasters/bin/answer.py?answer=35843&hl=fr*, voir figure 8-7).

Figure 8-7

Formulaire de demande de réexamen du site

Lisez bien le contenu de la page qui vous demande de faire amende honorable. Avant toute demande, vous devrez donc avoir enlevé de votre site tout ce qui a pu créer le problème (texte et lien caché, système de cloaking, pages satellites, etc.). Vous avez certainement une petite idée à ce sujet...

Indiquez ce que vous avez fait pour enlever le spam (ou en tout cas ce que vous pensez que Google considère ainsi) de vos pages. Expliquez votre vision de la chose (Pourquoi votre site a-t-il été exclu selon vous ? Qu'avez-vous fait pour corriger la situation ?).

Dites clairement que vous ne recommencerez plus. Oui, c'est vrai, c'est un peu puéril, mais ce *mea culpa* sera certainement nécessaire pour voir votre situation s'arranger (ceci dit, en cas de récidive, ne vous attendez pas à des miracles...).

Par ailleurs, vérifiez bien au préalable que votre site ne figure plus dans l'index. Demander qu'un site soit réintroduit dans l'index de Google alors qu'il n'en a jamais été exclu serait, à notre avis, assez mal perçu par le moteur…

Ensuite, le délai de prise en compte de votre demande dépendra de la programmation des robots. Si vos pages sont mises à jour très souvent et que le passage des robots est quasi quotidien, cela peut être rapide à partir du moment où Google prend en compte votre

message (il semblerait que ce soit fait rapidement) et accepte votre « rémission ». Si le robot ne passe que tous les mois sur votre site, le délai peut s'allonger et prendre de 6 à 8 semaines selon Google, notamment pour des pénalités considérées comme dures (comme la liste noire). Pour des pénalités plus douces (déclassement), le délai serait d'environ 2 à 3 semaines pour revenir à une situation normale (mais ce n'est en rien un délai contractuel).

Dernière information importante : afin de ne pas avoir de soucis avec votre prestataire, demandez-lui – le mieux étant de le faire au moment de la signature du contrat – de s'engager juridiquement en cas de blacklistage.

Enfin, sachez que la procédure ci-dessus ne fonctionne que sur Google, bien évidement. Les autres moteurs de recherche n'ont pas mis en place un tel système. Comme nous l'avons dit précédemment, vous recevez ensuite une notification dans le Message Center lorsque la demande de reconsidération de votre site a été prise en compte par les équipes de Google (*http://googlewebmastercentral.blogspot.com/2009/06/reconsideration-requests-now-with.html*).

Guide de lecture

Pour conclure, nous ne saurions trop vous recommander la lecture des *guidelines* des différents moteurs de recherche en ce qui concerne le spamdexing :

Google :

– http://www.google.fr/support/webmasters/bin/answer.py?answer=35769

– http://www.google.fr/support/webmasters/bin/answer.py?answer=35291

– *http://www.google.fr/webmasters/index.html*

Yahoo! :

– *http://help.yahoo.com/l/us/yahoo/search/deletion/*

Bing :

– *http://www.bing.com/webmaster*

Nous vous avons déjà donné ces adresses dans les chapitres précédents. Nous nous permettons cependant de vous les rappeler car leur contenu est primordial. À lire donc avec attention avant tout référencement !

Optimisez votre temps d'indexation

Comment accélérer l'acceptation d'un nouveau site dans les index des moteurs ? Comment faire pour qu'au lancement d'un site web, celui-ci soit déjà indexé par les spiders des différents moteurs, même si ce n'est que provisoirement, en attendant la mise à jour suite au rafraîchissement suivant de l'index, quelques jours, ou au pire quelques semaines, plus tard ? Voici quelques astuces qui vous permettront de gagner du temps en faisant en sorte que votre site web soit présent sur les moteurs, même en version minimale, dès son lancement.

Mettez en ligne une version provisoire du site

N'attendez pas le jour du lancement pour créer votre nom de domaine et proposer une page web en ligne. Au moins deux mois avant le lancement, créez un mini-site avec une page d'accueil provisoire. Exemple pour le site KSE du réseau Abondance, à l'adresse *http://www.keyword-search-engine.com/*, où l'on pouvait déjà trouver la page présentée en figure 8-8 bien avant son lancement officiel.

Figure 8-8

Version provisoire du site Keyword Search Engine

Le site fut alors rapidement indexé par Google comme le montre la figure 8-9.

Figure 8-9

Indexation du site par Google quelques jours après la mis en ligne du site

Une fois que les pages réelles seront mises en ligne et que le site sera officiellement lancé, il suffira d'attendre la prochaine mise à jour du moteur pour que les documents soient pris en compte dans leur contenu final. Mais ils seront au moins déjà présents dans les index, ce qui est loin d'être négligeable.

Mot de passe et page d'accueil

Ne protégez pas la page d'accueil de votre site par un mot de passe, car cela bloquera les spiders et donc l'indexation de vos pages. En revanche, vous pouvez bloquer l'éventuelle aspiration des autres pages (notamment les pages de test si elles sont en accès libre) par un fichier `robots.txt` ou une balise meta `robots`. Vous pouvez éventuellement le faire par un mot de passe, mais, dans ce cas, uniquement sur les pages que vous ne désirez pas voir indexées par les moteurs et laissez la page d'accueil libre d'accès. Sinon, préférez une balise meta `robots`, car la seule lecture du fichier `robots.txt` pourrait indiquer à un internaute les emplacements de votre site de test (voir chapitre 9).

Profitez de cette version provisoire

Vous avez mis en ligne une version plus ou moins expurgée de votre site afin que celle-ci soit, dans un premier temps, indexée par les moteurs ? Profitez-en pour en faire une première version attractive et efficace.

- Créez un jeu en demandant aux internautes de deviner de quoi parlera le site une fois lancé.
- Créez un *teasing* : « Rendez-vous sur ces pages dans 10 jours, 9 jours, etc. » Et soyez à l'heure le jour J !
- Mixez *Pull* (l'internaute va à l'information) et *Push* (l'information va à l'internaute). Demandez leur adresse e-mail aux internautes afin de les prévenir, le jour où le site sera disponible (voir figure 8-10).

Figure 8-10

Formulaire de saisie d'adresse e-mail afin d'être alerté de la sortie du livre

En revanche, n'en profitez pas pour revendre la base d'adresses e-mail au plus offrant ou pour l'utiliser à autre chose que l'alerte proposée au départ. Votre image de marque pourrait en souffrir.

- Faites un mini-site de 10 pages au maximum présentant votre projet, cela peut être intéressant et faire en sorte que l'internaute revienne d'autant plus facilement sur vos pages une fois le site officiel mis en ligne.
- etc.

Avec un peu de chance, ces tentatives de promotion génèreront quelques liens sur le web, qui favoriseront l'indexation de vos pages par les moteurs. Dans tous les cas, cela constituera une promotion intéressante pour votre futur site. Pour exemple, le formulaire présent sur la page d'accueil de la figure 8-10 avait généré à l'époque plus de 3 000 demandes d'alertes e-mail de la part d'internautes intéressés. Autant de lecteurs potentiels du livre…

Proposez du contenu dès le départ

Proposez du contenu sur la ou les pages provisoires, les moteurs de recherche en sont très friands, et optimisez déjà ces pages (voir chapitres 4 et 5) : titre, texte, lien, etc.

N'utilisez pas de spam, de lien caché, de texte invisible (blanc sur fond blanc, dans des layers, etc.) ! Bannissez toute méthode frauduleuse et optimisez loyalement votre première version de site. Ce serait quand même idiot de voir votre site directement inscrit en liste noire par les moteurs avant même qu'il soit créé !

Nous ne le répéterons jamais assez : il existe un très grand nombre de possibilités pour être bien référencé sans spammer, à partir du moment où on a pris les bonnes options d'optimisation des pages avant la création du site.

Faites des mises à jour fréquentes de la version provisoire

On sait que de nombreux moteurs calquent les intervalles entre deux visites de spiders sur les fréquences de mise à jour des pages web. Exemple : la page d'accueil du site Abondance, qui propose 5 jours sur 7 les titres de l'actualité des outils de recherche, est « aspirée » tous les jours par le spider de Google, notamment. Dans ce cas, la date à laquelle le robot de Google a crawlé la page est indiquée par le moteur dans sa page de résultats. Il en est de même pour toute page web, comme le montre la figure 8-11.

Figure 8-11

Indication de la date d'indexation de la page par Google

Mais il en est ainsi pour de nombreux moteurs. La figure 8-12 illustre l'exemple d'AltaVista et de sa mention « Mise à jour dans les dernières 48 h ».

Figure 8-12

AltaVista indique également lorsqu'un site a récemment été indexé.

N'hésitez donc pas à modifier tous les jours le contenu de votre site pour que le moteur prenne l'« habitude » de vous rendre visite plus souvent.

Raccourcir le délai entre deux visites du spider est en effet très intéressant : le jour où votre site sera en ligne, il ne s'écoulera, dans le meilleur des cas, que quelques heures pour que les nouvelles versions de vos pages apparaissent sur le moteur.

N'utilisez pas le revisit after !

N'utilisez pas la balise meta `revisit-after` pour indiquer au spider de revenir selon des délais prédéfinis, puisque cette balise ne sert à rien pour le référencement et n'est prise en compte par aucun moteur majeur. Un vieux serpent de mer comme on en compte quelques-uns sur le Web...

Générez les premiers liens

On l'a vu en début de chapitre, la plupart des moteurs de recherche indexent de nouvelles pages en suivant les liens des pages web rencontrées lors de la création de leur index. Pour favoriser cette « aspiration », créez, si vous en avez la possibilité, quelques liens vers votre nouveau site depuis des sites existants.

Exemple : nous avons créé, sur toutes les pages d'accueil des sites du réseau Abondance, des liens vers les différents sites du réseau. La figure 8-13 présente les liens créés sur la page d'accueil du site Abondance, en bas de page.

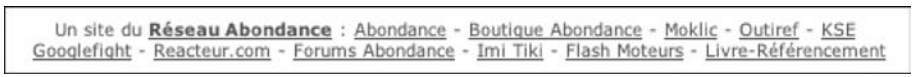


Figure 8-13

N'hésitez pas à créer les premiers liens vers un nouveau site.

Plus vous avez de sites web, plus vous pouvez multiplier ce type de signalement d'un nouveau site aux moteurs de recherche. Si vous n'avez pas d'autres sites web, essayez de contacter d'autres sites « amis » (si possible disposant de pages à fort PageRank) qui puissent faire un lien vers vous, à partir du moment où cette demande est légitime, bien sûr. Mais, là aussi, attention à l'effet Sandbox (voir précédemment)…

Inscrivez votre site sur certains annuaires dès sa sortie

Vous ne pouvez pas soumettre votre site aux principaux annuaires avant son lancement officiel, puisque les documentalistes des outils de recherche ne pourront alors pas l'évaluer dans sa version finale.

Mais n'hésitez pas à le soumettre dès sa mise en ligne. Plus votre site sera rapidement présent sur les principaux annuaires, plus vous augmenterez vos chances de le voir indexé par les principaux moteurs, qui scannent très souvent les nouveaux sites inscrits sur lesdits annuaires chaque mois. En même temps, nous avons vu au chapitre 3 que ce type de stratégie ne générait pas des miracles. Mais, si c'est bien fait, vous y gagnerez toujours quelques liens supplémentaires...

Créez des liens le plus vite possible

Comme pour l'inscription sur les annuaires, n'hésitez pas à demander des liens, dans le cadre d'échanges avec d'autres sites, vers le vôtre pour attirer les spiders vers vos pages. Comme vous vous en doutez (l'éternel *Content is king* – le contenu est le capital), plus votre contenu sera de bonne qualité, plus les liens seront faciles à obtenir. Voir chapitre 5 à ce sujet.

Évitez les systèmes de type *links farms* (échanges de liens automatiques), souvent mal vus par les moteurs et relativement inefficaces. Privilégiez les vrais liens, émanant de sites connus, sur des pages disposant d'un fort PageRank, populaires, notamment dans votre domaine d'activité. En effet, les moteurs s'orientent de plus en plus vers la contextualisation de l'information et un lien depuis un portail incontournable de votre thématique devient de plus en plus important, comparativement à un lien émanant d'un site plus généraliste.

Présentez votre site sur les forums et blogs

Les discussions des forums et les articles des blogs sont indexés par les moteurs qui suivent les liens qui y sont proposés (un site comme Googlefight – *http://www.googlefight.com/* – s'est principalement fait connaître ainsi).

Identifiez les forums et les blogs qui parlent de votre domaine d'activité et présentez votre site uniquement si cela ne passe pas pour une publicité gratuite et si cela présente une réelle information pour les internautes. Suivez la Netiquette de ces espaces communautaires et ne faites pas n'importe quoi, sinon abstenez-vous. Au pire, insérez l'URL de

votre site dans votre signature, sans en faire trop non plus. Agissez avec parcimonie sur les forums et blogs, les autres internautes ne sont pas là pour lire vos pubs !

En revanche, si vous faites bien votre travail, toujours de façon loyale et honnête, l'URL proposée sera suivie par les moteurs si le fil de discussion est indexé par les moteurs.

Votre site n'est toujours pas référencé ?

Vous vous débattez pour que votre site obtienne une meilleure présence dans les index des moteurs ? Vous avez suivi tous les conseils de ce chapitre mais vous voudriez qu'un plus grand nombre de vos pages soient indexées ? C'est une préoccupation assez normale pour un webmaster sain d'esprit. Nous allons essayer ici de lister les raisons pour lesquelles un site est mal pris en compte par les différents moteurs de recherche en tentant d'y apporter les remèdes adéquats.

Référencement n'est pas positionnement !

Attention : il ne s'agit pas ici d'essayer d'obtenir de meilleurs positionnements de vos pages en réaction à certains mots-clés, mais uniquement de faire en sorte que le plus grand nombre possible de vos pages soient référencées, donc « trouvables » sur les moteurs. Nous ne parlerons donc pas ici d'optimisation du code HTML ou de ce type de pratique visant à améliorer la réactivité de vos pages en regard des algorithmes de pertinence des outils de recherche. Chaque chose en son temps : avant d'être en tête de gondole, il faut déjà être dans les rayons…

Nous partirons donc, dans ce paragraphe, du constat que votre site n'est pas, ou est mal, indexé par les moteurs. Par exemple : seules 10 % de vos pages sont trouvées par les robots, ou seulement votre page d'accueil, etc. Il y a certainement une raison à cela. Nous allons essayer de la trouver et de vous indiquer comment faire pour améliorer la situation.

Comment lister les pages indexées par les moteurs de recherche ?

La première chose à faire, bien évidemment, est de lister toutes les pages de votre site qui sont prises en compte par les différents moteurs.

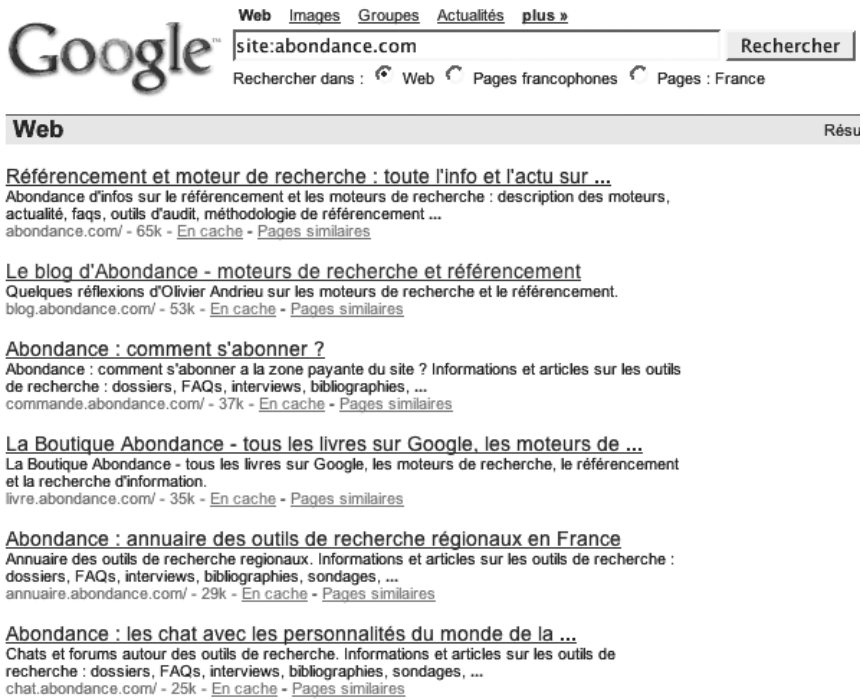
Pour cela, on utilisera le plus souvent la syntaxe « site: », comprise par la majorité des moteurs actuels et dont nous avons parlé à maintes reprises dans les pages précédentes.

Google

La commande « site:votresite.com » vous donnera la liste des pages indexées par ce moteur. En utilisant, par exemple, la requête « site:abondance.com » (sans les guillemets), nous obtenons la page présentée en figure 8-14.

Figure 8-14

La commande « site: » permet sur tous les moteurs de lister les pages indexées de votre site.



Gestion des sous-domaines pour la syntaxe « site: »

Attention cependant si vous utilisez, comme le site Abondance, des sous-domaines du type *annuaire.abondance.com*, *palmares.abondance.com*, *actu.abondance.com*, etc. La requête « site:abondance.com » vous donnera accès à toutes les pages (et donc tous les sous-domaines) du site demandé. Si vous désirez lister les pages d'un sous-domaine précis, tapez : « site:actu.abondance.com » par exemple. La requête « site:www.abondance.com » ne listera pas les pages des sous-domaines *actu.abondance.com* ou *palmares.abondance.com* mais uniquement celles présentes sous le nom de serveur *www.abondance.com*.

Yahoo! Search

La syntaxe à utiliser sur Yahoo! est également « site: ». Notre exemple reste donc « site:abondance.com » avec les mêmes restrictions et possibilités que celles évoquées précédemment.

Bing

Même motif, même punition pour Bing : « site:abondance.com »

Exalead

La syntaxe d'interrogation « site: » reste d'actualité sur Exalead : « site:abondance.com ».

Ask.com

En revanche, Ask ne comprend pas la syntaxe « site: » seule mais uniquement accompagnée d'un mot-clé de recherche comme « abondance site:abondance.com ». Ce bogue était présent sur Google pendant longtemps avant qu'il ne soit corrigé par les équipes techniques du moteur. Ce n'est pas (encore) le cas sur Ask : si aucun mot représentatif n'est présent dans vos pages, il ne semble pas qu'il soit possible de faire une requête exhaustive de ce type sur votre site, des recherches comme « * site:abondance.com » ou « ? site:abondance.com » ne donnant aucun résultat. Dommage…

Pas d'affolement !

Avant toute investigation dans une voie expliquée dans la suite de ce paragraphe, n'hésitez pas à faire un état des lieux de la présence de vos pages dans les index des moteurs grâce à ces syntaxes de recherche. Combien voit-on de messages dans les forums du type « mon site a disparu de Google ! » alors qu'il a simplement été rétrogradé, parfois temporairement, sur certains mots-clés…

Différentes raisons de non-indexation de votre site par les moteurs

Votre site est refusé par les moteurs de recherche ? Il est pourtant mis en ligne depuis plusieurs semaines ? Il peut y avoir plusieurs raisons à cela.

Blocage par un fichier robots.txt ou des balises meta robots

Cela peut paraître idiot, mais il existe de nombreux exemples de non-indexation de sites web en raison de la présence d'un fichier `robots.txt` (voir chapitre 9) empêchant les robots des moteurs d'indexer le site ou de balise meta `robots` ayant un attribut `noindex` dans les pages. Nous avons déjà évoqué ce problème auparavant.

Vérifiez donc vos fichiers `robots.txt` et vos balises meta `robots` afin d'éviter tout problème de ce type. Cela prend quelques secondes et peut éviter bien des désagréments !

De la même façon, si vous avez mis en place des redirections serveurs (301, 302) vérifiez la syntaxe des fichiers `.htaccess` utilisés, afin d'être sûr que les moteurs les lisent de façon adéquate.

Manque de backlinks vers votre site

Nous avons vu dans les pages précédentes que la meilleure façon d'accéder à une indexation rapide et efficace reste d'obtenir le plus vite possible un ou plusieurs liens (backlinks) vers votre page d'accueil depuis une page populaire, c'est-à-dire disposant d'un PageRank d'au moins 6. Ne connaissez-vous personne qui détienne ce type de page et qui puisse vous rendre ce service ?

Si cela vous semble difficile, tentez de multiplier les backlinks émanant de pages à PageRank le plus fort possible, inscrivez-vous dans des annuaires, etc. Faites donc en sorte d'obtenir, dès les premiers jours de vie de votre site, des liens émanant de pages les plus populaires possibles. Évitez cependant de faire n'importe quoi… Trop de liens obtenus trop rapidement peuvent aussi jouer dans l'effet Sandbox sur Google (voir précédemment). Tablez donc sur une dizaine ou une vingtaine de liens rapides qui feront logiquement indexer votre site en quelques jours. Encore une fois, ne soyez pas tenté par des sirènes commerciales comme l'achat de lien au *prorata* du PageRank ou ce type d'offre qui commence à fleurir sur le Web. Faites les choses proprement, de façon loyale et honnête et vous n'aurez aucun souci par la suite.

Utilisation de technologies bloquantes pour les robots

Il existe finalement assez peu de technologies qui soient réellement rédhibitoires pour les robots et leur interdisent l'accès à vos pages. Mais cela peut arriver. Par exemple, l'obligation d'accepter les cookies, l'emploi d'identifiants de sessions sur un site dynamique dès la page d'accueil, peuvent poser des problèmes et provoquer une non-indexation totale du site. Le chapitre 7 de cet ouvrage est entièrement consacré à ce point. Vérifiez donc que rien, techniquement parlant, ne peut bloquer un robot sur votre site, on ne sait jamais.

Autre site posant problème sur votre serveur

Autre point pour lequel on dispose d'encore peu d'informations aujourd'hui : si vous êtes sur un serveur mutualisé, donc partagé par plusieurs sites ne vous appartenant pas forcément, un moteur de recherche peut vous pénaliser en bloquant l'adresse IP d'un site ayant par trop spammé son index. Même si les moteurs s'en défendent, ce cas de figure peut arriver. Vérifiez donc, en utilisant des outils comme Whois.sc (*http://www.whois.sc/*) ou la recherche par adresse IP de Bing (syntaxe IP : en fournissant l'adresse IP de votre serveur, par exemple IP:236.23.67.89), quels sont les autres sites hébergés sur le même serveur que le vôtre et regardez s'ils ne sont pas potentiellement des sources de problèmes pour les moteurs.

Nous concluons ainsi ce chapitre sur le référencement de votre site web dans les index des moteurs de recherche, et d'éventuelles pénalités que ces outils lui auraient infligées (ce que nous ne vous souhaitons pas, bien entendu). En revanche, il existe un certain nombre de cas pour lesquels un webmaster ne désire pas que son site se retrouve dans les index des moteurs de recherche. Cela fera l'objet du chapitre suivant...

Un exemple de référencement effectué en quelques jours

Le référencement d'un site web et sa visibilité sur les moteurs de recherche est un levier essentiel de création de trafic à moyen et long terme. Mais on oublie parfois qu'il est possible de générer du trafic à court terme, voire avant même que le site web en question ne soit réellement en ligne. Pour appuyer ce fait sur un cas concret, nous allons voir comment un site a été positionné sur certains mots-clés en quelques heures, bien avant sa mise en ligne officielle. Suivez le guide dans ce petit « travail pratique »...

Voici le début de l'histoire : en 2008, un ami, expert-comptable à Barr (Bas-Rhin), demande à l'auteur de cet ouvrage s'il peut lui créer un « petit site web » pour présenter son activité, alors qu'il vient de se mettre à son compte dans le cadre d'une entreprise qu'il a baptisée « ECR Conseil ». La création de site n'est pas du tout l'activité de l'auteur, mais il lui arrive de dépanner un ami de cette sorte (tout en lui expliquant que s'il désire un site professionnel à forte ambition, il ne frappe pas à la bonne porte...).

Bref, l'idée, ici, est d'avoir quelques pages en ligne pour une clientèle locale et d'obtenir une visibilité sur Google pour sa marque (« ECR Conseil ») ainsi que sur des expressions correspondant à son activité comme « comptable Barr » ou « expert comptable Barr », voire « expert comptable Bas-Rhin »... Expressions peu concurrentielles s'il en est. Nous allons donc ici tenter de décrire la genèse du projet et d'expliquer les choix retenus pour le référencement du site afin d'arriver à cet objectif...

Étape 1 – Choix du nom de domaine

Première étape logique : le choix du nom de domaine, car il nous semble évident que si l'on veut avoir un minimum d'ambition sur le Web, l'achat d'un nom de domaine est une étape nécessaire et obligatoire. Au vu du coût d'une telle démarche (une dizaine d'euros par an), on voit mal pour quelles raisons on s'en passerait... La société s'appelant « ECR Conseil », il y avait plusieurs possibilités en termes de TLD (*Top Level Domain*) : *.fr, .com, .biz*, etc. Finalement, le *.com* a été retenu par rapport au *.fr* (ce qui n'amenait de toute façon rien de plus pour le référencement, les deux domaines étant globalement équivalents en termes de prise en compte par les moteurs).

Ensuite, le choix s'est porté sur le tiret éventuel entre les deux mots : fallait-il acheter *ecrconseil.com* ou *ecr-conseil.com* ? Pour les moteurs, il n'y a pas photo : *ecr-conseil.com*, avec un tiret pour séparer les deux mots, était préférable (voir chapitre 4). Mais l'expert-comptable a préféré opter pour l'option « sans tiret », plus facile à retenir pour l'internaute moyen, d'autant plus que sa clientèle n'est pas technophile et ne connaît pas obligatoirement les subtilités des « traits d'union », « tiret » ou autres « undescores », etc. Bref, le nom de domaine *ecrconseil.com* a été acheté, tout en étant conscient que cela n'était pas le choix optimal pour le référencement... De même, l'impasse a été faite sur des formes différentes de mot comme *ecrconseils.com* (avec un « s »). Ce type de question aurait pu être prise en compte dans le cadre d'un projet ambitieux à visibilité nationale, voire internationale. Ce n'était pas ici le cas...

Étape 2 – Création d'une maquette d'attente

Au niveau du contenu, l'ami n'était pas très pressé et n'avait pas encore préparé les textes et images à mettre en ligne. L'idée a alors été de voir s'il était possible, sur la base d'une simple « carte de visite », de commencer à faire venir du monde sur le site, pour proposer les coordonnées postales et l'adresse e-mail pour contacter l'entreprise, par le biais du référencement naturel. Voici comment nous avons procédé :

- Une page très simple (disponible à l'adresse *http://www.ecrconseil.com/*) a été créée, avec un simple logo et les coordonnées de la société.
- Le but était de positionner cette page sur les requêtes suivantes :
  - « ecr conseil » (6 590 000 résultats sur Google) ;
  - « comptable barr » (129 000 résultats sur Google) ;
  - « expert comptable barr » (60 800 résultats sur Google) ;
  - « expert comptable bas-rhin » (53 500 résultats sur Google).
- Le tout dans un délai qui soit le plus court possible.

Vous pourrez visualiser cette page, qui propose un aspect pour le moins spartiate, en figure 8-15.

Figure 8-15

Page d'accueil du site ecrconseil.com

L'optimisation suivante a été mise en place :

Balise `<title>` :

```
<title>ECR CONSEIL - Expert Comptable &agrave; Barr (67140 - Bas-Rhin)</title>
```

Tous les mots-clés pour lesquels un positionnement est demandé (« ecr », « conseil », « expert », « comptable », « barr », « bas-rhin ») s'y trouvent. Sa taille (7 mots descriptifs) est dans la « norme acceptable » (entre 7 et 10 mots) pour obtenir un bon référencement. Le nombre de caractères inférieur à 70 garantit qu'il sera entièrement affiché dans les pages de résultats de Google et donc tout à fait lisible et compréhensible par les internautes.

Balise meta `description` :

```
<meta name="description" content="ECR CONSEIL - Michel Richter, Expert Comptable
➥&agrave; Barr (67140 - Bas-Rhin). Adresse : 6, rue d'Alsace - ZAC du Muckental -
➥67140 Barr. T&eacute;l : 03 88 58 00 22 - Fax : 03 88 08 20 97">
```

Elle développe (190 caractères) la balise `<title>` et reste donc en cohérence avec son contenu tout en restant inférieure en taille à 200 caractères.

Texte de la page :

Il est assez simple : quelques mots comportant le nom et l'activité (« ECR CONSEIL - Expertise Comptable Richter - Expert Comptable ») en gras (balise `<strong>`) et les coordonnées de l'entreprise. Pas de contenu caché ou autres joyeusetés de ce type...

Est-ce que cela allait suffire ? La question restait posée car le volume de texte était très faible...

La page a été mise en ligne sous cette forme le 10 juin 2008.

Étape 3 – Détection du site par les moteurs de recherche

Comment faire connaître le site aux moteurs de recherche ? Rien de plus simple : sur la page d'accueil du site Abondance (*http://www.abondance.com/*), nous avons mis un lien dans le footer vers le site avec comme intitulé « ECR CONSEIL » (nom de l'entreprise et requête visée pour un bon positionnement), pointant bien entendu vers la page d'accueil (unique page d'ailleurs) du site.

Figure 8-16

Lien vers le site dans le footer du site Abondance.com

Résultat ? Moins de 12 h après la mise en place de ce lien, la page était indexée par Google et Yahoo! (il a en revanche fallu plus de deux mois pour qu'elle soit indexée par Live Search/Bing, mais ce moteur est connu pour la paresse de ses robots, hélas).

Fin juin, on peut voir sur la figure 8-17 que c'étaient les robots de Yahoo! (Slurp) et de Google (Googlebot) qui avaient été les plus actifs sur le site (mais on peut remarquer la présence de celui de Baidu en bonne position : bientôt l'entreprise sera connue en Chine ! mais mon ami n'en a cure, hélas...

Figure 8-17

Statistiques sur les venues des robots sur le site ECR Conseil

Visiteurs Robots/Spiders (Top 10) - Liste complète - Dernière visite

10 robots différents*	Hits	Bande passante	Dernière visite
Inktomi Slurp	56	29.90 Ko	30 Juin 2008 - 13:46
Googlebot	16	49.89 Ko	27 Juin 2008 - 22:28
BaiDuSpider	10	11.09 Ko	30 Juin 2008 - 16:27
Unknown robot (identified by 'crawl')	8	135.86 Ko	29 Juin 2008 - 02:07
Alexa (IA Archiver)	3	3.36 Ko	24 Juin 2008 - 10:02
GigaBot	2	2.19 Ko	22 Juin 2008 - 07:37
Unknown robot (identified by 'robot')	2	2.19 Ko	21 Juin 2008 - 13:55
Grub.org	1	1.17 Ko	25 Juin 2008 - 19:34
LinkWalker	1	1.17 Ko	18 Juin 2008 - 17:49
Walhello appie	1	1.02 Ko	17 Juin 2008 - 16:07

Étape 4 – Évaluation du travail effectué

Une fois ce lien mis en place, quel a été le résultat obtenu sur les moteurs de recherche ?

Sur Google :

- Requête « ecr conseil » : 1re position sur plus de 6 millions de résultats en 48 h.
- Requête « comptable barr » : 4e à 10e position (selon le jour) sur près de 130 000 résultats en 48 h également.
- Requête « expert comptable barr » : 1re position sur près de 60 000 résultats, en moins de 48 h.
- Requête « expert comptable bas-rhin » : 10e position sur plus de 54 000 résultats, en moins de 48 h.

Sur Yahoo! :

- Requête « ecr conseil » : 1re position sur 213 000 résultats en 48 h.
- Requête « comptable barr » : 14e position sur plus de 20 000 résultats en 48 h.
- Requête « expert comptable barr » : 5e à 7e position sur plus de 2 000 résultats, en moins de 48 h.
- Requête « expert comptable bas-rhin » : 9e position sur plus de 337 000 résultats, en moins de 48 h.

Les résultats sont donc plutôt bons et intéressants en termes de positionnement : quasiment que des premières pages et quelques premières positions. Et le tout en quelques heures et sans « recette miracle » autre que les conseils indiqués dans ce livre...

Ceci dit, le but de cet exemple est surtout de bien faire comprendre que :

- Il est possible d'obtenir d'excellents positionnements très rapidement avec très peu d'efforts (la création de la page web et sa mise en ligne n'ont pas duré plus de deux heures) à partir du moment où les requêtes visées ne sont pas très concurrentielles – comme c'est le cas ici – et que l'optimisation est « bien faite ».
- Bien sûr, si les mots-clés de positionnement espérés avaient été plus concurrentiels (plus de résultats renvoyés, plus de sociétés tentant de se positionner dessus), il aurait certainement fallu plus d'efforts dans deux voies différentes : plus de contenu textuel et plus de liens entrants. Logique. Évident. Dans notre cas, une configuration minimaliste a suffi. Tant mieux... Mais elle correspond à bon nombre de besoins d'entreprises qui travaillent sur un marché local et qui n'ont pas besoin de se faire connaître à Bruxelles, Washington ou Pékin...
- Ceci dit, obtenir une première place sur le nom d'une entreprise alors que Google renvoie 6 millions de résultats sur cette requête et le tout en moins de 48 h, cela reste intéressant. Pour cela, il a suffi de suivre quelques règles très simples énoncées dans cet ouvrage : balises `<title>` et meta `description` cohérentes, un minimum de texte, un lien depuis une page populaire avec un anchor text (texte du lien) correspondant au nom de la société (« ECR CONSEIL ») et rien de plus. Le rapport « temps de travail/positionnement obtenu » est pour le moins optimisé... Et cela prouve également que cet ouvrage ne raconte pas que des bêtises !

La conclusion de cet exemple pratique est simple : il est tout à fait possible de voir son site référencé – et obtenir de bons positionnements – avant même qu'il soit réellement en ligne, si les mots-clés visés ne sont pas trop concurrentiels dans un premier temps. Cela ne vaut-il pas la peine de « tenter le coup » grâce à quelques heures de travail en attendant que le contenu prévu soit finalisé et les pages du site terminées ? D'autant plus que quand ce contenu final sera mis en ligne, Google connaîtra déjà votre site, aura déjà indexé sa page d'accueil dans sa première version et une partie du travail sera donc fait... C'est toujours ça de gagné...

9

Comment ne pas être référencé ?

Dans les précédents chapitres de cet ouvrage, nous vous avons parlé des différentes manières de référencer votre site sur les moteurs de recherche. Mais il peut arriver qu'on ait besoin de déréférencer une source d'informations déjà indexée ou de signaler aux moteurs un certain nombre de pages ou d'informations à ne pas prendre en compte, notamment lors de tests. Un site ou une page en construction, par exemple, ne doit pas obligatoirement être la cible d'une « aspiration » par les spiders des moteurs. Idem pour des images qui seraient soumises à un copyright. Il faut alors empêcher certains robots de les prendre en compte. Il existe heureusement plusieurs façons de signaler ceci aux moteurs de recherche. Nous allons les passer en revue dans ce chapitre.

Fichier robots.txt

Si vous désirez que votre site ou certaines de vos pages ne soient plus pris en compte par les moteurs, la première possibilité est d'insérer un fichier `robots.txt` adéquat sur votre serveur. Ce fichier va donner des indications au spider du moteur sur ce qu'il peut faire et ce qu'il ne doit pas faire sur le site.

Dès que le spider d'un moteur arrive sur un site (par exemple, sur l'URL *http://www.monsite.com/*), il va rechercher le document présent à l'adresse *http://www.monsite.com/robots.txt* avant d'effectuer la moindre « aspiration ». Si ce fichier existe, il le lit et suit les indications qui y sont fournies. S'il ne le trouve pas, il commence son travail de lecture et de sauvegarde de la page HTML qu'il est venu visiter, considérant qu'*a priori* rien ne lui est interdit.

Quelques points à vérifier

Il est important que votre fichier `robots.txt` soit à la racine de votre site. Sans cela, il ne sera pas pris en compte par les moteurs de recherche. En outre, il ne peut exister qu'un seul fichier `robots.txt` sur un site. Enfin, le nom du fichier (`robots.txt`) doit toujours être en minuscules. Attention également à ne pas oublier le « s » final du nom du fichier : « robots ».

La structure d'un fichier `robots.txt` est la suivante :

```
User-agent: *
Disallow: /cgi-bin/
Disallow: /tempo/
Disallow: /abonnes/prix.html
```

Dans cet exemple :

- `User-agent: *` signifie que l'accès est accordé à tous les agents (tous les spiders), quels qu'ils soient ;
- le robot n'ira pas explorer les répertoires `/cgi-bin/` et `/tempo/` du serveur ni le fichier `/abonnes/prix.html`.

Le répertoire `/tempo/`, par exemple, correspond à l'adresse *http://www.monsite.com/tempo/*.

Chaque répertoire à exclure de l'aspiration du spider doit faire l'objet d'une ligne `Disallow:` spécifique. La commande `Disallow:` permet d'indiquer que « tout ce qui commence par » l'expression indiquée ne doit pas être indexé.

Ainsi :

- `Disallow: /perso` ne permettra l'indexation ni de *http://www.monsite.com/perso/index.html*, ni de *http://www.monsite.com/perso.html* ;
- `Disallow: /perso/` n'indexera pas *http://www.monsite.com/perso/index.html*, mais ne s'appliquera pas à l'adresse *http://www.monsite.com/perso.html*.

De plus :

- le fichier `robots.txt` ne doit pas contenir de lignes vierges (blanches) ;
- l'étoile (`*`) n'est acceptée que dans le champ `User-agent`. Elle ne peut servir de joker (ou d'opérateur de troncature) comme dans l'exemple : `Disallow: /entravaux/*` ;
- il n'existe pas dans le standard initial de champ correspondant à la permission, de type `Allow:` (même si certains moteurs le permettent maintenant : *http://actu.abondance.com/2008/06/microsoft-yahoo-et-google-sentendent.html*) ;
- enfin, le champ de description (`User-agent`, `Disallow`) peut être indifféremment saisi en minuscules ou en majuscules.

Les lignes qui commencent par un signe dièse (*#*), ou plus exactement tout ce qui se trouve à droite de ce signe sur une ligne, est considéré comme un commentaire.

La tableau 9-1 présente quelques commandes très classiques et non moins importantes du fichier `robots.txt`.

Tableau 9-1 Syntaxe d'utilisation du fichier robots.txt

Syntaxe	Explications
`Disallow:/`	Permet d'exclure toutes les pages du serveur (aucune aspiration possible).
`Disallow:`	Permet de n'exclure aucune page du serveur (aucune contrainte). Un fichier `robots.txt` vide ou inexistant aura une conséquence identique.
`User-Agent: googlebot`	Permet d'identifier un robot particulier (ici, celui de Google).

`User-agent: googlebot`
`Disallow:`
`User-agent: *`
`Disallow:/`
	Permet au spider de Google de tout aspirer, mais « ferme la porte » aux autres.

N'oubliez pas également que le fichier `robots.txt` permet de déclarer votre fichier Sitemap en ajoutant simplement cette ligne :

```
Sitemap: <sitemap_location>
```

Par exemple :

```
Sitemap: http://www.votresite.com/sitemaps.xml
```

Vous pouvez vous reporter au chapitre 8 pour plus d'informations.

Quelques liens utiles au sujet du fichier robots.txt

– Comment trouver les noms des robots des différents moteurs ?
http://www.robotstxt.org/orig.html

– Un autre article sur la syntaxe de ce fichier :
http://www.searchtools.com/robots/robots-txt.html

– Vérificateur de syntaxe pour votre fichier `robots.txt` :
http://tool.motoricerca.info/robots-checker.phtml

Notez que les Webmaster Tools de Google (*http://www.google.com/webmasters/tools/?hl=fr*) proposent également un générateur et un vérificateur de fichier `robots.txt`.

Balise meta robots

Nous avons vu dans le chapitre 4 qu'il existait des balises meta à insérer dans le code source de vos pages, permettant ainsi de délivrer un certain nombre d'informations, au travers des balises `description` et `keywords`, aux moteurs de recherche, et ce même si leur importance a fortement baissé depuis quelques années.

De très nombreuses autres balises meta sont disponibles et parfois visibles dans le code HTML des pages web : `revisit-after`, `classification`, `distribution`, `rating`, `identifier-URL`, `copyright`, etc. Rappelons ici qu'elles ne sont clairement prises en compte par aucun moteur de recherche majeur. Leur présence est donc superflue dans vos pages, si ce n'est pour un autre but que le référencement.

Seule la balise `<meta name="robots">` nous intéressera ici.

Elle ne sert jamais comme critère de pertinence pour les moteurs, mais elle permet de leur indiquer la façon dont ils doivent indexer la page. Une balise meta `robots` spécifique peut effectivement être utilisée – dans chaque document HTML – pour permettre ou interdire l'accès aux spiders des moteurs. Elle se présente sous la forme suivante :

```
<meta name="robots" content="attribut1,attribut2">
```

Où les champs `attribut1` et `attribut2` peuvent prendre les valeurs suivantes :

- `attribut1` :
  - `index` : page à indexer par le spider ;
  - `noindex` : interdiction d'indexer la page.
- `attribut2` :
  - `follow` : le spider peut suivre les liens contenus dans la page pour indexer d'autres documents ;
  - `nofollow` : le spider ne peut pas suivre les liens de la page.

Les indications `index`, `noindex`, `follow` et `nofollow` peuvent indifféremment être saisies en minuscules et en majuscules. Voici les différentes possibilités offertes par cette balise :

```
<meta name="robots" content="index,follow" />
<meta name="robots" content="noindex,follow" />
<meta name="robots" content="index,nofollow" />
<meta name="robots" content="noindex,nofollow" />
```

Ces balises meta, comme celles présentées auparavant, doivent se trouver dans l'en-tête du document HTML, entre `<head>` et `</head>` et si possible après la balise `<meta name="keywords">` (si elle existe). Elles doivent figurer dans tous les documents dont vous désirez filtrer l'accès, contrairement au fichier `robots.txt` qui prend en compte toute l'arborescence d'un site. Grâce à cette balise, l'accès aux robots est très finement filtrer, à la page près, ce qui est plus complexe (mais cependant pas impossible) avec le fichier `robots.txt`.

Enfin deux derniers points sont à préciser :

- Le premier exemple donné ci-dessus (`"index,follow"`) n'a pas d'application pratique. En effet, elle est équivalente à l'absence de balise meta `robots`.
- Les syntaxes suivantes sont équivalentes :

```
<meta name="robots" content="index, follow" /> et <meta name="robots" content="all" />
<meta name="robots" content="noindex, nofollow" /> et <meta name="robots" content="none" />
```

Tous les moteurs de recherche majeurs prennent en considération cette balise meta, tout comme ils explorent le fichier `robots.txt`.

> **Procédure d'urgence sur Google**
>
> On peut noter que Google propose une procédure d'urgence qui permet d'éliminer très rapidement des pages web de son index, dans les Webmasters Tools du moteur (*http://www.google.com/webmasters/tools/?hl=fr*).
>
> Vous trouverez également davantage d'informations à cette adresse : *http://www.google.com/support/webmasters/bin/answer.py?answer=35301*.

En mettant en ligne ces informations – fichier `robots.txt` ou balise meta `robots` – sur votre site, vous indiquerez clairement au spider, lors de son prochain passage, ce qu'il doit faire ou ne pas faire. Dans ce cas, il supprimera de son index (s'il est bien programmé) les pages que vous lui demandez de ne plus indexer ou il ne les indexera pas la première fois qu'il les rencontrera.

Fonctions spécifiques de Google

Google propose un certain nombre de fonctionnalités qui lui sont propres et qui permettent de mieux gérer la façon dont cet outil indexe vos pages. Voici un florilège des possibilités supplémentaires qu'offre ce moteur.

Balise meta robots spécifique

Pour empêcher uniquement les robots de Google d'indexer une page de votre site tout en autorisant cette opération à d'autres robots, utilisez la balise suivante :

```
<meta name="googlebot" content="noindex, nofollow" />
```

Bien entendu, vous pouvez également utiliser des attributs comme `index` ou `follow` dans cette balise (voir précédemment).

Suppression des extraits textuels (snippet)

Un snippet est, comme le montre la figure 9-1, un extrait de texte qui apparaît parfois sous le titre d'une page dans les résultats de recherche du moteur et qui décrit le contenu de la page en question.

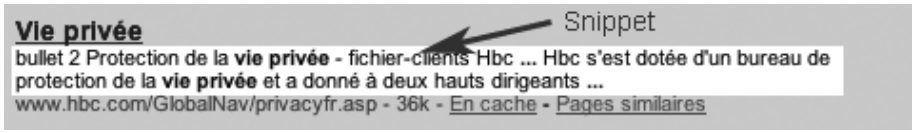


Figure 9-1

Exemple de snippet dans les résultats de Google

Pour éviter que Google affiche des extraits de votre page (ce qui serait dommage, mais vous faites comme vous le désirez...), placez la balise suivante dans la section `<head>` de la page :

```
<meta name="googlebot" content="nosnippet" />
```

Suppression des extraits issus de l'Open Directory

Google propose également une balise à insérer dans ses pages, permettant de ne pas afficher le descriptif issu de l'Open Directory pour une page donnée dans ses résultats. En effet, selon le cas, Google affiche trois sources d'informations différentes pour décrire une page (voir chapitre 4) :

- le contenu de la balise meta `description` de la page ;
- le descriptif issu de l'annuaire Open Directory (Dmoz) ;
- un snippet, extrait textuel de la page contenant le mot demandé (voir ci-dessus).

Les balises suivantes (au choix) permettront de refuser l'affichage du descriptif de l'Open Directory par Google :

```
<meta name="robots" content="noodp" />
<meta name="googlebot" content="noodp" />
```

Bing, le moteur de recherche de Microsoft, accepte également cette balise. Le même « tag » (dans sa première version ci-dessus) servira donc aux deux outils.

Yahoo! utilise son annuaire, Guide Web, et a donc sa balise particulière sous la forme :

```
<meta name="robots" content="noydir" />
```

ou :

```
<meta name="slurp" content="noydir" />
```

Suppression de contenu inutile

Yahoo! a également annoncé en 2007 que ses robots allaient dorénavant suivre des consignes que les webmasters pouvaient mettre dans leurs codes sources sous la forme d'une balise nommée `robots nocontent`. Cette balise, sous forme d'une classe CSS, conjuguée avec des balises comme `div`, `p` ou `span`, permet par exemple, d'indiquer aux robots que le contenu qui suit n'est pas le plus important dans la page et ne doit pas être indexé. Exemples fournis par Yahoo! :

```
<div class="robots-nocontent">
  This is the navigational menu of the site and is common on all pages. It contains
  ➥many terms and keywords not related to this site
</div>
<span class="robots-nocontent">
  This is the site header that is present on all pages of the site and is not related to
  ➥any particular page
```

```
</span>
<p class="robots-nocontent">
  This is a boilerplate legal disclaimer required on each page of the site
</p>
<div class="robots-nocontent">
  This is a section where ads are displayed on the page. Words that show up in ads may
  ➥be entirely unrelated to the page contents
</div>
```

Dans chacun de ces exemples, le texte contenu entre les balises indiquées ne sera pas pris en compte pour des recherches et ne sera pas affiché dans les snippets. Yahoo! était en 2009 le seul à proposer cette fonction qui disparaîtra peut-être avec l'abandon de sa technologie de recherche au profit de Bing.

En revanche, Google propose une balise meta intitulée `unavailable_after` qui indique aux moteurs qu'une page ne sera plus pertinente ou sera indisponible après une certaine date et qu'il ne sera plus la peine pour les spiders de l'indexer. Une façon, par exemple, d'informer Google qu'une page présentant une offre promotionnelle limitée dans le temps, n'est plus valable passé un certain délai. Exemple :

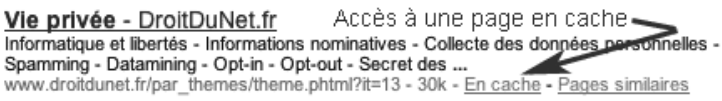
```
<meta name="googlebot" content="unavailable_after: 23-Jul-2007 18:00:00 EST" />
```

Suppression des pages en cache

Google prend automatiquement un instantané de chaque page explorée afin de l'archiver. Cette version en cache permet d'extraire une page web pour les utilisateurs finaux si la page d'origine vient à être indisponible (en raison d'un arrêt temporaire du serveur web de la page). La page en cache se présente exactement comme elle se présentait la dernière fois que Google l'a analysée. Seule différence : un message apparaît en haut de la page afin d'indiquer qu'il s'agit de la version en cache. Les utilisateurs peuvent accéder à la version en cache en cliquant sur le lien « En cache » sur la page des résultats de recherche (figure 9-2).

Figure 9-2

Exemple d'accès à une page en cache dans les résultats de Google



Pour empêcher tous les moteurs de recherche de proposer un lien en cache pour votre site, placez cette balise dans la section `<head>` de la page :

```
<meta name="robots" content="noarchive" />
```

Pour empêcher uniquement Google de proposer un lien en cache et autoriser cette opération pour les autres moteurs de recherche, utilisez la balise suivante :

```
<meta name="googlebot" content="noarchive" />
```

Suppression d'images

Pour supprimer une image de l'index des images Google (ou de tout autre moteur), le mieux est de vous servir du fichier `robots.txt` vu précédemment.

Par exemple, si vous souhaitez que les moteurs excluent l'image `chiens.jpg` qui apparaît sur votre site à l'adresse *www.votresite.com/images/chiens.jpg*, insérez dans votre fichier `robots.txt` le code suivant :

```
User-agent: *
Disallow: /images/chiens.jpg
```

Pour interdire uniquement au robot de Google l'accès à ce fichier :

```
User-agent: Googlebot-Image
Disallow: /images/chiens.jpg
```

Pour supprimer de l'index de Google toutes les images de votre site uniquement, placez le fichier `robots.txt` suivant à la racine de votre serveur :

```
User-agent: Googlebot-Image
Disallow: /
```

Vous pouvez également, si vous désirez interdire l'accès à toutes vos images (pour des questions de droit d'auteur, par exemple), stocker tous ces fichiers dans un répertoire unique (suggestion : `/images/`) et en interdire l'accès à tous les robots :

```
User-agent: *
Disallow: /images/
```

Spécificités de Google

Google a en outre accentué la souplesse d'utilisation du protocole `robots.txt` grâce à la prise en charge des astérisques. Les formats d'interdiction peuvent inclure le signe `*` pour remplacer toute séquence de caractères et se terminer par le symbole `$` pour indiquer la fin d'un nom. Pour supprimer tous les fichiers d'un type particulier, par exemple, pour inclure les images `.jpg` mais pas les images `.gif`, utilisez l'entrée de fichier `robots.txt` suivante :

```
User-agent: Googlebot-Image
Disallow: /*.gif$
```

Attention : cette syntaxe ne fonctionne qu'avec le moteur Google.

Pour le moteur de Yahoo!, la procédure est la même (sauf pour l'utilisation des caractères `*` et `$` qui est spécifique de Google) avec le user-agent :

```
User-agent: Yahoo-MMCrawler
```

Pour Exalead (qui prend en compte les caractères `*` et `$`), utilisez :

```
User-agent: Exabot
```

Bing et Ask.com, à qui nous avons osé la question, ne proposent en revanche pas de solution spécifique basée sur le fichier `robots.txt` et un user-agent spécifique de leur robot images. Mais ces deux moteurs nous ont proposé une voie différente pour que vos images ne soient pas indexées : intégrer les images dans un lecteur Flash, ce qui va empêcher les moteurs de les « trouver ».

C'est effectivement une possibilité intéressante. On préférera cependant les solutions basées sur le fichier `robots.txt`, plus pérennes à notre sens...

10

Méthodologie et suivi du référencement

Référencer son site web est une phase essentielle dans le cycle de promotion d'une source d'informations sur le Web. Mais cette stratégie demande une méthodologie efficace pour sa mise en place tout autant que pour la mesure de son efficacité. À quoi bon investir dans une action de promotion bien menée si vous ne savez pas ce qu'elle vous rapporte ? Ces deux points cruciaux (méthodologie et suivi) seront traités dans ce chapitre.

La règle des « 3C » : Contenu, Code, Conception

Pour optimiser au mieux le référencement d'un site, il est devenu important, voire primordial, de le penser dès le départ pour être compatible avec les différents moteurs de recherche qui vont venir le visiter, grâce à leurs spiders, robots qui viennent « aspirer » les pages web et suivre les liens qu'elles contiennent. Si vous avez lu assidûment les précédents chapitres de cet ouvrage, vous devez en être convaincu...

Pour qu'un site soit parfaitement « compris » et « analysé » par les moteurs de recherche, il faut donc qu'il ait été pensé pour être compatible avec les critères d'exploration et de pertinence de ces outils. Nous allons, dans ce début de chapitre, expliciter une règle qui nous est chère et que nous avons pu expérimenter sur de nombreux sites : celle des « 3C ».

Ces « 3C » sont les suivants :

- Contenu éditorial : parce que tout part de là. Un bon contenu, écrit pour les internautes tout en étant pensé – dans une certaine mesure – pour les moteurs, est primordial.

- Code HTML : car il doit être optimisé et permettre de mettre en exergue votre (excellent) contenu éditorial en le rendant, là aussi, réactif aux critères de pertinence des moteurs de recherche.
- Conception : parce qu'un site bien conçu doit proposer une « journée portes ouvertes » aux spiders des moteurs au travers d'une indexabilité sans faille.

Si on prend les 3C à l'envers, on peut dire que :

- les spiders doivent pouvoir accéder à toutes vos pages facilement et sans obstacle (Conception) ;
- une fois les pages trouvées, les moteurs doivent pouvoir lire leur code HTML et l'analyser facilement afin d'en extraire ce qui les intéresse le plus, le contenu éditorial (Code) ;
- une fois le contenu textuel trouvé, ils doivent le « comprendre » et pouvoir identifier aisément de quoi parle la page, quel est sa thématique principale (Contenu).

La règle des 3C fonctionne donc dans les deux sens !

Nous allons aborder ces trois concepts sous la forme de « mémentos » ou « pense-bêtes » afin de vous aider à ne rien oublier dans le cadre de la création – ou de la refonte – de votre site... Une bonne façon également de bien réviser tout ce qui a été dit dans les pages précédentes...

Contenu éditorial : tout part de là !

Votre contenu éditorial – ce que certains appellent le « text appeal » (*http://s.billard.free.fr/referencement/index.php?2006/11/30/319-optimisation-du-contenu-travaillez-votre-text-appeal*) est-il optimisé pour les moteurs de recherche ? Ce contenu éditorial est effectivement la source du trafic généré par la longue traîne (voir chapitre 3) qui va représenter près de 80 % du trafic « moteurs » total. Raison de plus pour bien le penser afin qu'il soit le plus réactif possible aux critères de pertinence des outils de recherche, sans jamais oublier toutefois que vous écrivez pour que vos textes soient lus avant tout par des internautes... Voici les questions qu'il faut se poser à leur sujet.

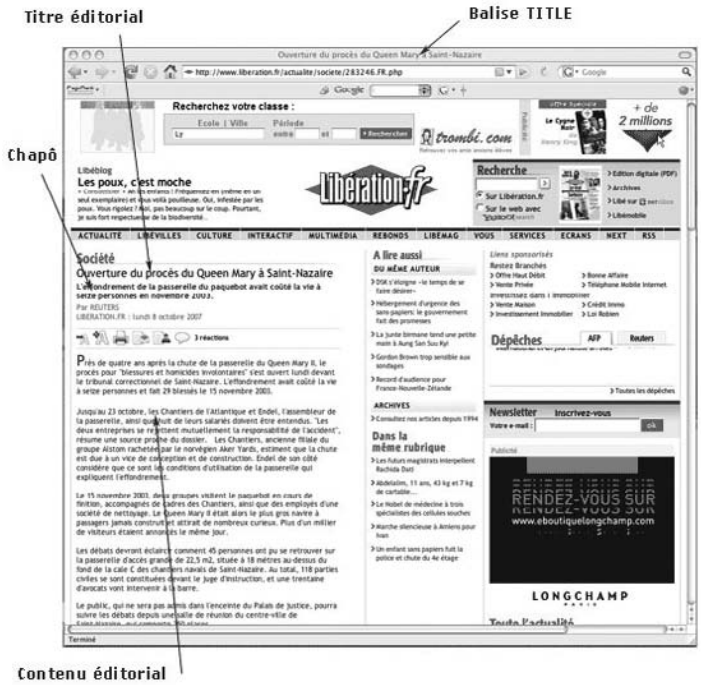
Contenu>Titre éditorial

Dans ce paragraphe, nous parlons du titre du contenu éditorial et non pas du contenu de la balise `<title>` de la page web (voir figure 10-1)...

- Le titre éditorial de la page est-il descriptif (factuel) et contient-il des mots-clés importants pour définir le contenu de la page ?
- Contient-il environ 5 à 7 mots descriptifs ?
- Est-il inséré dans une balise `<h1>` ?
- Y a-t-il une seule balise `<h1>` dans votre page (ce qui est le plus logique) ?

Figure 10-1

Différentes zones dans une page web



Contenu>Premier paragraphe du texte (chapô)

- Le premier paragraphe (deux à trois premières phrases, 100 premiers mots) du texte éditorial contient-il les mots-clés déjà présents dans le titre éditorial ? En propose-t-il d'autres, descriptifs de son contenu ?
- Ce premier paragraphe fait-il partie des 100 premiers mots de la page telle que lue par les spiders ?
- Les mots importants sont-ils mis en exergue (gras – balise `<strong>` – notamment) ?
- La mise en exergue (gras) est-elle insérée dans le code HTML lui-même (méthode compatible avec le référencement) ou dans les feuilles de styles (ce qui la rendra invisible pour les moteurs) ?
- Est-il éventuellement placé dans une balise `<h2>` ?

Contenu>Texte éditorial

- Les autres niveaux de titres éditoriaux sont-ils gérés par des balises `<hn>` (logiquement, de `h3` à `h6`) ?
- Si certains mots, contenus dans le texte, sont d'une extrême importance pour vous, en avez-vous calculé leur indice de densité grâce à des outils comme Outiref (*http://*

www.outiref.com/) ou Keyword Density (*http://www.seochat.com/seo-tools/keyword-density/*) ? Cet indice de densité est-il aux alentours de 2 à 3 % et en tous les cas inférieur à 5 % ?

- Avez-vous pensé à « disséminer » dans votre texte plusieurs occurrences et formes des mots importants : singulier/pluriel, féminin/masculin, accentuée ou non, etc. (exemple : hôtel, hotel, hôtels, hotel, hoteliere, etc.) ?
- Avez-vous fait en sorte que votre page soit le plus « monothème » possible et à ne pas disperser son contenu éditorial (ce qui impliquerait, alors, de créer plutôt des pages différentes pour chaque thème traité) ?
- Le texte éditorial est-il résumé par le titre éditorial et la balise `<title>` de la page ? Ces deux dernières zones sont-elles décrites en 200 caractères au maximum dans la balise meta `description` ?
- Le texte comporte-t-il des zones « descriptives » (en dehors des « effets de style », humour, nuances, etc.) pouvant être analysées sémantiquement par des robots pour comprendre « de quoi parle la page » ?
- Les règles journalistiques (5W : *who, what, where, when, why* et 2H : *how, how much*) sont-elles respectées pour fournir le plus d'informations possible aux spiders (et aux internautes) ?
- L'article ou contenu éditorial comporte-t-il plus de 100, ou mieux 200, mots descriptifs ?
- Lors de la rédaction, vous êtes-vous mis à la place du lecteur en utilisant ses propres mots, ou ceux qu'il emploierait pour rechercher sur un moteur, une page comme la vôtre ?
- Avez-vous fait une étude préalable sur les mots-clés les plus souvent utilisés par les internautes sur cette thématique, grâce au générateur de mots-clés de Google (voir chapitre 3) ? Avez-vous inséré ces mots importants dans votre texte ?
- N'avez-vous pas « suroptimisé » la page pour les moteurs, rendant son contenu éditorial complexe à lire pour les internautes (qui restent les lecteurs principaux de vos contenus) ?

Contenu>Liens textuels

- La partie éditoriale de vos pages contient-elle des liens sortants vers des pages (internes ou externes) du même domaine sémantique dans le cadre d'une rubrique « Pour en savoir plus » ?
- Avez-vous également ajouté des liens (notamment internes) au sein de votre texte, zone la plus importante pour les moteurs de recherche ?
- Les liens sont-ils textuels (toujours préférables à des liens images ou JavaScript) ?
- Les textes des liens sont-ils en rapport avec la page distante pointée (éviter les intitulés de type « Voir la suite » ou « Cliquez ici ») ?

Code HTML : les grands classiques

Bien sûr, une fois votre contenu bien « calibré », il va vous falloir mettre en place un code HTML optimisé pour les moteurs de recherche. C'est, on peut dire, l'enfance de l'art du domaine, historiquement parlant, mais cette phase reste aujourd'hui essentielle... Vous trouverez dans ce paragraphe quelques répétitions avec certaines recommandations édictées dans la partie « Contenu », mais nous avons préféré « enfoncer le clou » pour être sûr que tous les points étaient bien pris en compte. Veuillez donc nous en excuser par avance...

Code HTML>Header

- Votre code commence-t-il par les zones suivantes, ce qui sera garant d'une bonne prise en compte de la balise `<title>` par les différents moteurs de recherche ?

```
<html>
<head>
<title>Titre de votre page</title>
```

- La balise `<title>` contient-elle de 7 à 10 mots descriptifs ?
- Sa structure est-elle de la forme :

```
<title>[Contenu] - [rubrique] - [Source]</title>
```

  où `[Contenu]` reprend le titre éditorial (balise `<h1>`), `[Rubrique]` la rubrique de la page (pour les pages internes) et `[Source]` le nom ou la marque du site ?

- Les balises meta `keywords` (peu utiles aujourd'hui pour le référencement) et `description` (utiles pour mieux maîtriser la façon dont les moteurs affichent votre site dans leurs résultats) sont-elles prévues pour recevoir un contenu strictement en rapport avec le contenu de la page ?
- La balise meta `description` est-elle présentée sur 150 à 200 caractères ?
- Son contenu donnera-t-il « envie de cliquer » lorsqu'il sera repris dans les pages de résultats des moteurs ?
- Chaque page présente-t-elle un couple `<title>`/meta `description` spécifique de son contenu ?
- Votre header propose-t-il une balise meta `language` de la forme suivante ?

```
<meta http-equiv="content-language" content="fr">
```

  Cette balise est souvent lue par les moteurs, aussi il peut être intéressant de l'indiquer, notamment si vos pages sont dans une langue spécifique ou si l'ajout de mots dans une autre langue peut induire le moteur en erreur (cas d'un site en français qui vend du vin italien, par exemple).

- Idem pour les balises meta éventuelles :

```
<meta name="keywords" lang="fr" content="contenu de la balise en français">
<meta name="keywords" lang="en" content="contenu de la balise en anglais">
```

- Avez-vous prévu d'éventuelles balises meta `robots` pour indiquer certaines actions (indexation, suivi des liens) aux spiders des moteurs, notamment sur des pages de test ?
- Le codage des lettres accentuées a-t-il été pris en compte ? Si les pages de votre site sont codées en ISO-8859-1 (balise meta : `<meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1" />`) ou si vos pages sont en UTF-8 (balise meta : `<meta http-equiv="Content-Type" content="text/html; charset=utf-8" />`), les lettres accentuées doivent être codées de façon cohérente avec le codage choisi au préalable ainsi que pour le serveur web ou la base de données éventuellement utilisée. Si vous n'êtes pas sûr de vous, codez ces lettres en HTML (`&eacute;` pour le « é », par exemple).
- Externalisation des éléments CSS et JavaScript : plus la taille de la page est limitée à la portion de contenu textuel, plus la page est réactive. Donc CSS et JavaScript doivent être le plus possible appelés à l'aide de fichiers externes à la page. Vous externalisez également ainsi toute erreur possible d'analyse de votre code HTML par les moteurs. Exemples :

```
<link title="styles abondance" type="text/css" rel="stylesheet" href="styles-homepage.css">
<script language="javascript" src="http://www.abondance.com/js/scripts.js"></script>
```

Un header bien optimisé !

Voici un exemple de « header » HTML compatible avec les moteurs de recherche :

```
<html>
<head>
<title>R&eacute;f&eacute;rencement et moteur de recherche : toute l'info et
l'actu sur le référencement avec Abondance</title>
<meta name="description" content="Abondance d'infos sur le r&eacute;
f&eacute;rencement et les moteurs de recherche : description des moteurs,
actualit&eacute;, faqs, outils d'audit, m&eacute;thodologie de r&eacute;
f&eacute;rencement, articles, offres d'emploi, bibliographie, etc.">
<meta name="keywords" content="referencement, abondance, moteur de recherche,
recherche d'information, r&eacute;f&eacute;rencement">
<meta http-equiv="content-language" content="fr">
<meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1">
<link title="styles abondance" type="text/css" rel="stylesheet" href="styles-
homepage.css">
<link rel="shortcut icon" href="http://www.abondance.com/Bin/favicon.ico">
<script language="javascript" src="http://www.abondance.com/js/scripts.js">
</script>
</head>
</html>
```

- Si vous utilisez des frames, avez-vous fait tout ce qu'il fallait pour qu'elles soient bien prises en compte par les moteurs de recherche (voir chapitre 7) ? Idem pour le Flash et tous les autres critères freinants étudiés dans ce chapitre...

Code>Contenu Textuel

- Vos pages sont-elles prévues pour proposer au moins 200 mots en texte visible ?
- Avez-vous la possibilité (ou donnez-vous la possibilité aux rédacteurs) de créer/modifier des liens et le texte des liens affichés dans les pages ?
- Votre code HTML est-il prévu pour ne pas être surchargé en commentaires (non lus par les moteurs), ce qui peut en revanche alourdir le poids (en Ko) des documents ? Pas de balises commentaires (`<!-- -->`) donc, autres que les informations nécessaires au balisage utile au développement...
- Si vous utilisez du code JavaScript, avez-vous vérifié qu'il était compatible avec le fonctionnement des spiders des moteurs de recherche (voir chapitre 7) ?
- Avez-vous passé vos pages en test sur le site Spider Simulator (*http://www.spider-simulator.com/*) ou les avez-vous regardé dans la version « En cache » de Google avec l'option « Version en texte seul » afin de visualiser la façon dont le moteur comprend vos contenus ?
- Vos menus peuvent-ils être lus par les robots des moteurs (sont-ils lisibles dans la version en cache textuelle) ?
- Avez-vous vérifié, notamment grâce aux deux outils ci-dessus, que la mise en page de vos pages n'empêche pas le texte le plus important d'être proposé en premier (le « plus haut » possible) aux moteurs de recherche ?
- Utilisez-vous les div CSS, qui s'attachent à l'essentiel (le contenu) plutôt que les tableaux qui alourdissent la taille d'une page ?

Code>Mise en exergue des mots importants de la page

- Vos pages sont-elles prévues pour recevoir un titre (balise `<title>`) ? Et oui, on voit de tout sur le Web... Pourrez-vous le modifier au quotidien qu'elle que soit la page du site ?
- Les balises `<h1>`, `<h2>`, etc., sont-elles bien redéfinies dans les feuilles de styles pour apparaître de façon « agréable » dans un navigateur ?
- Vos images contiennent-elles des options `alt` dans les balises `<img>` ? Ces options textuelles contiennent-elles du contenu intelligible et descriptif pour les moteurs de recherche ? Comment les rédacteurs de contenu auront-ils accès à cette zone ?
- Les mots important décrivant la page sont-ils indiqués en gras ? Le gras est-il présent dans le code HTML, sous forme de balises `<strong>`, ce qui permettra aux moteurs de les prendre en compte comme telles, ou dans les feuilles de styles (que ne lisent pas les moteurs) ?
- Avez-vous vérifié que vous ne disposez aucun contenu caché dans vos pages, que l'internaute voit tout ce que le spider voit et vice versa ?

Code>Homogénéisation des différentes « zones chaudes »

- Avez-vous pensé l'organisation de votre code HTML pour que les « zones chaudes » suivantes soient toutes cohérentes au niveau de leur contenu, ce qui est essentiel pour un bon référencement ?
  - balise `<title>` ;
  - balise meta `description` ;
  - balise meta `keywords` (si vous l'affichez dans votre code) ;
  - titre éditorial (dans une balises `<h1>`) ;
  - début du texte visible de votre page (le premier paragraphe, les deux ou trois premières phrases, les 50 à 100 premiers mots).
- Si vous utilisez un CMS (*Content Management System*), l'avez-vous configuré pour que cette homogénéité se fasse de façon quasi automatique ?

Code>Compatibilité W3C

- Avez-vous passé les pages les plus importantes de votre site web, et notamment sa page d'accueil, au validateur de code du W3C (*http://validator.w3.org/*) ? Ont-elles passé ce test sans problème ? Avez-vous corrigé les éventuelles erreurs signalées qui pourraient gêner l'analyse de votre code par les moteurs ?
- Avez-vous testé le site une fois conçu avec au moins Firefox et Internet Explorer pour vérifier que tout s'affiche bien ? Si des éléments ne s'affichent pas, sous Firefox notamment, il y a de grandes chances que Google ne les voie pas non plus.
- Idéalement, votre code présente-t-il tous les atouts et critères d'accessibilité (*http://www.accessiweb.org/*) ? En effet, accessibilité et optimisation pour les moteurs de recherche sont souvent des concepts assez proches...

Bref, votre code est-il :

- parfaitement compatible avec toutes les contraintes induites par les moteurs ?
- optimisé par rapport aux critères de pertinence de ces mêmes moteurs ?

Si oui, et si vous avez déjà appliqué la première règle édictée auparavant, vous êtes sur la piste d'un excellent référencement. Il ne vous reste donc plus qu'à suivre nos conseils pour le troisième « C », la Conception...

Conception : l'essentielle indexabilité

Comment imaginer pour son site une structure qui soit totalement compatible avec son exploration par les spiders ? Comment mettre en place un réseau de liens aidant ces robots à obtenir une meilleure compréhension du maillage de votre source d'informations ? Comment faire pour éviter tout obstacle technologique freinant ou bloquant pour les moteurs ? Bref, nous allons lister ici, toujours sous la forme d'un

« mémo », une suite de « bonnes pratiques » pour mettre en ligne un site qui sera 100 % *spider friendly* dès son lancement...

Conception>Structure du site

Le premier des points à inspecter est la structure du site : les robots peuvent-ils aller partout de façon efficace et découvrir toutes vos pages ? Voici quelques points à vérifier :

Conception>Structure du site>Fichier robots.txt

- Votre site contient-il un fichier `robots.txt` ?
- Son nom est-il bien orthographié (« r » minuscule, « robots » au pluriel) ?
- Est-il disponible à la racine de votre site (*www.votresite.com/robots.txt*) ?
- En cas d'utilisation de sous-domaines (*motclé.votresite.com*), chaque sous-domaine dispose-t-il de son propre fichier `robots.txt` (*actu.votresite.com/robots.txt*, *produits.votresite.com/robots.txt*) ?
- Les zones « interdites aux robots », si elles existent, sont-elles bien listées dans le(s) fichier(s) `robots.txt` ?
- Certains spiders moins importants sont-ils pris en compte ou interdits, si nécessaire, ou si vous avez remarqué dans vos statistiques que leur venue gène votre serveur (trop de bande passante occupée lors du crawl, par exemple) ?
- Votre fichier `robots.txt` indique-t-il l'URL de votre fichier Sitemap (fonction `Autodiscovery`) ?
- Avez-vous vérifié la syntaxe de votre fichier `robots.txt` grâce à un outil disponible en ligne ?

Conception>Structure du site>Fichier Sitemap

Les cinq moteurs de recherche principaux se sont mis d'accord sur un format de fichier, nommé « Sitemap », décrit à l'adresse *http://www.sitemaps.org/*, et qui fournit aux moteurs un « plan du site », au format TXT ou, mieux XML, indiquant aux spiders, pour chaque page de votre site son URL, sa priorité d'indexation (0 à 1), sa fréquence de mise à jour (de `always` à `never` en passant par `hourly`, `weekly`, etc.) et sa date de dernière modification (voir chapitre 8).

Ce fichier est aujourd'hui hautement recommandé pour obtenir une bonne indexation quantitative de votre site. Voici la liste des questions qu'il faut se poser au sujet de vos fichiers Sitemaps :

- Votre site propose-t-il un fichier Sitemap ? Comme pour le fichier `robots.txt`, si vous utilisez des sous-domaines, n'oubliez pas de proposer un fichier Sitemap par sous-domaine.
- Le fichier Sitemap recense-t-il de façon exhaustive toutes les pages de votre site ?

- Les indications qu'il fournit pour chacune d'elles (date de dernière modification, fréquence de mise à jour) sont-elles conformes à ce que le robot découvrira de lui-même s'il parcourt votre site en suivant les liens mis à sa disposition ? Il est important que les informations que vous délivrez dans le fichier Sitemap soient le reflet de la réalité et soient cohérentes par rapport aux voies d'exploration de votre site par les robots...
- Avez-vous mis en place une procédure simple de gestion de votre fichier Sitemap pour les zones dynamiques, dont les informations changent souvent (nouvelles pages, modifications des pages existantes) ? Le but sera de fournir aux moteurs les données les plus fraîches possible...
- Avez-vous indiqué l'adresse du Sitemap dans le fichier `robots.txt` (voir paragraphe précédent) ?
- Avez-vous soumis votre fichier Sitemap à Google, Yahoo! et Bing au travers de leurs outils pour webmasters ? L'outil de Google, notamment, vous fournira des statistiques de visites ainsi que des diagnostics d'erreurs éventuelles sur vos fichiers, bien utiles parfois...

Conception>Navigation

Votre site doit proposer une opération « portes ouvertes » aux robots des moteurs de recherche (pour les zones auxquelles ils ont accès, bien sûr...). Aussi, il est important de répondre aux questions suivantes :

- Vos liens seront-ils tous compris et suivis par les spiders (sous la forme `<a href="http://www.votresite.com/...">`) même s'ils sont écrits en langage JavaScript ?
- Chaque page de votre site est-elle accessible en trois clics au maximum à partir de la page d'accueil, ce qui garantira une meilleure prise en compte quantitative de vos pages ?
- Les liens situés à l'intérieur du contenu éditorial sont-ils facilement configurables lors de la saisie du contenu (texte du lien, URL de destination, etc.) ?
- Un plan du site proposant des liens textuels simples est-il disponible ? Permet-il aux spiders d'accéder à chaque page de votre site en trois clics au plus ?
- Avez-vous prévu des pages d'erreur en cas de renvoi de code 404, 403, etc. ? Ces pages contiennent-elles des liens vers les différentes zones de votre site, autant de points d'entrée pour l'internaute mais également pour les spiders ?

Conception>Obstacles technologiques éventuels

- Si votre site propose des animations Flash, avez-vous fait le nécessaire (zone `noembed`, emploi de la norme sIFR, pages HTML complémentaires, etc.) pour le référencer au mieux ?

- Si votre site contient du JavaScript, l'avez-vous rendu compatible avec les moteurs de recherche, notamment au niveau des liens affichés (voir précédemment) ?
- Si votre site est réalisé avec des frames (ce qui peut paraître étrange car rares sont aujourd'hui les sites web qui utilisent encore cette technologie), avez-vous fait ce qu'il fallait pour pallier les inconvénients de ce type de système ?

Pour tous ces points (et bien d'autres), rendez-vous au chapitre 7.

Conception>Intitulés des URL et redirections

- Si vos URL contiennent des caractères `?` et `&`, avez-vous mis en place des techniques de réécriture d'URL ?
- Vos URL proposent-elles des mots-clés descriptif du contenu des pages (*www.votre-site.fr/epicerie/condiments/exotiques/sel-guyane.html*) ? Et notamment le nom de domaine et le nom de fichier (premier et dernier intitulé de l'URL) ? Ces mots sont-ils séparés par un tiret (-) ?
- Si vous désirez que vos pages soient indexées par Google News (*http://news.google.fr/*), vos intitulés d'URL contiennent-ils trois chiffres au minimum, condition demandée par Google pour leur indexation dans son moteur d'actualités ?
- Les redirections éventuelles d'une page vers une autre sont-elles réalisées à l'aide de codes 301 ?
- Avez-vous évité de mettre en place une redirection sur votre page d'accueil ?
- Avez-vous évité les redirections 301 en cascade (à la suite les unes des autres sur une même page) ?
- Votre site utilise-t-il des identifiants de session qui apparaissent dans l'URL ? Ceci peut poser de nombreux problèmes à certains moteurs.

Conception>Pages web

Enfin, vos pages web, dans leur structure, sont-elles compatibles avec les moteurs ?

- Par exemple, ont-elles été conçues pour être monolingues ? Une langue donnée doit être affichée majoritairement dans une page web donnée pour être « compréhensible » par les robots.
- De même, les pages doivent être conçues pour être le plus « monothèmes » possible. Peu importe leur longueur (dans ce cas, c'est plus la densité des mots-clés qui jouera pour leur pertinence), mais il est important que chaque page cerne un point donné, quitte à multiplier les pages sur votre site. En d'autres termes, si vous avez 10 produits à présenter, n'hésitez pas une seconde : créez une page de présentation pour chaque produit et surtout pas une longue page qui les présente tous sur un même document. Vous devez être capable de décrire le contenu d'une page en 10 mots au maximum (ce sont les termes qui constitueront son titre).

Conception>Rubriquage

- Avez-vous pensé les différentes rubriques du site en fonction de ce que les internautes recherchent et pas en fonction de la structure de la base de données ou de notions techniques (comme cela arrive parfois) ?
- N'oubliez pas de donner des noms explicites à chaque zone thématique de votre site. Ces noms explicites seront ensuite repris dans les intitulés de liens, primordiaux aujourd'hui pour un référencement...

Conclusion

Tous les points visés dans ce chapitre touchent au contenu, au code et à la conception du site web. Ils sont capitaux car ils représentent des étapes essentielles, fondamentales, sur lesquelles il sera très difficile, voire impossible de revenir en arrière lorsque le site sera en ligne. Et Dieu sait si, par le passé, ce type d'erreur a été commis, nécessitant parfois une remise en question complète du site pour le rendre compatible avec les moteurs. Autant penser à ces points dès le départ, cela facilite bien les choses et permet d'éviter des surprises désagréables...

Bien sûr, une fois que tous les points évoqués dans les paragraphes précédents auront été validés, il ne vous restera plus (*sic*) qu'à trouver d'excellents backlinks et à partir « à la pêche aux liens ». Un nouveau travail commencera alors, souvent difficile, long et chronophage, mais tellement indispensable à un bon référencement. Qui a dit que l'optimisation d'un site pour les moteurs de recherche était un long fleuve tranquille ?

Le retour sur investissement : une notion essentielle

Vous avez suivi les différentes étapes listées au début de ce chapitre ? Parfait, vous avez donc avancé sur le chemin, parfois laborieux mais si passionnant, de l'optimisation de site pour les moteurs de recherche. Mais une fois tout ce travail effectué, il va vous falloir mesurer l'efficacité de votre labeur. Au cours des dernières années, la mesure de la qualité du référencement s'est basée sur plusieurs notions successives :

- Le positionnement : il s'agit de positionner dans les résultats du moteur certaines pages du site à référencer pour tel mot-clé ou telle expression (suite de mots-clés). Le référenceur fournit à son client des tableaux indiquant les mots-clés visés, les moteurs pris en compte et les positions obtenues. De nombreux logiciels ont par ailleurs été créés pour automatiser cette tâche et vérifier les différents positionnements acquis. Principal inconvénient de ce système : il ne donne aucun renseignement sur le trafic obtenu. Ainsi, vous pouvez obtenir 50 premières positions sur 20 moteurs de recherche différents, ce qui réjouira l'éditeur du site, mais sur des mots-clés que personne ne saisit, ce qui est moins agréable du point de vue des statistiques. De plus, de nombreux moteurs sont mis au même niveau, ce qui correspond assez peu à la réalité. Comment, en effet, comparer – notamment en France – un positionnement sur Google par rapport au même rang obtenu sur Francité ou Lokace (nous parlons ici d'une autre époque), voire Bing ou Yahoo! ?

De même, la notion même de positionnement est devenue, avec le temps, problématique et sujet à caution. En effet, tous les moteurs de recherche majeurs ont mis en place une stratégie de personnalisation de leurs résultats en fonction de l'internaute qui effectue la requête. Ainsi, pour un même mot-clé, plusieurs personnes pourront obtenir des résultats différents en fonction de plusieurs facteurs :

- sa localisation géographique ;
- la version du moteur utilisée (anglaise, française, espagnole) et la langue choisie pour l'interface utilisateur ;
- la langue de son navigateur ;
- son historique de recherche ;
- les données et préférences « SearchWiki » sur Google ;
- etc.

Si chaque internaute, à terme, affiche une page de résultats différente pour un même mot-clé, comment peut-on restreindre la qualité d'un référencement à la seule notion de positionnement, puisque le site sera affiché différemment pour chaque utilisateur d'un moteur ou presque ?

On pourrait donc penser que ce type de méthodologie n'est plus utilisé par les référenceurs. Il n'en est rien. De nombreuses sociétés de référencement fonctionnent encore ainsi. Ce qui prouve que cette tendance est loin d'être obsolète et qu'elle répond à une réelle demande. Mais il est vrai que la mouvance actuelle n'est plus au strict positionnement de pages web pour mesurer l'efficacité d'un référencement.

Quelques outils de mesure du positionnement

Il existe plusieurs logiciels permettant d'automatiser la surveillance des positionnements obtenus sur les moteurs. Voici les principaux d'entre eux, classés par ordre alphabétique (vous trouverez d'autres outils, notamment des sites web, en annexe) :

- Advanced Web Ranking – *http://www.advancedwebranking.com/*
- Agent Web Ranking – *http://www.agentwebranking.com/*
- Indexweb – *http://www.indexweb.fr/*
- Search Engine Commando – *http://www.searchenginecommando.com/*
- SEO WebRanking – *http://www.seowebranking.com*
- Trellian – *http://www.trellian.com/fr/swolf/*
- Yooda SeeUrank – *http://www.yooda.com/*

- La vérification d'un référencement s'est alors tournée vers la notion de trafic généré. On était déjà plus proche, ici, de la réalité. Il ne s'agit plus « seulement » d'être bien positionné de façon plus ou moins artificielle, mais de créer un trafic (ou de l'augmenter s'il existe déjà) en provenance des outils de recherche. Les stratégies se modifient et s'affinent. On prend alors mieux en compte les différences entre moteurs

(autant se focaliser sur les 10 outils de recherche majeurs plutôt que sur les « petits », notamment pour le travail effectué manuellement). Les baromètres (Première Position/Xiti, adoc, voir chapitre 3) ont également beaucoup œuvré dans ce sens en donnant une bonne hiérarchie d'importance aux outils de recherche actuels dans le cadre du trafic moyen généré sur un site. Plusieurs stratégies peuvent alors être mises en place pour améliorer un trafic déjà existant. On peut essayer de demander une nouvelle catégorie pour son site sur les annuaires afin d'augmenter le trafic issu de Yahoo! Directory ou de l'Open Directory. Ou encore tenter un bon positionnement sur Google, Yahoo! ou Bing sur des mots-clés porteurs mais encore peu concurrentiels. On pourra ainsi rapidement s'apercevoir qu'une 11e position sur un mot-clé donné peut générer un trafic bien plus important qu'un 9e rang sur une autre requête… L'approche est bien plus fine (même si elle est un peu plus contraignante pour le client, qui doit mettre en place un outil d'analyse d'audience sur ses pages), mais il existe encore un inconvénient : elle ne prend pas en compte l'aspect qualitatif du trafic. Elle ne mesure qu'un aspect strictement quantitatif.

- Vient donc, de façon presque naturelle, la notion de retour sur investissement, très souvent appelée ROI (pour *Return on Investment*). Le but est de mesurer la qualité du référencement effectué en tenant compte de deux critères majeurs :

  – D'où vient l'internaute qui arrive sur mon site ?

  – Qu'y fait-il ?

  En d'autres termes : si mon site est marchand, les internautes achètent-ils ? Si je désire recruter des prospects, les visiteurs remplissent-ils le formulaire qui leur est proposé ? Si je désire présenter un produit, les visiteurs affichent-ils en majorité les pages qui le décrivent ?…

En effet, l'heure n'est plus à la création de trafic stérile (le terme n'est pas obligatoirement péjoratif), qui pourrait être intéressant dans le cadre d'un modèle économique basé sur l'affichage de bannières publicitaires au CPM. Ce modèle est de moins en moins répandu sur les sites francophones. Aujourd'hui, le responsable de site désirera mesurer la qualité de son trafic en ayant la meilleure vision possible de ce que le visiteur y a fait, et surtout, en sachant s'il y a effectué une action « profitable » pour le site.

Différents types de calcul du retour sur investissement

Cette notion de « profitabilité » peut être très vaste. Il ne s'agit pas uniquement d'un accroissement du chiffre d'affaires dû à un achat en ligne, qui sera certainement le but d'un site de commerce électronique. Les objectifs peuvent être très variés. En voici quelques-uns :

- **Vente en ligne.** L'éditeur du site veut que l'internaute achète chez lui ou prépare un achat pour une prochaine visite. Dans ce cas, le critère retenu est le chiffre d'affaires généré.

Le ROI sera alors égal à :

ROI = chiffre d'affaires généré par des visiteurs issus des outils de recherche/coût du référencement

Ce calcul peut être effectué en fonction du chiffre d'affaires, de la marge brute, de la marge nette, etc., en fonction des besoins de l'éditeur du site. Selon le paramétrage de son système de tracking, un client qui achète un mois après être venu sur le site pour la première fois sera pris en compte ou non.

- **Notoriété active.** L'éditeur veut que l'internaute vienne sur son site pour y passer du temps. L'objectif se situe au niveau de la marque, d'une bonne gestion de la notoriété de l'entreprise et de sa visibilité. Dans ce cas, le ROI sera calculé en fonction du temps passé par l'internaute sur le site, du nombre de pages vues par visite, etc.

  Exemple : Renault sort une nouvelle voiture et il veut que 10 000 internautes/mois se connectent sur la page qui la présente.

  Le calcul de son ROI s'effectue sur la base suivante :

  ROI = durée des visites des internautes issus des outils de recherche/coût du référencement

  ou :

  ROI = nombre de visites d'internautes issus des outils de recherche/coût du référencement

- **Notoriété passive.** L'éditeur veut que l'internaute voie sa marque sans pour autant vouloir amener l'internaute sur son site.

  Son ROI sera calculé en fonction du nombre d'affichages/jour, affichages/semaine ou affichages/mois. Cela peut, aussi, correspondre à une campagne de liens sponsorisés.

  Exemple : un constructeur automobile veut que sa marque soit affichée dans les 5 premiers résultats du mot « pick-up » sans être leader car il n'a pas un modèle récent dans cette catégorie de véhicule.

  Le calcul de son ROI s'effectue sur la base suivante :

  ROI = nombre de fois où le lien est affiché dans les pages de résultats/coût du référencement ou de la campagne de liens sponsorisés

- **Recrutement/actions opt-in.** L'éditeur veut que l'internaute effectue une action sur son site. Une action peut être le remplissage d'un formulaire, l'abonnement à une newsletter, l'inscription en tant que membre, une demande d'information sur un produit/service…

  Dans ce cas, le ROI sera calculé sur la base suivante :

  ROI = nombre d'actes effectués par des visiteurs issus des outils de recherche/coût du référencement

Il peut exister bien d'autres possibilités de calcul de ce ROI, en fonction du type de site et des informations, ressources, produits ou services qu'il propose.

Ce type de calcul peut également être effectué pour de nombreuses actions de promotion :

- bannières publicitaires ;
- référencement ;
- positionnement publicitaire (liens sponsorisés de type Microsoft adCenter, Google Adwords, Yahoo! Search Marketing) ;
- échange de liens : quel site pointant vers vous génère le trafic le plus qualifié ?
- insertion de liens dans une newsletter ;
- sponsoring de zones à l'intérieur d'un site (exemple : un fleuriste pour la Saint-Valentin) ;
- etc.

La mise en place de liens de tracking

Pour effectuer ces calculs, les outils de mesure du trafic – et donc du ROI – actuels se basent sur les *URL referrers*, c'est-à-dire l'URL de la page d'où vient l'internaute arrivant sur votre site, que votre serveur reçoit à chaque nouvelle visite.

Comment ces statistiques sont-elles calculées ? Sur quelles bases les informations sont-elles traitées ? Raisonnons sur un exemple : un internaute va sur Google et tape les mots-clés « veille technologique ». La page de résultats de Google s'affiche à l'adresse *http://www.google.fr/search?rls=ig&hl=fr&q=veille+technologique&btnG=Re-cher-che+Google&meta=&aq=f&oq=.*

Si l'internaute clique sur un des liens proposés, il arrrive sur le site affiché. Ce serveur reçoit alors l'URL referrer correspondant à l'adresse de la page de résultats de Google, indiquée ci dessus.

Cette URL contient de nombreuses informations intéressantes :

- le site de provenance (*Google*) ;
- la langue utilisée (*hl=fr*) ;
- la requête demandée sur le moteur (*q=veille+technologique*).

Les sociétés d'analyse d'audience insèrent des balises HTML et des scripts spécifiques dans les pages des sites qu'elles auditent. Ces bouts de programmes permettent de récupérer les URL referrers à chaque venue d'un visiteur. Les logiciels d'analyse de logs lisent directement ces données dans les fichiers logs (la mémoire des connexions) du serveur en question.

Mesure de l'efficacité d'un référencement au travers de la longue traîne

Lors des formations que nous animons, en France et ailleurs, une question revient souvent sous la forme de l'évaluation et de la mesure de l'efficacité d'une stratégie de référencement. Que le travail d'optimisation et de référencement naturel soit effectué en interne ou *via* un prestataire, la question revient toujours comme une antienne : Comment mettre en place un tableau de bord valable et efficace pour mesurer le travail effectué et être sûr de la progression (ou de la régression d'ailleurs...) de ses résultats en termes de visibilité sur les moteurs de recherche ?

À partir de 2004, le concept de « longue traîne » (voir chapitre 3) est arrivé dans le domaine du commerce électronique et son adaptation aux moteurs de recherche a vu le jour à partir de 2005/2006 en France.

Aussi, à la lumière de cette longue traîne, nous allons essayer, dans les paragraphes qui viennent, de définir une stratégie de mesure des efforts et des résultats du référencement d'un site web. Il ne s'agit que de propositions et vous pouvez vous en inspirer et les adapter à vos besoins, vos attentes, votre site, votre marché, bref, à votre propre vision de votre activité web...

Tête et queue de longue traîne

Tout d'abord, rappelons que le concept de longue traîne scinde le trafic issu des moteurs de recherche en plusieurs parties :

- La « tête », qui représente le plus souvent 20 % environ du trafic, et qui est constituée par le trafic généré par un nombre faible (quelques dizaines tout au plus) de requêtes « best sellers » (ou « best clicker » ?) :
  - Votre marque et ses dérivés (URL du site, différentes formes des mots qui composent votre sigle, fautes d'orthographe et de frappe, etc.).
  - Des mots-clés génériques qui décrivent votre métier, votre domaine d'activité.
  - Éventuellement, des mots-clés « éphémères », très liés à l'actualité et qui ont drainé sur votre site beaucoup de trafic pendant quelques jours mais sans lendemains.
- La « queue » de la longue traîne (que nous nommerons le plus souvent « LT » dans la suite de ce chapitre), qui représente 80 % du trafic « moteur », sous la forme d'un très grand nombre (parfois plusieurs dizaines de milliers, voire plus) de requêtes qui génèrent, chacune indépendamment, très peu de visites sur votre site.

Nous vous engageons vivement à relire, si nécessaire, le chapitre 3 de cet ouvrage qui présente de façon approfondie le concept de LT.

On notera que, longtemps, les prestations des sociétés de référencement ne se sont intéressées qu'à la tête de la LT : « Donnez-nous une liste de mots-clés qui représentent votre site, nous allons tenter de vous positionner dessus. » Le travail de référencement n'était

donc réalisé que sur 20 % du trafic « moteur »... La prise de conscience du concept de LT a donc fait fortement évoluer les esprits à ce sujet !

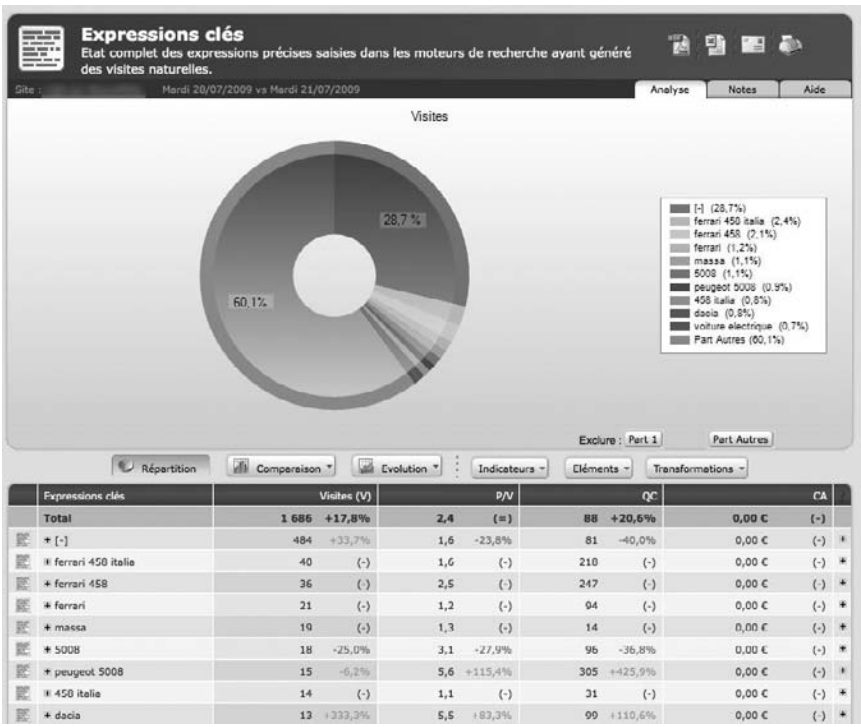
Une stratégie de référencement se doit donc, aujourd'hui, de prendre en compte cette LT en différenciant les deux sources de trafic :

- Stratégie « à l'ancienne », en optimisant certaines pages pour certains mots-clés définis au préalable et représentatifs de votre marque et de votre métier.
- Stratégie « Queue de LT », en donnant la meilleure visibilité possible au contenu éditorial des pages du site, qui « nourrit » ces 80 % de trafic « moteurs » indispensables.

Pour mesurer les retombées de cette stratégie, il sera logique, alors, de tenir compte de la même dichotomie des sources de trafic : tête et queue de LT. Voici comment...

Étape 1 – Différenciation des deux trafics

Dans un premier temps, il vous faut bien différencier les deux types de trafic généré : tête et queue de LT. Pour cela, vous allez vous armer des statistiques fournies par votre outil de mesure d'audience (XiTi, Weborama, eStat, Google Analytics, Wysistat, etc.), dont nous reparlerons plus loin dans ce chapitre.



Expressions clés	Visites (V)		P/V		QC		CA	
Total	1 686	+17,8%	2,4	(=)	88	+20,6%	0,00 €	(-)
+ [-]	484	+33,7%	1,6	-23,8%	81	-40,0%	0,00 €	(-)
+ ferrari 450 italia	40	(-)	1,6	(-)	210	(-)	0,00 €	(-)
+ ferrari 458	36	(-)	2,5	(-)	247	(-)	0,00 €	(-)
+ ferrari	21	(-)	1,2	(-)	94	(-)	0,00 €	(-)
+ massa	19	(-)	1,3	(-)	14	(-)	0,00 €	(-)
+ 5008	18	-25,0%	3,1	-27,9%	96	-36,8%	0,00 €	(-)
+ peugeot 5008	15	-6,2%	5,6	+115,4%	305	+425,9%	0,00 €	(-)
+ 450 italia	14	(-)	1,1	(-)	31	(-)	0,00 €	(-)
+ dacia	13	+333,3%	5,5	+83,3%	99	+110,6%	0,00 €	(-)

Figure 10-2

L'interface XiTi, par exemple, vous fournit nombre d'informations sur les mots-clés qui ont servi à trouver un site web.

Condition essentielle : cet outil doit vous fournir la **globalité** des statistiques de votre trafic « moteur », c'est-à-dire **tous** les mots-clés ayant servi à trouver votre site sur les différents moteurs de recherche du Web. Ce n'est malheureusement pas le cas de tous les outils de mesure d'audience du marché (certains ne donnent, par exemple, que le « Top 100 »). Si vous ne disposez pas de cette liste exhaustive, sachez qu'il vous sera difficile, voire impossible, de mettre en place la procédure expliquée dans la suite de ce paragraphe. Dans ce cas, il vous faudra peut-être envisager de changer d'outil d'analyse d'audience...

Notez que si vous pouvez disposer de ces chiffres pour chaque moteur de recherche (trafic généré par Google, par Yahoo!, par Bing, etc.), ce sera encore mieux et permettra des analyses plus fines.

Bref, vous partez donc d'un document Excel (ou autre), vous donnant, pour votre site, la liste des mots-clés ayant permis de trouver votre site sur les moteurs de recherche et, pour chacun d'eux, le nombre de visites générées. Exemple – semi-fictif – pour le site Abondance.com :

Tableau 10-1 Principaux mots-clés ayant généré au moins une visite sur le site Abondance.com en décembre 2008

Mots-clés	Nombre de visites générées en décembre 2008
abondance	4 533
google france	1 028
life magazine	1 020
abondance.com	1 002
google	968
google tv	809
lettre d information	800
google translate	785
robots.txt	652
geoportail	648
moteurs de recherche	528
olivier andrieu	444
lycos	421
miserable failure	412
moteur de recherche	102
teoma	100
meta keyword	98

Nota : pour le site Abondance.com, la liste complète contient environ 20 000 requêtes différentes chaque mois.

La première étape sera de différencier votre « tête » de votre « queue » (ok, ok, le premier qui fait une allusion quelconque sur cette phrase recevra un gage et sera condamné à reprendre la lecture de ce livre depuis le début...) de LT. En inspectant vos requêtes, vous allez vous apercevoir qu'à partir d'un certain classement, les mots-clés ne génèrent que peu de trafic. Pour le site Abondance.com, par exemple, nous estimons qu'en dessous de 100 visites générées par mois, le mot-clé fait partie de la « queue » de la LT, mais ce critère peut être différent pour chaque site et la barrière, sur certaines sources d'informations importantes, peut tout à fait être fixée à plusieurs milliers de visites. À vous, donc, de définir la « frontière » entre les deux zones de trafic.

Une fois ce travail effectué, faites les calculs des proportions de ces deux zones : statistiquement parlant, la tête représente la plupart du temps 20 % du trafic et la queue 80 %, bien que, là aussi, cela peut différer d'un type de site à un autre. Encore une fois, adaptez cette étape à votre propre site et réflexions en interne.

Récapitulatif de l'étape 1

Vous disposez donc maintenant :

- D'une liste complète, chaque mois, des mots-clés qui ont servi à trouver votre site sur les moteurs de recherche (si votre outil de mesure d'audience vous fournit cette liste par moteur de recherche, c'est encore mieux).
- Du pourcentage représenté par votre « tête » et votre « queue » de longue traîne.
- Vous pouvez passer à l'étape suivante...

Étape 2 – Tête de longue traîne : outils de positionnement et mesure du trafic

La deuxième étape est constituée par l'analyse, sur plusieurs mois, du classement des mots-clés représentatifs de votre « tête de longue traîne » : Sont-ils représentatifs ? Si vous avez travaillé sur certains mots-clés et expressions avec une société de référencement, les retrouvez-vous dans cette liste ?

À partir de cette analyse, vous vous poserez certainement plusieurs questions :

- Vous aviez imaginé au préalable certains mots-clés que vous ne retrouvez pas ici comme générateurs de trafic. Deux possibilités dans ce cas : soit votre site n'est pas bien optimisé pour ces requêtes, soit elles ne sont pas souvent saisies par les internautes sur les moteurs. L'utilisation du générateur de mots-clés de Google (*https://adwords.google.fr/select/KeywordToolExternal*) sera intéressante pour vérifier ce dernier point (voir chapitre 3). Si cet outil confirme que la requête est souvent demandée, c'est que votre site présente des lacunes à ce niveau en termes d'optimisation. À vous de les corriger...

- Certains mots-clés générateurs de trafic n'avaient pas été imaginés au départ : Quel est leur intérêt ? Est-ce une requête « porteuse » ? Dans ce cas, il faut renforcer l'optimisation de votre site pour elle et regarder son positionnement actuel dans les pages de résultats des moteurs de recherche. Est-ce une requête « accidentelle » ou « éphémère » ? liée à une actualité non récurrente ou n'ayant pas de rapport direct avec votre activité ? Dans ce cas, laissez tomber...

À vous de prendre les bonnes décisions sur la base de ces analyses. Dans tous les cas, il sera intéressant de constituer un ensemble de mots-clés et expressions constituant la trame de votre tête de longue traîne et contenant :

- Les mots-clés (marque et termes génériques) que vous désirez voir figurer dans votre « tête » et qui se retrouvent bien dans vos statistiques (trafic « acquis »).
- Les mots-clés (marque et termes génériques) que vous désirez voir figurer dans votre « tête » et que ne se retrouvent pas dans vos statistiques (trafic « espéré »).
- La troisième partie de votre tête de LT sera donc constituée par les mots-clés que vous ne désirez pas obligatoirement voir figurer dans votre « tête » (trafic « secondaire »). Par exemple, les mots-clés liés à l'actualité du moment.

Pour chacune de ces parties, il sera intéressant de suivre, grâce à un logiciel comme Agent Web Ranking, SeeUrank (voir encadré « Quelques outils de mesure du positionnement ») ou *via* les comptes-rendus fournis par votre prestataire de référencement, vos positionnements sur les divers moteurs de recherche pour ces termes et vérifier la progression, mois après mois, de chacun d'entre eux.

Vous pouvez aussi – et nous vous y encourageons fortement – aller plus loin en mesurant la qualité du trafic généré : combien de personnes viennent-elles sur mon site après avoir tapé « moteurs de recherche » sur Google ou Yahoo! et qu'y font-elles ? Quel est le taux de rebond constaté (fort taux de rebond = les internautes viennent et repartent aussitôt du site, sans aller plus loin) ? etc. Rappelons que le taux de rebond mesure le nombre de visiteurs partant du site en n'en ayant visualisé qu'une page.

En effet, voici une question souvent posée au sujet de la qualité des deux sources de trafic de la longue traîne : Est-ce que l'une génère un meilleur retour sur investissement que l'autre ? Pour l'instant, nous n'avons pas encore vu et lu d'études globales fiables à ce sujet. Mais vous pouvez tout à fait mesurer ce point sur votre site en configurant au mieux votre outil de mesure d'audience...

Récapitulatif de l'étape 2

L'analyse de votre trafic de « tête de longue traîne » vous a permis de différencier :

– Le trafic acquis : mots-clés identifiant votre marque ou votre métier, générant du trafic et pour lesquels vous êtes bien positionné. À vous de continuer vos efforts pour garder cet acquis...

– Le trafic espéré : mots-clés identifiant votre marque ou votre métier, générant du trafic et pour lesquels vous n'êtes pas encore bien positionné. À vous de mieux optimiser votre site pour ces mots-clés et de gagner des places.

☞

Récapitulatif de l'étape 2 *(suite)*

Pour les deux types de trafic ci-dessus, vous vous devez de suivre, grâce à des outils de suivi du positionnement, l'évolution de votre présence sur les moteurs de recherche.

– Le trafic secondaire : mots-clés éphémère (liés à l'actualité « chaude ») ou n'ayant pas de rapport avec votre activité, pour lesquels vous êtes bien positionné.

A priori, ce trafic étant éphémère ou inintéressant pour vous, il n'a pas besoin d'un suivi particulier...

Il est aujourd'hui important de mesurer également la qualité du trafic généré par ces mots-clés grâce à vos outils de mesure d'audience et leurs fonctionnalités de calcul du ROI. La plupart des outils majeurs de ce type permettent cela aujourd'hui.

Étape 3 – Queue de longue traîne : outils de mesure du trafic généré et de sa qualité

Parlons maintenant des 80 % de trafic (généralement constatés) représentés par la « queue » de votre LT, donc des milliers (voire plus) de requêtes générant chacune un faible trafic. Ici, ce sera le contenu de vos pages qui sera « moteur », qui « nourrira » cette source de trafic.

Dans ce cas, les outils de mesure du suivi du positionnement, de type Agent Web Ranking ou SeeUrank, ne vous seront que de peu d'utilité. L'intérêt de suivre le positionnement de votre site sur 20 000 mots-clés est faible et, de toute façon, irréalisable en pratique. Comment lire tous ces rapports ?

L'idée, alors, sera de travailler plutôt sur l'analyse du fichier renvoyé par votre outil de mesure d'audience de façon globale (Quels mots-clés ? Combien de visites pour chacun d'eux ?) :

- Quel est le volume du trafic généré pour cette queue de LT ? 80 % ? Vous êtes alors dans la taille statistiquement constatée (mais, là encore, cela peut varier selon le type de site web pris en compte). Dans tous les cas, il faut se poser la question de ce pourcentage et se demander s'il est « logique ».
- Le volume constaté augmente-t-il avec le temps ? Si par, exemple, votre queue de LT est représentée par N visites, quelle est la variation de N d'un mois sur l'autre ? Il sera ici très important de surveiller la croissance – ou non – de ce trafic d'un mois sur l'autre, notamment si vous avez fait des travaux d'optimisation sur votre site dernièrement, pour voir s'ils ont porté leurs fruits.
- Quels sont les mots-clés identifiés ? Il ne s'agit pas d'inspecter plusieurs milliers de termes en détail mais de regarder, d'un coup d'œil rapide, si ces mots-clés ont, globalement parlant, un rapport avec votre site. Si non, il peut y avoir un souci de compréhension de votre source d'informations par les moteurs de recherche. À vous de revoir vos contenus et l'optimisation de vos pages HTML présentant ces contenus (balises `<title>`, `<h1>`, `<strong>`, etc.)...
- Il peut être intéressant, ici aussi, de mesurer la qualité du trafic généré. En effet, il ne servira pas à grand--clés qui « transforment » le mieux est donc essentielle !

À vous de configurer votre outil d'analyse d'audience (s'il propose des fonctionnalités de mesure du ROI) pour entrevoir et mesurer la qualité du trafic généré par votre « queue de LT ». Vous pourrez également identifier, ici, quels sont les moteurs qui offrent la meilleure qualité de trafic. Par exemple, il est possible de s'apercevoir que le trafic émanant de Bing est beaucoup plus faible en quantité, mais bien mieux qualifié que celui de Yahoo! ou de Google. Il vous faudra donc peut-être intensifier vos efforts de référencement et de visibilité sur ce moteur. C'est un exemple...

Récapitulatif de l'étape 3

L'analyse de votre trafic de « queue de longue traîne » vous a permis de :

– Comprendre quel était le volume généré par cette source de trafic (l'augmentation ou la diminution de ce volume proprement dit ainsi que son pourcentage comparé à la « tête » de l'étape 1).

– Analyser les mots-clés de cette « queue de LT » : Font-ils partie, en général, de votre univers sémantique ou sont-ils « à côté de la plaque » ?

– Calculer la qualité du trafic généré au travers du ROI correspondant : Que font, une fois arrivés sur votre site, les internautes qui ont saisi des mots-clés de « queue de LT » ?

Conclusion : la longue traîne, le futur du positionnement

Au vu de ce qui est décrit dans ce chapitre, il semble clair que la mesure professionnelle de l'efficacité d'un référencement ne peut plus, en 2009, se contenter de regarder et vérifier les positionnements d'un site pour quelques dizaines de mots-clés sur quelques moteurs de recherche majeurs. Cela peut bien sûr correspondre à des besoins « légers » (vous n'avez pas de temps à consacrer à ce type d'analyse, par exemple) ou à des besoins de *benchmarking* (vérifier votre visibilité par rapport à vos concurrents sur un ensemble de mots-clés précis), mais les résultats qui en résulteront ne pourront être que parcellaires.

L'idéal, dans cette stratégie de mesure des résultats de votre stratégie, sera bien de prendre en compte les différents « ingrédients » entrevus dans ce paragraphe :

- différenciation du trafic de « tête » et de « queue » de la LT ;
- analyse – par logiciels de suivi du positionnement – des différentes composantes du trafic de « tête » ;
- analyse globale du trafic de « queue » ;
- mesure de la qualité du trafic généré par les différentes parties de la LT.

À vous, bien sûr, d'adapter la proposition de procédure décrite dans ces pages à vos besoins et à vos attentes. Sur la base proposée, ce sera donc à chacun de faire ses choix en fonction de sa vision et du temps qu'il peut consacrer à cette mesure, bien entendu... Mais la mise en place d'un tableau de bord de ce type peut, nous en sommes certains, vous faire gagner beaucoup de temps et d'énergie, et surtout éviter de faire partir votre stratégie de référencement sur des voies de garage...

Mesure d'audience : configurez bien votre logiciel

En règle générale, on utilisera, pour mesurer le ROI d'une action de promotion, un logiciel d'étude de logs, ou d'analyse d'audience à base de marqueurs (de type « Site Centric »).

Quelques liens pour en savoir plus sur la mesure d'audience

Pour avoir plus d'informations sur ce type d'outil, voici quelques liens qui devraient vous aider :

– *http://www.lesmoteursderecherche.com/ressources_maintenance_stat.htm*

– http://www.journaldunet.com/solutions/dossiers/audience/sommaire.shtml

– *http://www.journaldunet.com/dossiers/audience/index.shtml*

– *http://www.journaldunet.com/0103/010306audiencepanel.shtml*

– *http://www.journaldunet.com/0103/010308audienceoutils.shtml*

– *http://www.commentcamarche.net/contents/web/mesure-audience.php3*

Si vous utilisez ce type de logiciel : Êtes-vous sûr qu'il prenne bien en considération les outils de recherche francophones comme google.fr, voila.fr, orange.fr ou encore aol.fr ? Souvent, ces logiciels sont fournis par défaut sous une configuration correspondant aux outils de recherche anglophones, voire américains, majeurs, mais ont tendance à oublier les outils francophones. Pour prendre en compte ces derniers, il vous faudra alors configurer vous-même le logiciel.

Vérifiez donc bien cet état de fait pour être sûr que la globalité du trafic provenant des outils de recherche a été évaluée.

Logiciels de suivi du ROI

Certains logiciels d'analyse d'audience (analyse de logs, type Webtrends ou mise en place de tags HTML dans les pages pour analyse ultérieure, de type Xiti, Weborama ou eStat) proposent des versions spécifiques ou des fonctionnalités supplémentaires très performantes pour ce qui est du suivi net et précis du parcours d'un visiteur, depuis sa provenance jusqu'à son action sur le site, permettant de calculer de la façon la plus fine possible le ROI.

Les outils tels que XiTi (*http://www.atinternet.com/*), Weborama (*http://weborama.com/2/*), eStat (*http://www.estat.com/*) ou Wysistat (*http://www.wysistat.net/*), entre autres, servent au tracking des campagnes de liens sponsorisés mais aussi à celui du référencement naturel. N'hésitez pas à les tester, ils pourront vous être très utiles pour avoir à un instant T la visibilité de votre site sur les moteurs et la qualité du trafic généré.

Les outils pour webmasters fournis par les moteurs

Enfin, vous l'avez certainement compris en lisant cet ouvrage, il est très important, lorsqu'on s'intéresse au référencement, de créer des comptes sur les outils pour webmasters que proposent les outils de recherche majeurs. Le suivi du bon référencement de votre site est souvent indissociable de l'utilisation de ces outils.

Les Google Webmaster Tools (GWT pour les intimes, *http://www.google.com/webmasters/*) sont les plus importants. Tout d'abord, c'est de loin l'interface qui propose le plus d'outils utiles et parfois indispensables. Et, ça tombe bien, il est fourni par le leader actuel des moteurs de recherche. Il est inimaginable aujourd'hui de gérer son référencement sans un accès GWT car la richesse de ses informations et de ses diagnostics devient vite indispensable.

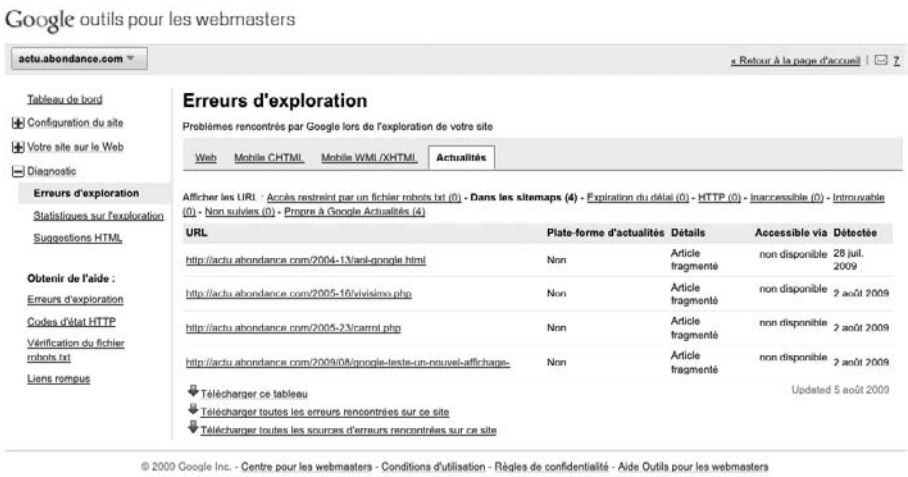


Figure 10-3

Exemple d'informations fournies par les Webmaster Tools : les problèmes que le spider a connu en parcourant vos pages

Yahoo! propose également l'outil Site Explorer (*http://siteexplorer.search.yahoo.com/)* mais nul ne sait s'il va perdurer après l'accord entre Yahoo! et Microsoft.

Microsoft propose également son propre site, similaire à celui de ses deux concurrents (*http://www.bing.com/webmaster*).

Ces outils sont des espaces absolument indispensables pour toute personne s'intéressant aux moteurs de recherche et au référencement. Ils sont gratuits et ne demandent que la création d'un compte spécifique (que vous détenez peut-être déjà). Foncez !

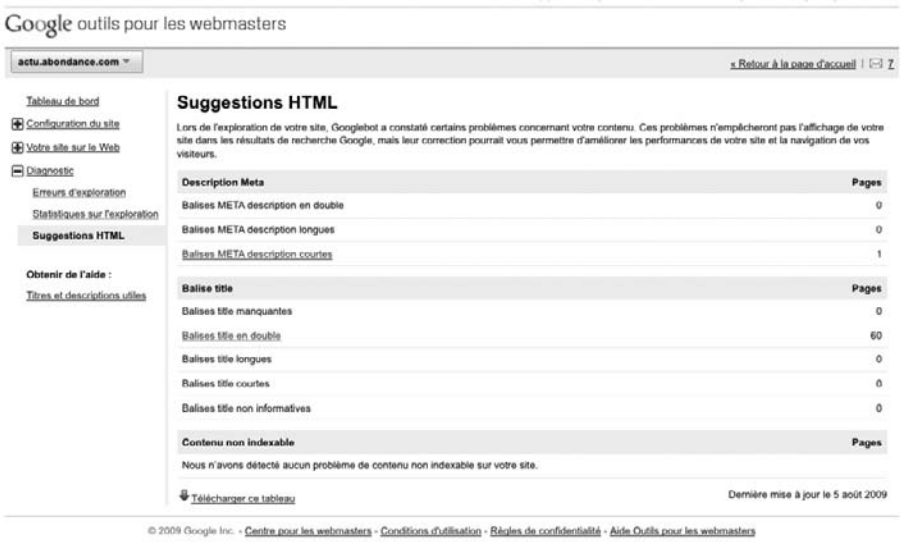


Figure 10-4

Autre information capitale : vos zones HTML trop courtes, trop longues, en double, etc.

Conclusion

La mesure de l'efficacité d'un référencement a beaucoup évolué depuis quelques années, certainement grâce à l'éclosion du positionnement publicitaire. Aujourd'hui, les clients veulent savoir où va leur argent et dans quels types d'actions ils investissent. Il paraîtrait même que ce n'est pas spécifique au monde du référencement ni à Internet…

À l'heure actuelle, les outils existent, comme nous venons de le voir. Ils permettent d'obtenir une vision pertinente de la qualité du trafic généré et des actions menées par les visiteurs issus des outils de recherche. Bien sûr, il existe toujours un décalage entre le discours et la réalité sur le terrain. Peu de sites web utilisent encore ces outils de tracking de façon approfondie, même en 2009. La faute, le plus souvent, à un manque de temps pour analyser les résultats fournis. Mais il y a fort à parier qu'ils vont devenir de plus en plus répandus car ils rendent de vrais services aux éditeurs de sites et… aux référenceurs !

11

Internalisation ou sous-traitance ?

En termes de référencement et de moteurs de recherche, plusieurs questions reviennent souvent de la part d'éditeurs de sites web, notamment de PME connaissant parfois mal ce marché si complexe à appréhender : Combien coûte un référencement ? Comment choisir un prestataire ? Est-il possible de garantir un positionnement ? Est-il possible de gérer soi-même son référencement, en interne de l'entreprise ?, etc.

Ces questions sont complexes et n'amènent pas toujours de réponse unique et évidente. Savez-vous répondre à la question « Combien coûte un site web ? » autrement que par l'affirmation : « ça dépend » ? Idem pour une question du type : « Dois-je faire mon site moi-même ou le confier à un professionnel ? » qui amènera invariablement le même type de réponse… Bien sûr, cela dépend de nombreux critères !

Le coût d'un retour en arrière...

Dans la pratique, on s'aperçoit vite que le coût d'un référencement revient surtout à la question : Combien cela va-t-il me coûter de revenir en arrière pour rendre mon site web compatible avec les critères de pertinence des moteurs de recherche ?... Quand on vous dit que plus l'aspect « référencement » est pris en amont dans le cahier des charges, mieux cela vaut...

En effet, plusieurs facteurs entrent en jeu pour évaluer le coût d'un référencement. Parmi ceux-ci, nous pouvons citer :

- votre structure : site personnel, professionnel indépendant, PME, grand groupe, filiale d'une société étrangère, etc. ;

- votre budget : de 0 à plusieurs centaines de milliers d'euros parfois ;
- le temps que vous avez à consacrer à ce projet en interne ;
- vos besoins : en nombre de sites, de langues, de mots-clés, etc. ;
- votre volonté de mesure du travail effectué : positionnement, trafic, calcul du ROI (retour sur investissement), etc. ;
- votre connaissance du domaine.

On pourrait multiplier les critères à l'envi, nous en trouverions encore bien d'autres…

Il est ainsi très difficile de répondre à certaines questions sur le référencement. Dans ce chapitre, nous tenterons cependant de vous apporter un certain nombre d'éléments qui vous permettront de prendre vos décisions en toute connaissance de cause…

Faut-il internaliser ou sous-traiter un référencement ?

Avant d'essayer de répondre à la question d'une éventuelle sous-traitance, il est nécessaire de bien prendre en considération le fait qu'un site web doit avant tout être « prêt » et donc optimisé pour être réactif aux critères de pertinence des moteurs dès sa mise en ligne. Les chapitres précédents de cet ouvrage devraient vous y aider.

Tout d'abord, nous prendrons bien entendu comme postulat que vous ne désirez pas spammer les moteurs de recherche et que vous renoncez donc à toute idée de passer par des rustines de type pages satellites ou autre système « industriel » de création de pages en très grand nombre (visibles ou non) et destinées aux moteurs de recherche. Vous avez donc en tête l'idée d'optimiser votre site *in situ* et donc de modifier vos pages afin qu'elles deviennent plus réactives par rapport aux moteurs de recherche. C'est une excellente idée...

Plusieurs situations sont alors possibles :

- **Vous créez votre site web de toutes pièces.** Vous en êtes à la phase de définition du cahier des charges. Dans ce cas, il est absolument nécessaire de prendre en compte les contraintes dues au référencement – et donc aux moteurs de recherche – dès le départ. Il est surtout primordial de ne pas faire d'erreurs technologiques comme la création d'un site 100 % Flash ou l'absence de réécriture d'URL en cas de site dynamique (voir chapitre 7). Vous devrez donc penser au référencement dès le début mais également tout au long de la mise en place conceptuelle et technique de votre site. Pour cela, vous pouvez vous aider de cet ouvrage ou vous faire assister par une société spécialisée dans le référencement qui va suivre, en tant que conseil externe, la réalisation de votre site en inspectant les maquettes successives imaginées par la société qui crée les pages (si vous ne les faites pas vous-même). Nous insistons sur cette notion de suivi dans le temps car le référencement est souvent synonyme de contraintes que l'on oublie parfois en s'en mordant les doigts par la suite lorsque des choix trop définitifs ont été faits.

- **Vous passez par un prestataire spécialisé.** Vous pouvez choisir une société de création de sites web sensibilisée aux techniques de référencement. L'expérience nous montre qu'elles sont assez peu nombreuses mais que, bien heureusement, elles existent ! Cela pourra donc être un critère de choix important lors de votre appel d'offre. Cela vous évitera également de gérer des conflits avec cette entreprise lorsqu'il faudra lui expliquer les contraintes dues aux moteurs.

  La situation idéale sera donc constituée d'une société spécialisée dans la création de sites, assistée tout au long du projet d'un expert du référencement qui pourra apporter des conseils au fur et à mesure. Vous aurez ainsi la quasi certitude qu'une fois en ligne, vos pages seront bien prises en compte par les moteurs. Mais, encore une fois, rien ne vous empêche de tout faire vous-même… Vous êtes votre seul maître.

- **Votre site web est déjà en ligne et n'est pas optimisé.** Dans ce cas, vous pouvez entrevoir plusieurs solutions intermédiaires, comme la réécriture des titres des pages (balises `<title>`, qui ne modifient pas la charte graphique de votre site) et un certain nombre d'autres pratiques classiques dans le domaine du référencement. Vous pouvez bricoler (mais on peut faire du bricolage très professionnel) des solutions provisoires en attendant des jours meilleurs, c'est-à-dire une refonte future de votre site, afin de prendre en compte la problématique du référencement dès le départ. Nous sommes persuadés que l'on peut fortement augmenter le trafic issu du référencement uniquement en « colmatant quelques brèches » comme le fait de revoir les titres, de réviser certains textes, de réécrire les textes des liens, de mieux gérer les liens entrants vers le site (*backlinks*) en général, etc. Pour le reste, vous pourrez vous en occupez plus tard, aucun souci.

  Dans ce cadre, vous avez le choix entre sous-traiter ces opérations ou les effectuer tout seul, tout en sachant que l'intervention et les conseils d'un expert du domaine peut vous faire gagner pas mal de temps. Et chacun sait que le temps, c'est de l'argent…

  Le « plus » de la sous-traitance par une société ou un expert spécialisé sera également parfois apporté par les outils de suivi : un extranet où vous pourrez analyser vos positionnements au jour le jour, le trafic généré avant et après la prestation, etc. Bref, les rapports d'évaluation du travail effectué seront accessibles grâce à des outils professionnels, ce qui est loin d'être négligeable.

  De plus, le référenceur pourra également intervenir en cas de problème épineux : URL Rewriting très technique, référencement dans des langues moins connues comme le roumain, le chinois ou le russe, etc. Enfin, son métier est également de mener une veille continuelle sur les moteurs de recherche et leurs algorithmes car c'est un domaine qui bouge beaucoup. En avez-vous le temps de votre côté ?

Pour résumer, on peut dire qu'il est possible de faire beaucoup de choses soi-même, en interne, dans le cadre d'un référencement de site web, mais qu'il sera préférable, dès que l'on désire effectuer un travail réellement professionnel, de s'adjoindre les conseils d'une société experte en référencement afin d'éviter toute erreur technologique et de suivre au plus près l'évolution d'un projet de mise en ligne d'un site.

Dans ce cadre, quelle organisation peut être mise en place, mixant ou non sous-traitance et internalisation, pour obtenir le meilleur référencement possible ? Le mieux est certainement de reprendre, étape par étape, votre projet de référencement et de tenter de positionner la sous-traitance et/ou l'internalisation des tâches dans chacune d'entre elles.

Audit et formation préalable

Chaque projet de référencement demande un audit préalable de la situation.

S'il s'agit d'un site qui n'est pas encore en ligne (ou que vous désirez refondre en profondeur), vous devrez mentionner dans un cahier des charges les grandes lignes stratégiques que vous désirez suivre pour votre référencement : prestations demandées, outils utilisés, objectifs et garanties demandés, délais, etc. Cette phase peut tout à fait être réalisée en interne si vous connaissez un tant soit peu le domaine du référencement. Difficile, en revanche, si ce monde vous est inconnu ou mal connu, de vous passer d'un prestataire qui va vous conseiller pour savoir jusqu'où aller (ou ne pas aller), ce qui vous évitera de demander à des prestataires éventuels des objectifs qu'ils ne pourront pas atteindre. En tout état de cause, il semble préférable et salutaire de se renseigner dans un premier temps sur ce qu'est le référencement au moment de votre prise de décision (ce domaine évolue tellement vite) et sur les différentes manières d'optimiser un site, même si vous ne rentrez pas dans la technique pure. Une connaissance, même générale, du référencement sera intéressante car elle vous permettra de poser les bonnes questions et de savoir si les réponses apportées par d'éventuels prestataires sont fantaisistes (cela arrive) ou sérieuses.

Il sera important et primordial, avant de prendre quelque décision que ce soit, d'avoir une connaissance générale de la façon dont un référencement s'effectue aujourd'hui. Il s'agit d'un monde qui évolue vite. Quelques lectures récentes et une journée de formation n'en seront que plus salutaires. Vous pouvez également assister à un séminaire, par exemple, il en existe de nombreux sur ce sujet.

Formation au référencement

La plupart des formations sur le sujet sont… référencées dans la rubrique Agenda de la lettre gratuite et hebdomadaire Actu Moteurs du site Abondance. Pour les consulter, rendez-vous à l'adresse suivante : *http://lettres.abondance.com/actumoteurs.html.*

Il se peut également que le site soit déjà en ligne et que vous ne désiriez pas le refondre de fond en comble. Dans ce cas, il est nécessaire d'effectuer un audit complet qui comprendra plusieurs rubriques :

- problèmes actuels d'optimisation (titres mal rédigés, liens inefficaces, utilisation de JavaScript, frames, Flash, etc.) à résoudre ;
- « état de l'art » de votre site en termes de nombre de pages indexées par les moteurs et de *baclinks* déjà obtenus.

- positionnements déjà acquis : la conservation de l'acquis sera un point important à traiter lors de votre projet. Comment améliorer votre référencement général sans perdre le travail déjà effectué et sans revenir en arrière sur le trafic déjà généré ?
- trafic sur votre site actuellement apporté par les moteurs de recherche et sur quels mots-clés ?
- actions à entreprendre pour améliorer la situation et délais de mise en œuvre.

Là encore, toutes ces données vous seront utiles pour établir votre cahier des charges de référencement. Vous pouvez bien entendu effectuer ce travail d'audit vous-même, mais nous vous conseillons de le faire réaliser par une société extérieure, qui aura plus de recul sur un site qu'elle ne connaît pas et qui pourra, dans le cadre d'une prestation unique, vous donner bon nombre d'informations sur tout le travail qu'il reste à effectuer. Ces renseignements vous permettront de repartir du bon pied sur la base d'informations pertinentes et réelles.

Un audit de trafic d'un site existant est toujours très intéressant à mettre en place avant une action de référencement, ne serait-ce que pour pouvoir estimer le travail effectué une fois l'optimisation mise en ligne. Une bonne base de réflexion, en quelque sorte...

Élaboration du cahier des charges

Une première phase préalable de définition des besoins est donc mise en place, et va déboucher sur l'élaboration du cahier des charges « référencement ». La réflexion effectuée précédemment doit vous aider à réaliser ce document. Cependant, nous ne saurions que trop vous conseiller soit de le réaliser avec une société du domaine spécialisée dans le conseil, soit de le montrer, une fois rédigé, à une telle structure. Là encore, un œil extérieur sera salutaire.

Si vous travaillez en termes de conseil et d'audit avec une société qui propose aussi des prestations de référencement proprement dit, vous courrez le risque qu'elle soit juge et partie et qu'elle vous oriente vers ses prestations plutôt que vers celles de la concurrence. Dans ce cas et tant que le cahier des charges final ne sera pas rédigé, peut-être sera-t-il préférable d'opter pour des structures plus spécialisées en conseil (même s'il en existe hélas peu en France malgré la demande grandissante) qu'en prestations de référencement proprement dites.

L'intervention sur le cahier des charges « référencement » d'une société extérieure de conseil est toujours intéressante, soit pour la rédaction elle-même, soit pour la vérification que tout a bien été pris en compte. Les erreurs commises dès le départ seront d'autant plus complexes à solutionner plus tard.

Définition des mots-clés

Tout d'abord, il est évident que vous maîtrisez un aspect essentiel : votre métier. Vous êtes le mieux placé pour définir, par exemple, les mots-clés qui correspondent à votre domaine professionnel et sur lesquels vous désirez vous positionner. Cependant, un

prestataire extérieur qui connaît bien le référencement et ses outils disponibles en ligne pourra vous apporter un éclairage intéressant sur plusieurs aspects :

- L'intérêt des requêtes que vous avez identifiées. Les mots-clés que vous avez imaginés sont-ils souvent saisis sur les moteurs de recherche ? La connaissance des nombreux générateurs de mots-clés disponibles en ligne sera un plus dans ce cadre.
- La vision que vous avez de votre métier au travers d'un certain nombre de termes est-elle la même que celle d'un internaute extérieur, pas obligatoirement spécialiste de votre métier, qui recherche une société comme la vôtre ou des produits/services comme ceux que vous proposez ?
- Quels sont la faisabilité technique et les délais prévisionnels pour espérer atteindre de bons résultats en référencement naturel ? Par exemple, si vous désirez être en première page des moteurs sur des termes génériques et concurrentiels comme « hôtel », « tourisme » ou « santé », cela peut prendre des années sans aucune garantie de résultats.

C'est pourquoi nous pensons que cette définition optimale des mots-clés se doit d'être accompagnée par des spécialistes du domaine. Et ce, d'autant plus que c'est une phase cruciale et essentielle dans votre projet. Combien de référencements ont-ils été ratés dans le passé car portant sur des requêtes que personne ne saisissait ou sur des mots-clés sur lesquels des bonnes positions étaient trop complexes à atteindre en peu de temps ?

Vous connaissez bien votre métier, mais un intervenant extérieur vous donnera un éclairage important sur l'intérêt et la faisabilité d'un référencement sur les requêtes imaginées et vous donnera une visibilité sur des termes souvent utilisés auxquels vous n'aviez pas pensé.

Mise en œuvre technique du référencement

Une fois le cahier des charges réalisé, vous aurez logiquement une vision un peu plus claire de ce qu'il vous faut faire pour améliorer votre référencement. Cette mise en œuvre technique s'accompagne alors le plus souvent de plusieurs étapes :

- Rédaction de meilleurs titres, textes, intitulés de liens, etc. Ici, vous êtes sans doute le mieux placé pour effectuer ce travail en interne puisque c'est vous qui gérez au quotidien votre site web. Le prestataire extérieur pourra en revanche intervenir sur deux domaines précis :
  - La rédaction d'un cahier de préconisations qui vous indiquera ce que vous devez faire pour optimiser vos pages actuelles au niveau éditorial (employer des titres plus longs et plus descriptifs, soigner les chapôs des articles en y insérant les termes importants en gras, etc.) et technique (utiliser la balise `<h1>` pour les titres rédactionnels, insérer les attributs `alt` pour chaque image, veiller à la compatibilité W3C des pages, insérer des balises meta `description` spécifiques à chaque page, etc.). Ce cahier constituera en quelque sorte un manuel des bonnes pratiques vous permettant

d'écrire vos contenus afin qu'ils soient réactifs au mieux par rapport aux critères de pertinence des moteurs.

– Éventuellement, la formation de vos équipes rédactionnelles pour leur apprendre non seulement à « écrire pour le Web » mais également à « écrire pour les moteurs ». Cela s'apprend et il n'est pas forcément évident d'appréhender les « gestes qui sauvent » lors de l'écriture de contenus, notamment lorsqu'on a l'habitude d'écrire pour le format papier. Le Web est un autre monde…

  Vous êtes certainement le plus à même d'intégrer l'optimisation des pages dans votre site. Mais un prestataire extérieur peut vous guider en vous expliquant comment le faire au mieux de vos intérêts.

- Prestations plus techniques comme de l'URL Rewriting, la mise en place de redirections 301, etc. Si vous avez besoin de telles prestations, nous vous conseillons de faire appel à une société spécialisée car ces domaines techniques sont souvent très « sensibles » et la moindre mauvaise manipulation peut donner des résultats catastrophiques. Le plus souvent et si besoin est, il vaut donc mieux s'en remettre à des entreprises qui maîtrisent cet aspect technique sur le bout des doigts.

  Un intervenant extérieur peut éventuellement être d'une aide estimable si le site est réalisé par une agence web. Celle-ci peut tout à fait envoyer des maquettes successives à un spécialiste du site et de sa charte graphique pour avis. Le référenceur indique alors ce qui, éventuellement, peut poser problème aux moteurs. Cela permet d'avancer rapidement et d'éviter tout écueil, le site web étant alors réactif dès sa mise en ligne.

- Dernier point : aujourd'hui, on ne soumet plus son site aux moteurs de recherche à l'aide d'un formulaire de type Add URL (voir chapitre 8). La meilleure façon de voir son site indexé est d'obtenir rapidement des liens émanant de pages populaires. Si un référenceur peut vous fournir un tel lien, la société ne vous aidera pas beaucoup dans le cadre de soumissions désormais obsolètes (quel crédit apporter aujourd'hui à une société de référencement qui facture des prestations de soumission d'un site sur les moteurs de recherche ?).

Suivi du référencement

Une fois votre site optimisé, il vous faudra bien évidemment suivre le travail effectué. L'audit du trafic, effectué dans la phase préalable (voir précédemment) vous sera bien utile dans ce cas pour comparer les étapes avant et après le référencement. Pour cette phase de suivi, tout dépendra de votre volonté et de la stratégie que vous désirez mettre en œuvre. En effet, vous pouvez vous contenter (mais cela serait dommage) de vérifier les positionnements obtenus grâce à des outils adéquats (logiciels, sites web, outils fournis par le prestataire). Cette phase, si elle est essentielle, ne sera la plupart du temps pas suffisante comme nous l'avons vu au chapitre précédent.

L'étude du trafic généré par les différents moteurs, des mots-clés qui ont servi à trouver vos pages, voire le calcul du retour sur investissement sont des étapes successives qui peuvent vous permettre de grandement affiner votre référencement et votre visibilité sur

les moteurs de recherche. Encore faut-il avoir du temps pour analyser toutes ces données… Ceci dit, vous serez le mieux placé pour faire ce travail qu'il semble difficile de sous-traiter. En revanche, une prestation de conseil sur le choix des outils de vérification de positionnement, de mesure d'audience et de ROI ne sera peut-être pas négligeable. De même, une prestation de conseil sur l'interprétation des résultats (pas toujours très évidente), pourra s'avérer un choix judicieux. Mais la plupart du temps, vous pourrez peut-être préférer des outils indépendants des sociétés de référencement afin d'être sûr que les résultats fournis sont objectifs.

Si un prestataire extérieur peut vous aider dans le choix des outils de suivi du référencement, l'analyse – indispensable – vous en revient en premier lieu.

Bien entendu, il faudra prévoir le cas où les résultats ne sont pas à la hauteur : nouveau travail sur le cahier de préconisations, nouvel audit, etc. Vous pourrez, dans ce cas, repartir en arrière vers une étape précédente.

Coûts

L'aspect financier est également à prendre en compte dans votre réflexion, comme vous pouvez vous en douter. Nous pouvons voir sur le tableau 11-1 quelques chiffres sur l'internalisation d'un spécialiste du référencement à un coût qui peut varier bien évidemment suivant le type d'entreprise et sa localisation géographique. Pour une TPE ou une PME, l'intégration d'un spécialiste peut devenir au final très coûteuse par rapport au choix d'un prestataire spécialisé. Dans le cadre des grands comptes, une internalisation peut en revanche être plus envisageable.

Tableau 11-1 Coût d'une prestation de référencement en interne et en sous-traitance

	Internalisation d'une personne dédiée spécialisée (salaire moyen Paris et province)	Externalisation (prestation annuelle – prix moyens)
Grands comptes	Salaire entre 30 000 et 40 000 €/an + primes	Entre 20 000 et 50 000 €
PME	Salaire entre 20 000 et 30 000 €/an (primes comprises)	Entre 3 000 et 10 000 €
TPE	-	Entre 1 000 et 3 000 €

Bien sûr, ces chiffres sont donnés à titre indicatif puisque les salaires d'une personne embauchée, comme le tarif des sociétés de référencement, dépendent de nombreux facteurs. Mais, là encore, il sera intéressant pour vous d'effectuer un tel calcul par rapport à votre propre contexte pour savoir dans quoi vous vous aventurez.

Nos conseils :

- **TPE/PME.** Orientez-vous plutôt vers un webmaster qui pourra être le parfait relais avec un prestataire de référencement, mais attention aux « Je sais tout faire, et je peux m'occuper du référencement ». Le webmaster ne doit pas, de son coté, engager des

actions de référencement sans en avoir averti le prestataire éventuel. Formez-le plutôt à optimiser votre linking, à trouver de bons partenaires, à mettre en place des échanges de contenu, etc.

- **Grands comptes.** Si vous choisissez d'internaliser toute la chaîne de référencement, assurez-vous que vous maîtrisez bien tous les aspects techniques de votre site (notamment pour les entreprises internationales pour lesquelles, par exemple, le site web est géré depuis les États-Unis) et que vous avez entièrement la main dessus, sans quoi votre spécialiste interne risque vite d'être frustré de ne pas pouvoir appliquer tous ces petits trucs et astuces et de ce fait risque de vous quitter très rapidement… La tendance en 2009, chez les grands comptes, est clairement à l'internalisation de cette fonction, avec beaucoup de difficultés, d'ailleurs, à trouver des profils compétents en la matière... Les « experts SEO » sont rares et donc chers.

Le métier de référenceur en France

La très active association française de référenceurs SEO Camp a publié en octobre 2008 une intéressante étude sur le métier de référenceur en France. Selon cette étude :

– Plus de 54 % des référenceurs recrutés en France sont autodidactes. Certains sont plus spécifiquement autodidactes en référencement, disposant par ailleurs des diplômes les plus divers. Tous les chemins semblent mener au SEO : il ne se dégage pas aujourd'hui de filière, ni de parcours type, pour exercer ce métier. (notons cependant la création d'une formation intitulée « Référenceur & Rédacteur Web » à l'IUT de Mulhouse pour laquelle l'auteur de cet ouvrage a eu le privilège d'être le parrain de la première année d'activité : *http://blog.abondance.com/2008/06/formation-de-rfrenceur-mulhouse-plus.html*).

– Deux tiers des entreprises qui recrutent ne reçoivent en moyenne que cinq curriculum vitae par courrier. De l'aveu même des employeurs, les profils des candidats présentent des niveaux très divers, souvent insuffisants. Bref, il existe plus d'offre que de demande et les « bons » référenceurs restent des perles rares...

– Les salaires des référenceurs juniors sont compris entre 20 et 30 k€ /an, 80 % des salaires étant compris dans la fourchette 23-29 k€/an. Les salaires de personnes confirmées se situent plutôt dans la fourchette 32–35 k€/an et les « experts » reconnus (mais qui constituent un club très fermé) se voient plutôt offrir des salaires entre 45 et 52 k€/an.

Pour plus d'informations sur cette étude, consultez l'adresse suivante : *http://actu.abondance.com/2008/10/seo-camp-publie-une-tude-sur-le-mtier.html*.

Préconisations

En règle générale, il nous semble judicieux de mettre en œuvre une stratégie qui mixe vos propres interventions (vous connaissez bien votre métier, vous avez la main sur votre site web) et une prestation extérieure (cette société aura souvent le recul nécessaire et les connaissances techniques indispensables pour mener à bien le projet) afin d'optimiser votre projet de visibilité sur les moteurs de recherche. Le fait de travailler avec une agence extérieure permet aussi de bénéficier de l'expérience de différentes problématiques rencontrées sur la globalité de la clientèle et donc d'avoir éventuellement beaucoup plus de solutions personnalisées à proposer, contrairement à une personne ne travaillant que sur un seul dossier en interne. Il en sera ainsi pour une veille constante orientée sur le

fonctionnement et les évolutions annoncées ou constatées des outils de recherche (et tout particulièrement Google). En effet, il sera dans certains cas primordial de réagir (modification technique, travail sur l'environnement de liens, sur la structure du site…) afin de ne pas se voir touché (sanctionné) par le nouveau filtre d'un outil de recherche (Google, Yahoo! et Microsoft Bing évoluent régulièrement). En sachant qu'une réaction après coup sera souvent synonyme d'un délai de plusieurs semaines, voire plusieurs, mois avant d'être efficace, à vous de choisir un prestataire qui effectue une véritable veille quotidienne sur le monde des moteurs de recherche.

Autre point qui peut jouer en faveur de l'externalisation : la notion de garanties ou de tiers responsable. En effet, externaliser le travail, c'est aussi externaliser les risques et avoir de meilleurs moyens juridiques pour se défendre, ce qui n'est pas négligeable.

Conclusion

Aujourd'hui, il n'est plus question de livrer un site, quel qu'il soit, à un référenceur qui fabrique de façon industrielle des centaines de pages satellites ou qui « bidouille » des astuces plus ou moins vaseuses dans son coin sans vous en faire part. Un travail important et commun devra être mis en place entre le client (vous), la société qui réalise le site et le référenceur (si les trois entités sont différentes, bien sûr). De la parfaite adéquation, entente, maîtrise et communication des uns avec les autres tout au long de la mise en place du projet dépendra la qualité du travail final.

Voici un tableau récapitulatif des différentes étapes de votre projet de référencement avec, pour chacune d'entre elles, ce que vous pouvez internaliser et ce qu'il est préférable, à notre avis, de sous-traiter.

Tableau 11-2 Différentes étapes du projet de référencement

Étape du projet	Ce que vous pouvez éventuellement internaliser	Ce que vous pouvez éventuellement externaliser
Formation préalable	Veille personnelle, lecture de guides, d'ouvrages, de newsletters, visites de salons, conférences, colloques, etc. Veille régulière sur le référencement et les moteurs de recherche (si vous en avez le temps).	Formation ou conseil auprès d'un organisme spécialisé du domaine. Veille régulière sur le référencement et les moteurs de recherche (si vous n'avez pas le temps). Analyse concurrentielle.
Audit du site si déjà en ligne		Réalisation du guide d'audit et de préconisations techniques. Analyse de l'acquis et propositions d'améliorations.
Audit du site si pas encore en ligne	Réflexion sur le CDC (cahier des charges).	Aide à la réflexion sur le CDC.
Écriture du cahier des charges	Rédaction du CDC.	Corédaction du CDC ou avis sur le CDC une fois rédigé.

Tableau 11-2 Différentes étapes du projet de référencement *(suite)*

Étape du projet	Ce que vous pouvez éventuellement internaliser	Ce que vous pouvez éventuellement externaliser
Définition des mots-clés	Définition de mots-clés « de départ ».	Conseil sur d'autres mots-clés. Évaluation de l'intérêt et de la faisabilité des mots-clés identifiés par le client.
Optimisation des pages et du site	Intégration des préconisations.	Rédaction du cahier de préconisations. Suivi des maquettes réalisées par l'agence web. Réécriture et optimisation des contenus existants. Formation des équipes rédactionnelles. Mise en place de fonctions techniques complexes (redirections, fichiers `robots.txt`, URL Rewriting…). Mise en place de liens.
Suivi du référencement	Interprétation des résultats fournis par les outils de suivi.	Mise à disposition d'outils de suivi. Conseil sur le choix des outils de mesure d'audience et/ou de calcul de ROI. Aide à l'analyse des données.

Combien coûte un référencement ?

Malheureusement, cette question sera vite traitée car il est impossible d'y répondre, tout comme à la question « Combien coûte un site web ? ». Nous l'allons vu – et le paragraphe précédent vous donne déjà quelques chiffres de prestations type –, vous pouvez mener de nombreux travaux par vous-même et cela ne vous coûtera que le temps que vous allez y passer…

Certaines offres basiques que l'on trouve en ligne, comprenant quelques conseils et un suivi simple vont coûter quelques centaines d'euros, d'autres offres plus élaborées, avec du conseil sur le choix des mots-clés, un important travail d'optimisation de vos pages, un suivi par extranet des positionnements et du trafic généré, etc., seront facturées plusieurs milliers d'euros. Il arrive également de voir passer des budgets de plusieurs centaines de milliers d'euros pour des prestations de référencement dans le cadre de projets importants mis en place par de grands groupes.

Puisqu'il faut indiquer des chiffres, nous dirons qu'une prestation de base, très simple, coûtera environ de 500 à 2 000 euros HT par an (mais ce seront des offres très basiques, ne vous y trompez pas) et qu'une prestation plus évoluée tournera autour de quelques milliers, voire dizaines de milliers d'euros en fonction des prestations proposées. Sur cette base, il est possible de tout imaginer et également des sommes bien plus importantes.

Un référencement gratuit est-il intéressant ?

Certaines offres de référencement gratuit sont disponibles sur le Web. Ayez conscience que, dans ce cas, vous en aurez pour votre argent et qu'il ne faudra pas attendre grand-chose de ce type de prestation : pas de conseil, pas de suivi, la soumission de votre site à quelques annuaires et moteurs de recherche secondaires sera, la plupart du temps, effectuée automatiquement par des logiciels, ce qui, il faut bien l'avouer (et nous espérons que vous en serez convaincu après la lecture de cet ouvrage) n'est que peu intéressant. Bref, disons que ces offres ont le mérite d'exister, mais ne venez pas vous plaindre par la suite si elles ne vous apportent que peu de trafic…

Comment choisir un prestataire de référencement ?

Là encore, la réponse n'est pas simple et peut être rattachée à d'autres questions comme « Comment choisir la société qui va créer mon site ? » ou, encore de manière plus générale, « Comment choisir un prestataire informatique ? ».

En plus des critères généraux (bonne santé de la société, nombre de personnes, CA, existence d'un numéro SIRET, coût de la prestation, présence d'une adresse postale et d'un numéro de téléphone, etc.), qui ne sont pas spécifiques du référencement, il existe cependant certains critères à retenir qui nous semblent importants et propres à ce métier.

- Les références et les possibilités de questionner ses clients : il est souvent très intéressant d'aller voir le site proposé en référence par le prestataire envisagé et de regarder comment celui-ci est optimisé : titres, liens, balises meta `description`, etc.
- Pages satellites : il nous semble évident qu'il est indispensable d'écarter aujourd'hui toute entreprise qui baserait sa stratégie de référencement sur la création de pages satellites, technique obsolète et interdite par les moteurs. De façon plus générale, la société doit s'engager à ne pas utiliser de techniques considérées comme étant du spam par les moteurs (voir plus loin). Mais, bien évidemment, peu diront qu'elles en utilisent… Difficile également de considérer comme crédible une société qui facture la soumission de votre site aux moteurs de recherche. Le référencement a évolué depuis dix ans... De même, certaines entreprises estiment que le fait d'insérer (ou de vous demander d'insérer) des mots-clés dans les balises meta `keywords` de vos pages revient à faire du référencement et vous font payer ce travail. Là encore, fuyez !!!
- Le site du référenceur lui-même : même si, on le sait, les cordonniers sont souvent les plus mal chaussés, certains points sont intéressants à observer sur le site de la société de référencement : rédaction des titres, du texte des liens, etc. Par exemple, si le code HTML de la page d'accueil du site contient une balise meta `revisit-after` (vieux « serpent de mer » totalement inutile dans le cadre d'un référencement), vous pouvez avec raison vous poser quelques questions sur le sérieux et les compétences de la société. Passez votre chemin. C'est un exemple parmi tant d'autres, bien sûr… On voit également des sites web soi-disant professionnels vous expliquant encore dans leurs

pages que les balises meta `keywords` sont essentielles pour le référencement. Là encore, allez voir ailleurs… On peut parfois également détecter dans les pages du prestataire quelques signes de spam…

- Le nombre et le nom des outils de recherche proposés. Une grosse dizaine d'outils de recherche thésaurisent l'immense majorité du trafic. Ce n'est donc pas la peine de privilégier des offres quantitatives qui vous font miroiter un référencement sur plusieurs centaines, voire milliers d'outils. Mieux vaut peu de moteurs, mais qu'ils soient bien traités. Ce point a été évoqué au chapitre 3.
- L'honnêteté du prestataire : s'il vous promet la Lune et la première position sur des mots-clés comme « voyage » ou « hôtel », c'est de toute évidence un escroc ou alors il est en train de vous vendre du lien sponsorisé en appelant cela « référencement ». Malheureusement, cela arrive beaucoup plus souvent qu'on ne le croît. Attention également aux garanties proposées, dont nous parlerons plus loin dans ce chapitre.
- Le conseil, notamment à la définition des mots-clés, qui reste une étape primordiale dans la stratégie de référencement d'un site, comme on l'a vu précédemment. Le client a souvent besoin d'un œil extérieur pour bien choisir les termes sur lesquels il va tenter de se positionner. Ce recul nécessaire est souvent amené par le référenceur qui connaît bien les méthodologies de choix des termes à envisager. Mais la société que vous consultez vous propose-t-elle une prestation dans ce sens (ou vous demande-t-elle, de façon basique, une liste de mots-clés ?) ?
- Le contrat et la propriété des démarches effectuées : si certaines optimisations sont effectuées sur votre site, elles doivent vous appartenir, même si vous changez de prestataire par la suite, comme dans le cadre de toute prestation informatique.
- La qualité du suivi : extranet, rapports envoyés par e-mail, indicateurs de suivi fournis, périodicité des envois, calcul du retour sur investissement, etc. Vous devez absolument pouvoir contrôler la prestation effectuée quand vous le voulez.
- La veille : le prestataire vous fournira-t-il des informations sur les nouveaux moteurs, les évolutions du marché, les nouveautés du domaine ? Cela peut jouer et peser dans la balance.
- Suivi humain : votre projet sera-t-il suivi par un chef de projet ? Pourrez-vous le contacter directement ?
- Charte de qualité (voir plus loin) : la société en propose-t-elle une ou en a-t-elle signé une générique ? Que contient-elle ?

Avant de choisir votre prestataire, vous devrez également avoir travaillé de votre côté et avoir explicité clairement à la future société choisie vos besoins sous la forme d'un cahier des charges qui reprendra, même de façon synthétique, ce que vous voulez – ou ne voulez pas – en termes de référencement : optimisation naturelle, pas de pages satellites, URL Rewriting éventuel… Vous devrez donc jouer cartes sur table afin d'aider l'entreprise à vous aider dans votre projet…

Deuxième point important, vous devez obtenir des sociétés consultées des devis clairs en termes de prestations effectuées : conseil, réalisation de pages de contenus spécifiques, différentes étapes du projet, technologies mises en œuvre, etc. Fuyez les « boîtes noires » qui sont le plus souvent remplies de vent. Bref, il va certainement vous falloir comparer des offres diverses et vous devrez demander le plus de détails possible. Vous vous apercevrez bien vite que de nombreux devis sont assez abscons et qu'il est difficile de les comparer d'une entreprise à l'autre. Demandez le plus de précisions possibles dès le départ sous peine d'irrecevabilité de la proposition.

Où trouver une liste de prestataires de référencement ?

Il existe un très grand nombre de sociétés spécialisées en référencement. Certains sites tentent de les… référencer, en voici quelques-uns :

- *http://prestataires.journaldunet.com/competence/27/1/referencement.shtml*
- *http://www.seorch.com/*
- *http://www.annuaire-referencement.com/*
- *http://partenaires.abondance.com/*

Parfois, vous serez démarché directement, par téléphone ou par e-mail, par l'une d'entre elles. Mais le plus souvent, c'est le bouche à oreille, redoutable sur le Web, qui fera son office et qui vous indiquera les coordonnées d'entreprises sérieuses. Enfin, vous pouvez tenter de taper des mots-clés comme « référencement » ou « prestataire en référencement » pour voir qui se positionne sur ces termes… très concurrentiels !

Autre possibilité pour trouver une entreprise spécialisée : passer par une des associations les regroupant. Il en existe peu en France, et SEO Camp (*http://www.seo-camp.org/*) est de loin la plus active, mais d'autres, au niveau international comme le SEMPO (*http://www.sempo.org/*), pourraient vous être utiles.

Quelles garanties un référenceur peut-il proposer ?

On voit fleurir, dans les offres des différents prestataires du marché, de nombreuses notions de garanties, le plus souvent demandées à cor et à cri par les clients eux-mêmes. Il est cependant important de se souvenir d'une chose essentielle qui peut être résumée en une phrase : **il est absolument impossible de garantir en référencement naturel un positionnement sur un mot-clé donné pour un moteur de recherche donné** !

Cette phrase est un axiome évident, à partir du moment où :

- on ne connaît pas parfaitement – loin de là – les algorithmes de pertinence des moteurs de recherche ;
- ceux-ci sont très souvent modifiés (parfois de façon quotidienne) par leurs propriétaires ;

- de nouvelles pages optimisées peuvent venir modifier une situation que l'on croyait établie ;
- les moteurs de recherchent personnalisent de plus en plus leurs résultats en fonction de l'internaute (géolocalisation de l'ordinateur, langue du navigateur utilisé, requêtes saisies précédemment, etc.)
- etc.

En partant du fait qu'une garantie « individuelle » (un mot-clé/un moteur) est impossible (tout référenceur proposant ce type d'offre ne pourrait clairement pas obtenir votre assentiment), il existe pourtant un certain nombre de garanties « acceptables » que proposent les entreprises spécialisées, de façon plus globale : par exemple, une cinquantaine de positions en première page sur 100 mots-clés et 10 moteurs… Il s'agit ici de garanties statistiques qui semblent plus honnêtes. On peut imaginer également une garantie sur l'augmentation globale du trafic issu des moteurs de recherche suite à la prestation effectuée.

Malheureusement, il est impossible pour un prestataire de fournir plus de garanties sur le résultat de son travail, puisque celui-ci est basé sur des technologies qui appartiennent aux moteurs de recherche et, donc, qu'il ne contrôle pas. N'hésitez pas cependant à clarifier ce point avec la société avec qui vous envisagez de traiter car, dans certaines propositions, la notion de garantie n'est absolument pas expliquée par écrit alors qu'il en est fait souvent mention dans les argumentaires commerciaux oraux. N'oubliez pas non plus que les résultats dépendent de votre site et de son contenu. Vous ne pourrez obtenir des résultats que si vous suivez les conseils des sociétés de référencement avec qui vous décidez de traiter.

Enfin, faites attention à ce que la garantie proposée ne soit pas synonyme d'achat de lien sponsorisé mais bien de travail de positionnement organique et d'optimisation de site web. On voit, hélas, de tout sur le Web et si certaines sociétés de référencement sont très sérieuses, il existe un certain nombre de charlatans et d'apprentis sorciers à qui il vaut mieux ne pas confier son site.

Chartes de déontologie

En 2000, l'auteur de ce livre a travaillé pour le compte d'une association de référenceurs (IPEA), sur une charte du référencement. Ce document résumait la plupart des indications que nous avions jugées nécessaires d'insérer dans le cadre d'une charte de qualité et de déontologie du métier de référenceur.

Le but de cette charte était de faire en sorte que les outils de recherche (dont l'objectif est de proposer des réponses pertinentes aux internautes) et les référenceurs (dont l'objectif est d'assurer à leurs clients – disposant la plupart du temps de sites web à fort contenu de qualité – une visibilité optimale sur le Web) travaillent ensemble pour bâtir de meilleurs outils de recherche et, par là même, fournissent de meilleures réponses aux visiteurs des annuaires et moteurs.

Cette charte avait le mérite de fixer un certain nombre de points. Elle avait également l'avantage, à l'époque, d'avoir été acceptée par bon nombre d'outils de recherche et de référenceurs. Dans le cadre du site Abondance.com, nous l'avons rajeunie pour la proposer dans une version plus actualisée. Elle est disponible sur le Web à l'adresse *http://partenaires.abondance.com/charte.html* et est explicitée ci-après.

Charte de déontologie du métier de référenceur

Les signataires de la *Charte de qualité et de déontologie sur le référencement de sites web* devront accepter les conditions suivantes (O-R = outils de recherche) :

Tableau 11-3 Charte de qualité du métier de référenceur

Titre	Contenu
Réalisme/Garanties	Les signataires s'engagent à une obligation de moyens à mettre en œuvre et à ne pas promettre (garantir) de résultats de positionnement limités à une requête et un moteur, et plus généralement ne pas promettre de résultats qui ne pourront être tenus ou vérifiés par le client. Des garanties statistiques (X % de positionnements sur Y mots-clés et Z moteurs) pourront cependant être proposés.
Transparence	Les signataires s'engagent à tenir à disposition de leurs clients un document clair et précis présentant leur méthodologie de travail : technologies mises en œuvre, méthodes d'optimisation, procédures de référencement, etc.
Conseil	Les signataires s'engagent à aider leurs clients dans la réflexion sur les informations qui seront fournies aux O-R (descriptif, mots-clés, optimisation des titres et textes des pages, etc.) lors du référencement.
Méthodologie	Les signataires restent libres de la méthodologie mise en place pour référencer les sites de leurs clients, à partir du moment où elle respecte la présente charte, notamment en ce qui concerne la lutte contre le spam de la part des O-R (voir plus loin).
Loyauté	Les signataires s'engagent à suivre strictement les indications des O-R, publiées de façon spécifique sur leurs sites (et reprises ci-après dans la présente charte), dans le but d'effectuer une soumission efficace d'un site web dans leur index ou bases de données.
Suivi	Les signataires s'engagent à remettre à leurs clients de façon périodique des rapports clairs sur l'avancée des travaux de référencement de leur site web (suivi du positionnement, du trafic généré, du retour sur investissement, etc.), sous la forme qui leur semble la plus appropriée (tableaux Excel, extranet, outils en ligne, etc.).
Autonomie	Les signataires s'engagent à remettre tous les éléments relatifs aux travaux réalisés dans le cadre de la prestation de référencement de façon à permettre à leurs clients de changer de prestataire s'ils n'étaient pas satisfaits de la prestation effectuée. Comme pour toute prestation informatique ou de service, les travaux effectués appartiennent au client qui en a payé le montant.
Veille	Les signataires s'engagent à mettre en place des mécanismes de veille afin de se tenir au courant de l'évolution des outils de recherche et à en faire profiter leurs clients.
Qualité	Les signataires s'engagent auprès des O-R à ne soumettre à leur indexation que des sites dont le contenu et la pertinence sont suffisamment riches pour alimenter leur base de données en vue d'apporter une information utile au visiteur.

Tableau 11-3 Charte de qualité du métier de référenceur *(suite)*

Titre	Contenu
Mode de fonctionnement	Les signataires s'engagent à n'effectuer pour leurs clients que des prestations de référencement manuel, sans l'aide d'aucun logiciel de soumission automatique, sauf dans le cas où cette prestation est explicitement indiquée dans la proposition commerciale, et uniquement si l'utilisation de logiciels n'intervient qu'en complément d'une prestation manuelle majeure et ce aussi bien en phase de référencement qu'en phase de suivi et de veille.
Respect de la concurrence	Les signataires s'engagent à ne pas nuire au référencement d'un concurrent pour le compte d'un client et à ne pas utiliser la marque de concurrents pour le référencement de leurs clients. En règle générale, les signataires s'engagent à ne pas nuire au référencement d'un site pour lequel ils n'auraient pas été mandatés.
Combat contre le spam	Les signataires acceptent de ne pas réaliser d'action de spamdexing (fraude sur les O-R). La notion de spamdexing (ce qui est considéré comme tel et ce qui ne l'est pas) est explicitée ci-après. Les signataires s'engagent notamment à ne cacher aux O-R aucun contenu destiné au référencement à l'intérieur du code HTML (balise `noscript`, utilisation de zones invisibles : `visibility-hidden`, `display:none`, etc.) du site de leur client.
Information	Les signataires s'engagent à remettre à leurs clients et prospects, dans leurs propositions commerciales, un exemplaire de la Charte ici présente, accompagnée d'un document expliquant le fonctionnement des O-R, détaillant en quoi consiste un référencement de site web, ainsi que ses contraintes.
Clarté	Le signataire s'engage à expliciter de façon claire les actions effectuées sur les moteurs de recherche et notamment à ne pas entretenir de flou entre des prestations de référencement manuel et d'éventuelles actions d'achat de mots-clés dans le cadre de campagnes de liens sponsorisés. Les deux types d'actions, si elles cohabitent dans la prestation proposée au client, devront être clairement mentionnées et démarquées.
Blacklistage	Le signataire s'engage à rembourser intégralement la prestation réalisée s'il est avéré que le site est exclu d'un moteur suite à une faute de ses services.

Définition du spamdexing

Les actions suivantes sont considérées comme étant du spam (ou spamdexing), cette liste n'étant pas exhaustive (voir chapitre 8 pour davantage d'informations à ce sujet) :

- référencement d'un site faisant la promotion d'activités illicites ou dégradantes (piratage, racisme, sectes, pédophilie, etc.) ;
- mise en place d'une stratégie de référencement basée sur plusieurs noms de domaine différents (chaque nom de domaine pointant sur la même page) ou sous plusieurs adresses ou sous-domaines différents (*www.site.com* et *actu.site.com* par exemple, chaque adresse pointant sur la même page) ;
- répétition abusive des mots-clés à l'intérieur d'une page, que ce soit dans des zones cachées ou non du code HTML ;
- répétition abusive de balises `<title>`, meta ou autres (commentaires) ;

- utilisation de mots-clés inadéquats avec le contenu du site, utilisation de marques ou de noms sans liaison avec le site ou faisant mention de marques déposées sans autorisation ;
- utilisation de texte invisible (blanc sur fond blanc, par exemple) ou caché à l'intérieur du code HTML (balise `noscript`, utilisation de zones invisibles : `visibility-hidden`, `display:none`, etc.) ;
- cloaking (technique installée sur les serveurs web et permettant de délivrer aux spiders des pages spécifiquement réalisées pour les moteurs de recherche) à des fins de positionnement ;
- copie sur son propre site de code HTML existant et n'appartenant pas au signataire ;
- création de pages HTML ou de liens ayant pour unique but d'augmenter l'indice de popularité d'un site ou d'une page ;
- achat de liens au *prorata* d'un indice de popularité (PageRank ou autre) ;
- pages de redirection (balise meta `refresh`, JavaScript ou autre) notamment dans le cadre de la création de pages satellites ;
- popularité : toute manœuvre visant à tenter d'augmenter l'indice de popularité (ou PageRank) d'une page de façon artificielle ;
- pages satellites, alias, fantômes, *doorway pages*, etc., et toutes techniques d'optimisation permettant de positionner une page dans les pages de résultats des moteurs, cette page redirigeant automatiquement l'internaute sur une autre page réelle ;
- toute méthode de création massive et automatisée de pages web spécifiquement optimisées pour les moteurs de recherche, même sans redirection automatique, dans le cadre d'une stratégie de référencement ;
- d'une manière générale, non-respect des limites indiquées par les O-R.

Autres chartes de référencement

Plusieurs sociétés de référencement ont mis en place, sur leur site web ou dans le cadre de « livres blancs », des chartes de qualité qui leur sont propres ou qui sont destinées à chapeauter l'activité de leur profession. Il en existe de nombreuses, mais en voici quelques exemples :

– 1ère Position – *http://www.1ere-position.fr/livre-blanc-referencement-naturel.pdf*

– Brioude Internet – *http://www.referencement-2000.com/charte-referencement.html*

N hésitez pas à les lire et à vous en inspirer au moment de choisir votre prestataire.

Conclusion

Vous voici arrivés à la fin de cet ouvrage. Nous espérons qu'il vous a apporté une meilleure compréhension du monde du référencement et qu'il vous a fourni une aide pour mener à bien vos futurs projets dans ce domaine.

Si vous l'avez lu attentivement, vous aurez certainement retenu quelques « grandes idées fortes » que nous nous résumons dans cette ultime partie. N'hésitez pas à les relire souvent, elles sont garantes de vos futurs positionnements !

Les 12 phrases clés du référencement

1. Il est nécessaire de prendre en compte les contraintes du référencement dès l'élaboration du cahier des charges d'un site Web. Plus vous tarderez, plus la complexité augmentera et plus les moyens humains et financiers à mettre en œuvre seront importants.
2. Il est important de gérer au mieux un arbitrage entre la réalisation d'un site pour les internautes (faire « beau », proposer une navigation intuitive, etc.) et pour les moteurs de recherche (options technologiques à consommer avec modération comme Flash, JavaScript, etc.)
3. Choisissez avec soin les mots-clés sur lesquels vous désirez vous positionner. Il est dommage de tenter un positionnement sur un terme trop concurrentiel ou sur une expression jamais saisie sur les moteurs...
4. N'oubliez jamais que le trafic généré par les moteurs de recherche sur un site est de deux ordres : la tête de la longue traîne (mots-clés que vous avez définis au préalable et pour lesquels vous avez optimisé une ou plusieurs pages de votre site) et la queue de cette même longue traîne (mots-clés issus du contenu du site, non définis au préalable, mais mis en valeur par une optimisation du code HTML et de la conception même du site).
5. Soignez les titres, les URL, le texte visible de vos pages. Insérez-y les mots-clés importants pour votre activité.
6. Soignez le plus possible les liens de vos sites ainsi que les liens entrants (*backlinks*) sur ceux-ci. Ils sont actuellement la *killer application* du référencement.

7. Suivez votre référencement pour être toujours « au top », même si une optimisation bien faite est très souvent extrêmement pérenne !
8. Le positionnement n'est plus une stratégie efficace de mesure d'un référencement. Préférez l'analyse du trafic généré par les moteurs de recherche dans une optique de longue traîne.
9. N'hésitez pas à vous faire aider par un professionnel du domaine, qui peut vous faire gagner beaucoup de temps et d'argent... Mais choisissez-le soigneusement...
10. Évitez tout système de référencement reposant sur des pages satellites ou toute « rustine » de ce type... Le meilleur référencement, le plus efficace, le plus pérenne, est basé sur l'optimisation des pages de votre site.
11. Ne trichez jamais, ne cachez rien dans vos pages, on arrive à de superbes résultats en optimisant ses pages de façon propre, loyale et honnête !
12. En tout état de cause, *Content is KING* et, mieux, *Optimized content is EMPEROR!* (en français, *Le contenu est votre capital et son optimisation lui donne la meilleure visibilité* !) Le référencement peut donc être vu comme le moyen de donner une bonne visibilité sur les moteurs de recherche à un contenu de bonne qualité. Tout commencera donc toujours par la qualité de ce contenu.

Et maintenant, à vos claviers et n'hésitez pas à nous envoyer un petit message à l'adresse *livre-referencement@abondance*.com si cet ouvrage vous a aidé. Nous le publierons, éventuellement, sur le site *www.livre-referencement.com* (vous y gagnerez un lien vers votre site, ce qui est toujours bon pour votre indice de popularité...) qui présente ce livre.

D'ici là... bon référencement !

Olivier Andrieu

www.abondance.com

www.livre-referencement.com

E-mail : *livre-referencement@abondance.com*

Annexe

Webographie

Voici quelques adresses de sites web et outils qui pourraient s'avérer très intéressants pour votre référencement. N'hésitez pas à les consulter...

La trousse à outils du référenceur

Tout webmaster qui s'intéresse au référencement a besoin, au quotidien, d'utiliser un certain nombre d'outils qui vont lui faciliter la vie. Nous en avons signalé bon nombre tout au long des chapitres précédents et ne nous reviendrons pas dessus ici. Voici quelques sites, logiciels et add-ons pour Firefox que nous n'avons pas mentionnés auparavant et qui peuvent vous aider dans vos quêtes et vos analyses.

Add-ons pour Firefox

- **Web Developer** (*http://chrispederick.com/work/web-developer/*), indispensable outil qui propose de nombreux utilitaires de diagnostic et test de page web : désactivation du JavaScript, des images, des CSS, recherche de texte caché, etc. Difficile de s'en passer...

Figure 12-1

Web Developer, indispensable compagnon de route du référenceur

- **FireBug** (*http://www.joehewitt.com/software/firebug/*) permet de déboguer les pages HTML et d'en lire rapidement le contenu.
- **Edit Config Files** (*http://extensions.geckozone.org/EditConfigFiles*) permet de voir les liens en `nofollow` d'une page grâce à une petite astuce (décrite ici : *http://www.webrankinfo.com/actualites/200801-comment-voir-facilement-les-liens-nofollow-dans-firefox.htm*).
- **SEOpen** (*http://seopen.com/firefox-extension/index.php*) permet de voir les backlinks et de très nombreuses informations sur une page donnée.
- **Search Status** (*http://www.quirk.biz/searchstatus/*) donne, entre autres, le PageRank d'une page, le fichier `robots.txt` d'un site, etc.
- **SEO Link Analysis** (*http://yoast.com/tools/seo/link-analysis/*) fournit des informations sur les backlinks d'une page.
- **SEO for Firefox** (*http://tools.seobook.com/firefox/seo-for-firefox.html*) propose de très nombreuses fonctionnalités également. À tester !
- **SEOQuake** (*http://www.seoquake.com/*) propose de très nombreuses informations sur une page web en cours d'affichage.
- **SEOmoz Toolbar** (*http://www.seomoz.org/mozbar*) met à votre disposition une barre d'outils entièrement créée pour les référenceurs, par l'incontournable site SEOmoz...

Il existe des dizaines d'extensions pour Firefox orientées SEO. Nous n'indiquons ici qu'un échantillon car il est impossible de toutes les décrire et de détailler les fonctionnalités de chacune d'entre elles. N'hésitez pas à les tester pour voir si leurs fonctionnalités vous conviennent, et à fouiller sur le Web, vous y découvrirez certainement quelques pépites !

Test de validité des liens

- **Link Valet** (*http://htmlhelp.com/tools/valet/*) teste si les liens dans une page sont valides ou cassés.
- **W3C Link Checker** (*http://validator.w3.org/checklink*) effectue le même type d'action que Link Valet.
- **Xenu's Link Sleuth** (*http://home.snafu.de/tilman/xenulink.html*) est un incontournable du domaine.

Analyse du header HTTP

- **L'analyseur de header HTTP** de WebRankInfo (*http://www.webrankinfo.com/outils/header.php*) vous indiquera le code de retour du serveur pour une URL donnée. Une bonne façon de savoir, par exemple, si une redirection est faite en 301, en 302 ou autres.

Sites web d'audit et de calcul d'indice de densité

- **Outiref** (*http://www.outiref.com/*) propose un audit complet du site de façon automatique, avec notamment un calcul de l'indice de densité des mots de la page.
- **Yagoort** (*http://outils.yagoort.org/compteurmots.html)* permet de calculer le nombre d'occurrences d'un mot et son indice de densité.
- **Outils référencement** (*http://www.outils-referencement.com/outils/mots-cles/densite*) est un autre outil de calcul de l'indice de densité utilisant un dictionnaire de mots vides.
- **WebSiteGrader** (*http://www.websitegrader.com/*) envoie par e-mail un audit complet.

Positionnement

Outre les logiciels de calcul et de suivi du positionnement de vos pages dans les résultats des moteurs, déjà évoqués dans cet ouvrage, il existe également un certain nombre de sites web qui peuvent vous aider dans ce domaine. En voici quelques-uns (il en existe beaucoup) :

- **Visiref** (*http://visiref.com/*)
- **Tests de positionnement WebRanInfo** :
  - *http://www.webrankinfo.com/outils/positionnement-google.php*
  - *http://www.webrankinfo.com/outils/positionnement-yahoo.php*
- **Ranks.fr** (*http://www.ranks.fr/*)
- **Référencement SEO** (*http://www.referencement-seo.fr/Positionnement-Google.seo*)

Il existe encore de très nombreux outils pouvant vous aider au quotidien dans vos optimisations de site et votre référencement. La liste ci-dessus n'est qu'une liste non exhaustive de quelques-uns d'entre eux qui nous ont semblé intéressants. N'hésitez pas à en rechercher d'autres sur le Web !

Les musts de la recherche d'informations et du référencement

Il n'y a pas qu'Abondance.com (le site web de l'auteur de cet ouvrage) dans la vie. Voici une liste d'autres sites et de blogs fournissant bon nombre d'informations sur les moteurs de recherche et le référencement de sites web.

En français

- Abondance : *http://www.abondance.com/*
- Affordance.info : *http://affordance.typepad.com/mon_weblog/*
- Blog d'Abondance : *http://blog.abondance.com/*
- Google XXL : *http://googlexxl.blogspot.com/*
- MoteurZine : *http://www.moteurzine.com/*

- Motrech : *http://motrech.blogspot.com/*
- Oseox : *http://oseox.fr/*
- Outils Froids : *http://www.outilsfroids.net/*
- Référencement, Design et Cie : *http://s.billard.free.fr/referencement/*
- Secrets2Moteurs : *http://www.secrets2moteurs.com/*
- Technologies du Langage : *http://aixtal.blogspot.com/*
- Urfist Info : *http://urfistinfo.blogs.com/*
- WebRankInfo : *http://www.webrankinfo.com/*
- Zorgloob : *http://www.zorgloob.com/*

En anglais

- Google Blogoscoped : *http://blogoscoped.com/*
- John Battelle's Searchblog : *http://battellemedia.com/*
- Matt Cutts Gadgets, Google, and SEO : *http://www.mattcutts.com/blog/*
- Pandia : *http://www.pandia.com*
- SEOmoz : *http://www.seomoz.org/*
- Search Engine Guide : *http://www.searchengineguide.com*
- Search Engine Land : *http://www.searchengineland.com*
- Search Engine Showdown : *http://www.searchengineshowdown.com*
- Search Engine Watch : *http://www.searchenginewatch.com/*
- SEO by the Sea : *http://www.seobythesea.com/*

Blogs officiels des moteurs de recherche

Les sites officiels des moteurs de recherche sont légion dans ce domaine...

- Google (la liste complète des – nombreux – blogs de Google se trouve sur la droite de la page d'accueil du blog général) : *http://googleblog.blogspot.com/*
- Google Webmaster Tools (indispensable) : *http://googlewebmastercentral.blogspot.com/*
- Google Inside AdWords : *http://adwords.blogspot.com/*
- Google Inside AdSense : *http://adsense.blogspot.com/*
- Yahoo! Search Blog : *http://www.ysearchblog.com/*
- Bing : *http://www.bing.com/community/blogs/search/default.aspx*

- Ask.com : *http://blog.ask.com/*
- Exalead : *http://blog.exalead.fr*
- Kartoo : *http://blog.kartoo.com/fr/*

Les forums de la recherche d'information et du référencement

Il existe de nombreux forums, en français et en anglais, sur lesquels vous pouvez partager votre passion des moteurs de recherche.

Forums en français sur les outils de recherche et le référencement

- Forums Abondance : *http://www.forums-abondance.com/*
- Forum Référencement : *http://www.forum-referencement.net/*
- Promoweb : *http://www.promo-web.org/*
- Outils de recherche et référencement : *http://forum.taggle.org/*
- WebMaster Hub : *http://www.webmaster-hub.com/*
- WebRankInfo : *http://www.webrankinfo.com/*
- Zorgloob : *http://www.zorgloob.com/forum/index.php*

Forums en anglais sur les outils de recherche et le référencement

- Cre8asite : *http://www.cre8asiteforums.com/*
- HighRankings : *http://www.highrankings.com/forum/*
- IHelpYou : *http://www.ihelpyouservices.com/forums/*
- Search Engine Forums : *http://www.searchengineforums.com/*
- SearchEngineWatch.com Forums : *http://forums.searchenginewatch.com/*
- Searchguild : *http://www.searchguild.com/*
- WebMasterWorld : *http://www.webmasterworld.com/*

Les associations de référenceurs

Ces associations regroupent les référenceurs professionnels en France, en Europe et dans le monde.

- SEO Camp : *http://www.seo-camp.org/*
- Sempo : *http://www.sempo.org/*
- SEOPros.org : *http://www.seopros.org/*

Les baromètres du référencement

Ces sites tentent de fournir des informations sur les parts de marché des différents outils de recherche sur le Web.

Baromètres français

- Wysistat : *http://www.wysistat.net/panorama/*
- Secrets2moteurs : *http://barometre.secrets2moteurs.com/*

Baromètres anglophones

Les sites ci-dessous publient parfois des chiffres sur les parts de marché des outils de recherche dans le monde anglophone.

- ComScore : *http://www.comscore.com/*
- Hitwise : *http://www.hitwise.com/*
- Keynote : *http://www.keynote.com/*
- Nielsen NetRatings : *http://www.nielsennetratings.com/*
- OneStat.com : *http://www.onestat.com/*

Lexiques sur les moteurs de recherche et le référencement

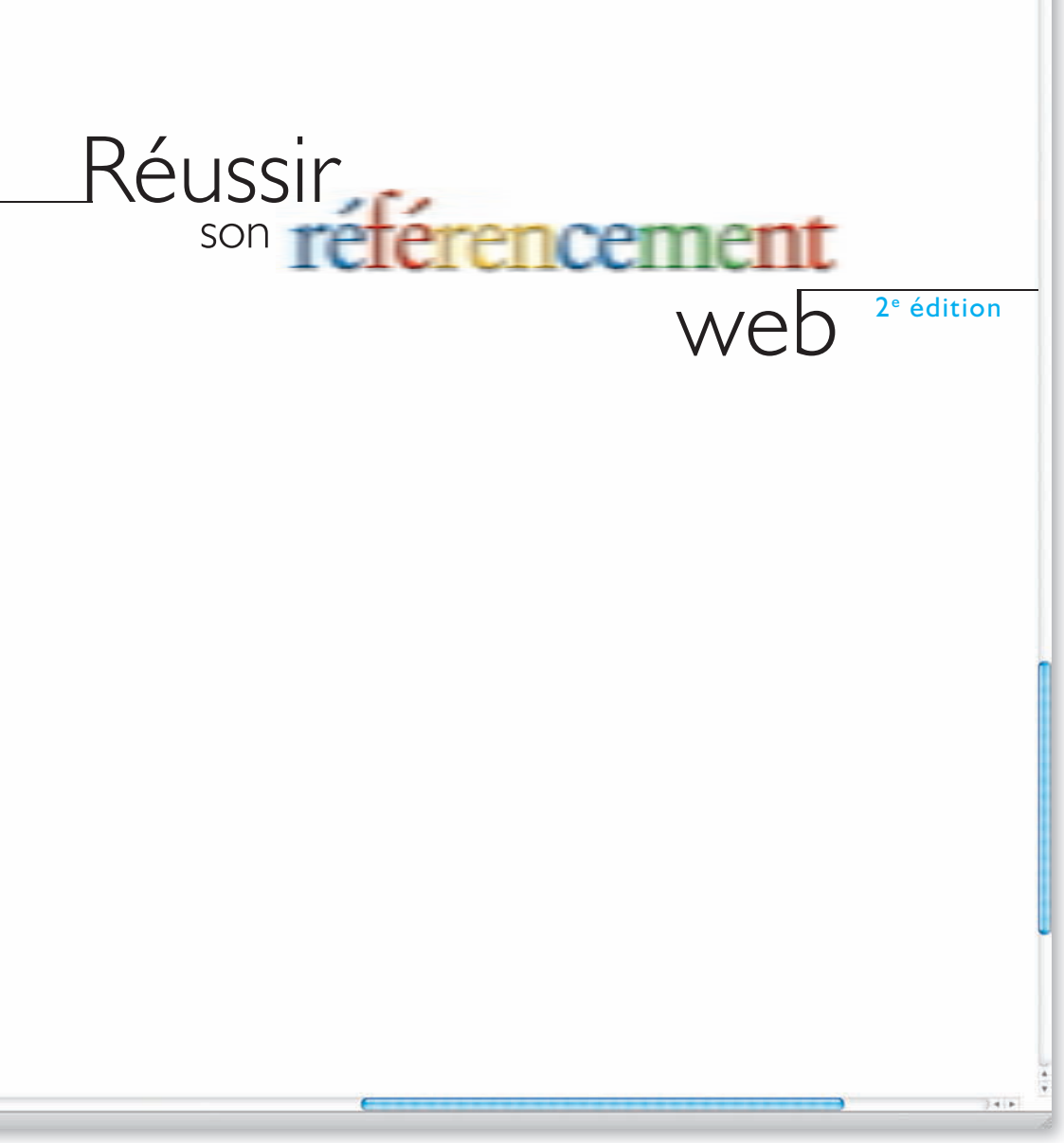
Ces lexiques en ligne vous permettront d'obtenir des définitions plus précises sur certains termes ayant trait au monde des moteurs de recherche et du référencement.

- Dicodunet : *http://www.dicodunet.com/definitions/moteurs-de-recherche/*
- SumHit : *http://www.sumhit-referencement.com/savoir-lexique.asp*
- WebRankInfo : *http://www.webrankinfo.com/lexique.php*

Index

Réussir son référencement web

2e édition



Olivier Andrieu est l'un des experts français les plus renommés du référencement et des moteurs de recherche sur Internet. Fondateur du site Abondance, qui est considéré comme le portail de référence sur ces thèmes dans le monde francophone, il a écrit plus d'une quinzaine d'ouvrages sur le sujet, dont le best-seller *Créer du trafic sur son site web* (Éditions Eyrolles). Il est également l'auteur du premier livre en langue française sur Internet, paru en 1994.

www.abondance.com

Au sommaire

Conception : Nord Compo