



Livre blanc

Référencement
Ce qu'il faut savoir

Patrice Bertrand – Directeur des Opérations

Version 1.1

Pour plus d'information : www.smile.fr

Tél : 01 41 40 11 00

Mailto : sd@smile.fr

PREAMBULE

Smile

Fondée en 1991, Smile est une société d’ingénieurs experts dans la mise en œuvre de solutions Internet et intranet. Smile compte 240 collaborateurs au 1^{er} janvier 2008.

Le métier de Smile couvre trois grands domaines :

- La conception et la réalisation de sites Internet haut de gamme. Smile a construit quelques uns des plus grands sites du paysage web français, avec des références telles que Cadremploi ou Explorimmo.
- Les applicatifs Intranet, qui utilisent les technologies du web pour répondre à des besoins métier. Ces applications s’appuient sur des bases de données de grande dimension, et incluent plusieurs centaines de pages de transactions. Elles requièrent une approche très industrielle du développement.
- La mise en œuvre et l’intégration de solutions prêtes à l’emploi, principalement dans les domaines de la gestion de contenus, des portails, du commerce électronique, du CRM et de la Business Intelligence.

Smile ne propose pas de prestation spécifique de référencement, considérant que, pour beaucoup de sites, l’essentiel tient en quelques bonnes pratiques, énoncées ici, et surtout dans l’intelligence des contenus, qui est avant tout du ressort de nos clients.

Néanmoins, Smile crée et met en œuvre des sites qui intègrent dès leur conception, l’essentiel des bonnes pratiques permettant un référencement optimal.

Quelques références de Smile

Intranets - Extranets

- Société Générale - Caisse d'Épargne - Bureau Veritas - Commissariat à l'Energie Atomique
- Visual - Vega Finance - Camif - Lynxial - RATP - AMEC-SPIE - Sonacotra - Faceo - CNRS
- AmecSpie - Château de Versailles - Banque PSA Finance - Groupe Moniteur - CIDJ - CIRAD
- Bureau Veritas - Ministère de l'Environnement - JCDecaux - Ministère du Tourisme
- DIREN PACA - SAS - Institut National de l'Audiovisuel - Cogedim - Ecureuil Gestion
- IRP-Auto - AFNOR - Conseil Régional Ile de France - Verspieren - Zodiac - OSEO - Prolea
- Conseil Général de la Côte d'Or - IPSOS - Bouygues Telecom - Pimkie Diramode
- Prisma Presse - SANEF - INRA

Internet, Portails et e-Commerce

- cadremploi.fr - chocolat.nestle.fr - creditlyonnais.fr - explorimmo.com - meilleurtaux.com
- cogedim.fr - capem.fr - editions-cigale.com - hotels-exclusive.com - souriau.com - pci.fr
- gdf.fr/presse - dsv-cea.fr - egide.asso.fr - osmoz.com - spie.fr - nec.fr - sogeposte.fr
- metro.fr - stein-heurtey-services.fr - bipm.org - buitoni.fr - aviation-register.com - cci.fr
- schneider-electric.com - calypso.tm.fr - inra.fr - cnil.fr - longchamp.com - aesn.fr
- Dassault Systemes 3ds.com - croix-rouge.fr - worldwatercouncil.org - projectif.fr
- editionsbussiere.com - glamour.com - fraterl.org - tiru.fr - faurecia.com - cidil.fr - prolea.fr
- ETS Europe - ecofi.fr - credit-cooperatif.fr - odit-france.fr - pompiersdefrance.org
- watermonitoringliance.net - bloom.com - meddispar.com - nmmedical.fr - medistore.fr
- Yves Rocher - jcdecaux.com - cg21.fr - Bureau Veritas veristar.com - voyages-sncf.fr
- eurostar.com - AON conseil - OSEO - cea.fr - eaufrance.fr - banquepsafinance.com
- nationalgeographic.fr - idtgv.fr - prismapub.com - Bouygues Construction
- Hachette Filipacchi Media

Applications métier

- Renault - Le Figaro - Sucden - Capri - Libération - Société Générale - Ministère de l'Emploi
- CNOUS - Neopost Industries - ARC - Laboratoires Merck - Egide - Bureau Veritas
- ATEL-Hotels - Exclusive Hotels - Ministère du Tourisme - Groupe Moniteur - Verspieren
- Caisse d'Épargne - AFNOR - Souriau - MTV - Capem - Institut Mutualiste Montsouris
- Dassault Systemes - Gaz de France - CFRT - Zodiac - Croix-Rouge Française

Systemes documentaires Xml

- Centre d'Information de la Jeunesse (CIDJ) - Pierre Audoin Consultants - EDF R&D

Ce livre blanc

C'est à dessein que ce livret ne s'intitule pas « Référencement – secrets d'experts » : son but est bien de présenter les principes fondamentaux du référencement, tant du point de vue des techniques sous-jacentes que des démarches visant à l'optimiser.

Avant de faire appel à un prestataire spécialisé dans l'optimisation du référencement, il conviendrait que chaque responsable de site connaisse ce *minimum* que nous présentons ici.

Il y a beaucoup d'idées fausses concernant le référencement, l'une d'elles étant qu'il suffit de payer un bon prestataire pour être dans les premières pages de Google.

La première chose que nous aimerions transmettre dans ce recueil est que le référencement n'est pas une sorte de sorcellerie aux recettes cryptiques et mystérieuses, mais un processus tout à fait raisonné, qui consiste plutôt à *mettre en avant* la pertinence réelle de votre site plutôt qu'à *faire croire* à une pertinence qu'il n'aurait pas.

SOMMAIRE

PREAMBULE.....	2
SMILE.....	2
QUELQUES REFERENCES DE SMILE	3
CE LIVRE BLANC	4
SOMMAIRE.....	5
LES BASES.....	7
LE SERVICE AUX INTERNAUTES	7
REFERENCEMENT POURQUOI ?	8
UN JEU A SOMME NULLE ?.....	9
L’ORDRE DE TRI.....	10
INDEXATION	12
LE CRAWLER	12
LES LIMITES DU CRAWLER	13
LES FRAMES.....	14
ATTENTION AUX LIENS CASSES.....	15
LE FICHIER ROBOTS.TXT	16
PERTINENCE.....	18
LE POIDS DES MOTS	18
LES URLS	19
TITRES.....	21
TEXTE DES LIENS.....	22
META/KEYWORDS	23
LES OUTILS DE GESTION DE CONTENU.....	24
URL STABLES, SIGNIFIANTES ET UNIQUES.....	25
NOTORIETE ET PAGERANK.....	27
L’ALGORITHME DE PAGERANK.....	27
UN CRITERE DIFFICILE A TROMPER	28
LES ECHANGES DE LIENS.....	29
QUELQUES ASTUCES DE LA GESTION DES LIENS	30
LA BARRE GOOGLE	30
GOOGLE SITEMAPS.....	31
LA DEMARCHE.....	33
LA VRAIE PERTINENCE.....	33
QUELS MOTS POUR ARRIVER A MON SITE ?.....	34
QUELS MOTS RECHERCHAIENT MES VISITEURS ?.....	35
QUELS LIENS POINTENT VERS MON SITE ?.....	36
CONTENU ORIGINAL ET COPYRIGHT.....	36
LE VOLUME COMPTE.....	37
LES RUSES.....	39

LES MOTS INVISIBLES 39
DES RESEAUX DE PAGES CREUSES 40
LES PAGES SPECIALES MOTEUR 41
LA PUNITION DES FRAUDEURS 42
EN CONCLUSION 44
ANNEXE – FORMULE DE PR 45

LES BASES

Le service aux internautes

Mettons-nous un peu à la place d'un moteur de recherche du web. Son objectif est de servir ses visiteurs, en les aidant à trouver rapidement l'information qu'ils recherchent. Donc de présenter les résultats de recherche dans l'ordre de *pertinence*. Bien sûr la notion de *pertinence* est très subjective, et la tâche du moteur est précisément de quantifier cette pertinence d'une manière qui corresponde *le plus souvent* aux attentes des internautes.

Si vous tapez 'Microsoft Word' dans un moteur de recherche, il est probable qu'il faut vous présenter en première position la page du site Microsoft consacrée à Word. Il y a quantité de sociétés qui proposent de vendre des licences Word, du consulting ou des add-ons sur ce produit, et les pages de leurs sites peuvent contenir les mots Microsoft Word autant que celles du site Microsoft. D'autant plus que s'il s'agit de compter les mots, ces sociétés feront en sorte d'y mettre les mots qu'il faut en nombre suffisant.

Le travail du moteur de recherche est de parvenir à distinguer la page Word du site Microsoft et les pages éventuellement consacrées à Word sur le site www.smile.fr.

Ce travail doit obligatoirement être totalement automatisé, puisqu'il porte sur des milliards de pages : il est hors de question qu'un intervenant humain passe 15 secondes à évaluer la pertinence de chaque page.

Enfin, la tâche du moteur de recherche est rendue plus difficile encore par le fait que les gestionnaires de sites ont pour objectif avoué de le tromper ! Le moteur veut établir de manière automatique la vraie pertinence de chaque page, le gestionnaire du site veut faire croire que son site est plus pertinent qu'il ne l'est réellement.

On a donc une vraie opposition, une guerre interminable, entre moteurs et webmasters. Si le moteur se laisse tromper par les sites, il perd sa crédibilité. Il lui faut donc trouver toujours plus d’algorithmes qui ne pourront être abusés par les webmasters.

Dans cette course, les moteurs de recherche ont pris une longueur d’avance depuis l’arrivée de Google : il est devenu sensiblement plus difficile de faire croire à une pertinence qui serait absente.

Cela a fait la réussite de Google, mais cela a été aussi un bénéfice pour l’Internet en général, en redonnant sa place à la vraie pertinence.

Référencement Pourquoi ?

Les internautes accèdent à un site de trois manières : (a) en tapant directement l’URL ou en la sélectionnant dans un signet (*bookmark*), (b) en suivant un lien depuis un autre site, et (c) par une recherche sur un moteur de recherche.

Pour trouver un site qu’ils ne connaissaient pas auparavant, seules restent les voies (b) et (c), et différentes études estiment que le moteur de recherche est la manière utilisée dans plus de 70% des cas pour découvrir un site que l’on ne connaissait pas.

Lorsqu’ils utilisent un moteur de recherche, il est évident que les internautes ne peuvent parcourir plus de quelques pages de réponse, et qu’en conséquence seuls les sites figurant sur les premières pages seront visités.

Il est donc d’une importance primordiale de figurer en bonne place dans les résultats de ces recherches si l’on veut attirer des visiteurs de cette manière. Tout le monde le sait, et c’est la raison pour laquelle le référencement est devenu une spécialité à part entière dans le monde des technologies Internet. Les anglophones appellent cette activité *Search Engine Optimization*, c’est-à-dire optimisation pour les moteurs de recherche, ce qui est plus explicite finalement car il ne s’agit pas d’être référencé, mais bien d’optimiser ce référencement.

Etant donné les milliards de pages indexées par un moteur de recherche – environ 10 milliards pour Google à la date d’impression –, il est naturellement de plus en plus difficile d’espérer figurer sur la première page pour des recherches larges, disons par exemple « télévision » pour un vendeur de télévisions. Pour la plupart des sites, il vaut mieux se fixer des objectifs moins ambitieux, et viser un bon rang pour des recherches plus ciblées, sur des couples ou des triplets de mots. De plus en plus, les internautes chevronnés savent qu’une recherche trop vague ne sera pas utile, et ils saisissent dès le début une petite liste de mots-clés. Ainsi, les recherches portant sur 3 mots seraient passées de 12% en 2002 à 17% en 2005 (source Ad’Oc).

Un jeu à somme nulle ?

Figurer en première page est un peu un jeu de dupe.

Vos 10 principaux concurrents ont payé une agence de référencement 5000 Euros chacun pour être en première page et ils y sont. Vous payez vous-mêmes votre dû et vous voilà en première page, éjectant l’un de vos concurrents en deuxième page. Furieux, celui-ci appelle l’agence chargée de son référencement, qui lui refacture un peu, bricole un peu plus ses pages, et le ramène victorieusement en première page. Vous appelez votre agence, dépensez encore un peu d’argent, et ainsi de suite...

L’amélioration du référencement est ce qu’on appelle un jeu à somme nulle. Plus précisément, c’est la somme des gains de position qui est nulle, pas les sommes dépensées.

Et pour autant, personne ne peut se permettre d’abandonner le combat.

L'ordre de tri

Il y a trois facteurs dans la qualité du référencement.

Le premier est bien sur « l'indexabilité », c'est à dire la compatibilité avec les mécanismes utilisés par le moteur pour parcourir et indexer les sites.

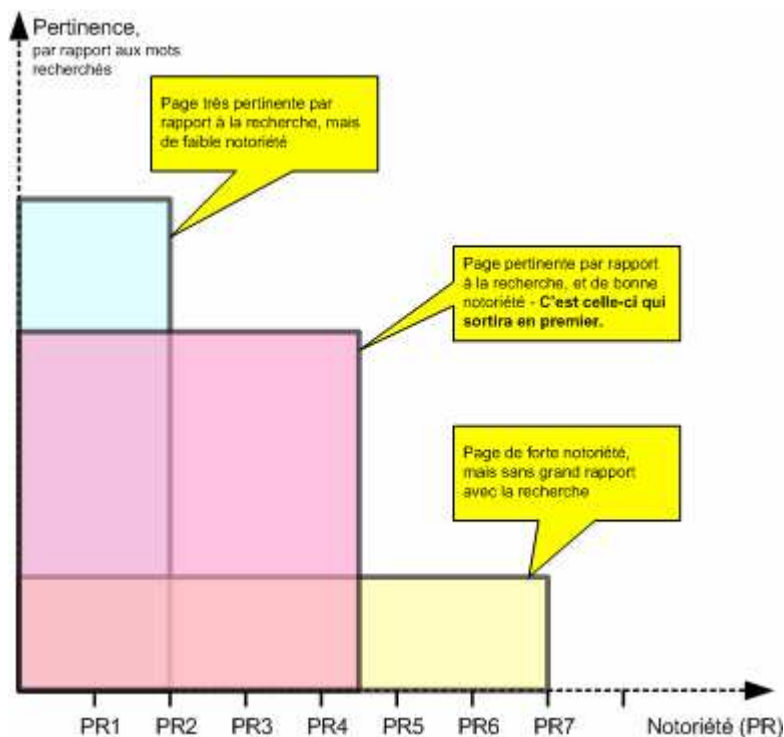
C'est la condition nécessaire. Une fois le site référencé, c'est à dire connu du moteur, la question est bien sûr celle de l'ordre de tri pour une recherche donnée.

L'ordre de présentation des résultats d'une recherche dépend de deux facteurs :

- La **notoriété**, qui est une valeur intrinsèque de la page, une mesure *sans rapport avec les mots recherchés*. C'est ce que Google appelle le *Page Rank*, ou « *PR* ».
- La **pertinence** de la page **pour les mots recherchés**, c'est à dire sa plus ou moins grande correspondance avec ce que recherche l'internaute.

Ces deux notions sont totalement transverses, indépendantes l'une de l'autre. Ce sont les deux notions clés qui interviennent dans le référencement, et un chapitre spécifique est consacré à chacune d'elles.

La manière dont ces deux facteurs sont combinés pour produire l'ordre de tri des résultats n'est pas connue. Elle est même jalousement protégée afin de ne pas donner trop de billes aux tricheurs.



La figure ci-dessus traduit la formule :

Ordre de sortie = pertinence x notoriété

En toute rigueur, ce n'est pas une simple multiplication, et la vraie formule n'est pas connue. Mais du moins la formule de multiplication donne une bonne approximation et traduit les faits suivants :

- A *contenu égal*, les pages de notoriété plus élevée viennent en tête ;
- A *notoriété égale*, les pages les plus pertinentes viennent en tête.

Le plan de notre document reprend chacun de ces trois facteurs : indexation, pertinence, notoriété.

INDEXATION

Le Crawler

A la base du référencement il y a le *robot d’indexation*, appelé encore *Crawler* (« celui qui avance en rampant »). C’est un petit programme, qui se comporte comme un internaute qui suivrait tous les liens qu’il rencontre. Il lit une page, analyse le contenu et indexe les mots rencontrés, puis suit tous les liens de cette page pour lire d’autres pages. Et ainsi de suite.

Normalement, le crawler devrait découvrir ainsi pratiquement tous les sites, puisqu’il suffit d’un lien vers une page de votre site pour qu’il « entre » et parcoure alors l’ensemble des pages en suivant ainsi tous les liens.

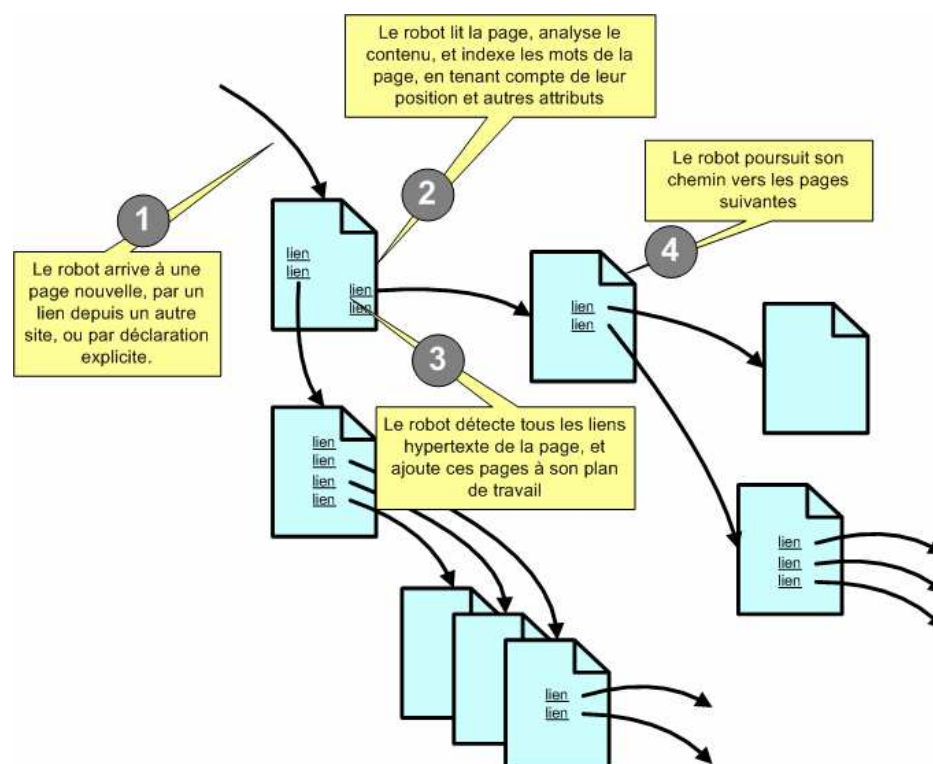


Figure 1 : Principe du Crawler

Mais si votre site est tout neuf, aucun autre site n'a encore de lien vers le vôtre. Il est toujours possible de signaler explicitement l'existence d'un nouveau site. Le moteur ne garantit pas qu'il viendra le visiter et l'indexer rapidement, mais sous quelques semaines il le fera. **C'est donc bien sûr la première étape pour indexer un site que de signaler son existence aux principaux moteurs de recherche du web.**

Certains recommandent plutôt, pour cette étape, de faire enregistrer le site de manière indirecte, par exemple en proposant la nouvelle adresse sur un site déjà référencé. Une fois le site enregistré par les principaux moteurs, un ajout manuel complémentaire peut alors être réalisé.

La fréquence de visite du crawler n'obéit pas à des règles publiées. Elle dépend du moins du taux de mise à jour du site : si le crawler voit que les contenus du site sont modifiés fréquemment, il revient fréquemment. La fréquence de visite dépend certainement aussi du *Page Rank*, de la notoriété du site : le site de Microsoft sera indexé plus fréquemment que celui des pêcheurs de la Marne. Pour autant, il serait tout à fait inutile de chercher à être indexé plus souvent, puisque cela ne donnerait en rien un meilleur référencement.

Les limites du Crawler

Le minimum requis pour que toutes les pages d'un site soient référencées est qu'il soit *crawlable*, c'est-à-dire qu'il ne présente pas d'impasse pour le fonctionnement du *crawler*.

Il faut donc bien comprendre ce que le crawler peut et ne peut pas faire.

Il suit très facilement les liens hypertextes standards (balise <a>). **Le crawler ne suit pas les liens qui résultent de l'exécution d'instructions Javascript. Et encore moins les liens inclus dans un programme Flash.**

Le crawler ne peut franchir aucun formulaire, même très simple. Aussitôt que le visiteur doit saisir un champ ou bien sélectionner dans une liste, le crawler ne passera pas. Il arrive que ce soit souhaité : certains utilisent un petit formulaire simple, par exemple une liste déroulante, pour bloquer

l’indexation de certaines pages, bien qu’il existe des méthodes plus élégantes, comme nous le verrons plus loin. Mais à l’inverse, certains sites mettent en place sans le vouloir une navigation qui bloque tout référencement.

Au strict minimum, un site doit pouvoir être visité de manière complète par le crawler.

Pour cela, il faut privilégier les liens Html naturels, interdire les liens résultant de javascript ou de Flash, et interdire les formulaires qui seraient le point de passage obligé vers certaines branches du site.

Les Frames

Cela fait quelques années déjà que l’utilisation des *frames* est abandonnée. Rappelons qu’il s’agit d’une disposition du Html qui permet de diviser la page web en sous-parties, haut/bas ou droite/gauche par exemple, chacune correspondant à une URL particulière.

Cette technique était utilisée en particulier pour conserver un bandeau permanent en entête, ou bien un menu stable en colonne gauche, ce qui économisait également la charge du serveur, du client, et la bande passante puisque tout le contenu n’était pas rechargé à chaque fois. L’avènement des outils de gestion de contenus a permis de parvenir au même rendu, mais sans les frames.

Cela étant, pour les nostalgiques, il est bon de garder à l’esprit les inconvénients des *frames* : le principe même des *frames* ne suit pas la correspondance URL-Page, qui est le fondement du web, et qui est aussi la base du fonctionnement des moteurs de recherche. Sur un site utilisant des frames, l’URL ne décrit plus qu’un morceau de page. Et il est courant dès lors que le lien cliqué depuis Google ne permette pas de retrouver la page d’origine dans son ensemble.

**Attention aux liens
cassés**

Imaginons que votre site ait conquis une petite notoriété. Vous aviez une page passionnante sur les tapis persans du XVIIème siècle et plusieurs sites spécialisés y ont fait référence, apportant ainsi un peu de leur propre notoriété à cette page. Et de là, cette notoriété se propage, comme on le verra, à l’ensemble de votre site.

Mais un jour, vous réorganisez tout cela et ladite page change d’URL. Ou bien pire, vous changez de technologie et ce sont *toutes vos pages* qui changent d’URL. Les liens entrants tombent alors en erreur (NOT FOUND !) et n’apportent plus leur poids à votre site. Après quelques passes, votre *ranking* s’effondre.

Il est fondamental d’analyser toutes les erreurs NOT FOUND (404) générées par votre site et d’en faire une chasse implacable.

Cela à la fois pour le confort de vos visiteurs – concernant les liens internes – et pour la qualité de votre référencement, concernant les liens entrants.

Il faut conserver la plus grande stabilité dans l’organisation de votre site et ses URLs. Si vous modifiez une page, elle doit conserver la même URL. Si une page est supprimée, elle doit être remplacée systématiquement par une instruction de redirection vers une autre page de votre site, par exemple l’accueil.

Si les URLs ont changé mais que les pages sont les mêmes, alors la bonne pratique est de retourner un code « HTTP 301 : moved permanently » signifiant le changement d’adresse définitif de la page.

Et en outre, il faudra prévenir les sites qui avaient des liens vers le votre. Cela demande de savoir les identifier, ce que nous verrons plus loin.

Il faut souligner également que l’utilisation d’un CMS (outil de gestion de contenu) peut être un atout important pour le référencement, dans la mesure où il permet de *tisser des liens* assez denses entre les articles tout en respectant la cohérence de la navigation, ce qui est difficile à réaliser en statique.

Le fichier Robots.txt

Les robots crawlers sont bien élevés. Surtout ceux des grands moteurs de recherche.

D'une part ils se signalent au site, c'est-à-dire qu'ils ne se font pas passer pour un utilisateur normal utilisant un navigateur normal. Ils se font connaître en renseignant dans leurs requêtes un champ particulier (user-agent), qui permet de les reconnaître. Ainsi, un site peut analyser ce champ, identifier le crawler, et présenter le cas échéant des pages différentes de celles que voient les visiteurs normaux. Nous verrons que cela peut faire partie des techniques visant à optimiser le référencement.

D'autre part, les robots respectent scrupuleusement les consignes qui leurs sont données par le site visité. Avant de visiter un site, le crawler demande à lire un fichier situé à la racine du site, et nommé *robots.txt*. Ce petit fichier, lorsqu'il existe, donne des instructions au robot, en particulier pour lui préciser le rythme de requête qu'il doit respecter, afin de ne pas submerger le serveur, ainsi que les 'branches' du site qu'il ne doit pas indexer.

Les indications peuvent distinguer l'un et l'autre des robots visiteurs.

Par exemple les lignes suivantes interdisent les répertoires /cgi-bin/ et /images/ aux robots.

```
User-agent: *  
Disallow: /cgi-bin/  
Disallow: /images/
```

Tandis que la ligne suivante interdit seulement le document email.htm au robot de Google :

```
User-agent: googlebot  
Disallow: email.htm
```

Un autre exemple, pour un site voulant rester « secret » :

```
User-agent: *  
Disallow: /
```

Autrement dit : « à tous les robots : n'indexez rien ! ».

Pour plus d’informations :

http://www.searchengineworld.com/robots/robots_tutorial.htm, par exemple.

En fait, il semble que l’utilisation de robots.txt tombe un peu en désuétude ; on observe que beaucoup de grands sites n’en font pas usage. Sans doute considèrent-ils que le fonctionnement par défaut des robots les satisfait.

PERTINENCE

Le poids des mots

Comme on l'a indiqué plus haut, deux mécanismes se combinent pour déterminer l'ordre des résultats d'une recherche : la pertinence *par rapport aux mots recherchés*, et la notoriété des pages.

Un premier principe, fondamental, est que les mots n'ont pas le même poids selon qu'ils apparaissent dans le titre d'une page, dans une entête, ou dans le corps d'un article.

L'ordre de préséance au sein de la page est le suivant :

1. Dans le nom de domaine. Ce n'est pas le plus facile à 'travailler', mais certains s'appliquent à définir des sous-domaines portant des mots choisis.
2. Dans l'URL
3. Dans le titre de la page, au sens Html (TITLE)
4. Dans des titres intermédiaires, selon leur importance (H1, H2, ...)
5. En caractères accentués (gras, ou 'strong').
6. Les mots du haut de page ont un poids plus important que les mots du bas de page.

Ce qui est représenté sur la figure suivante.

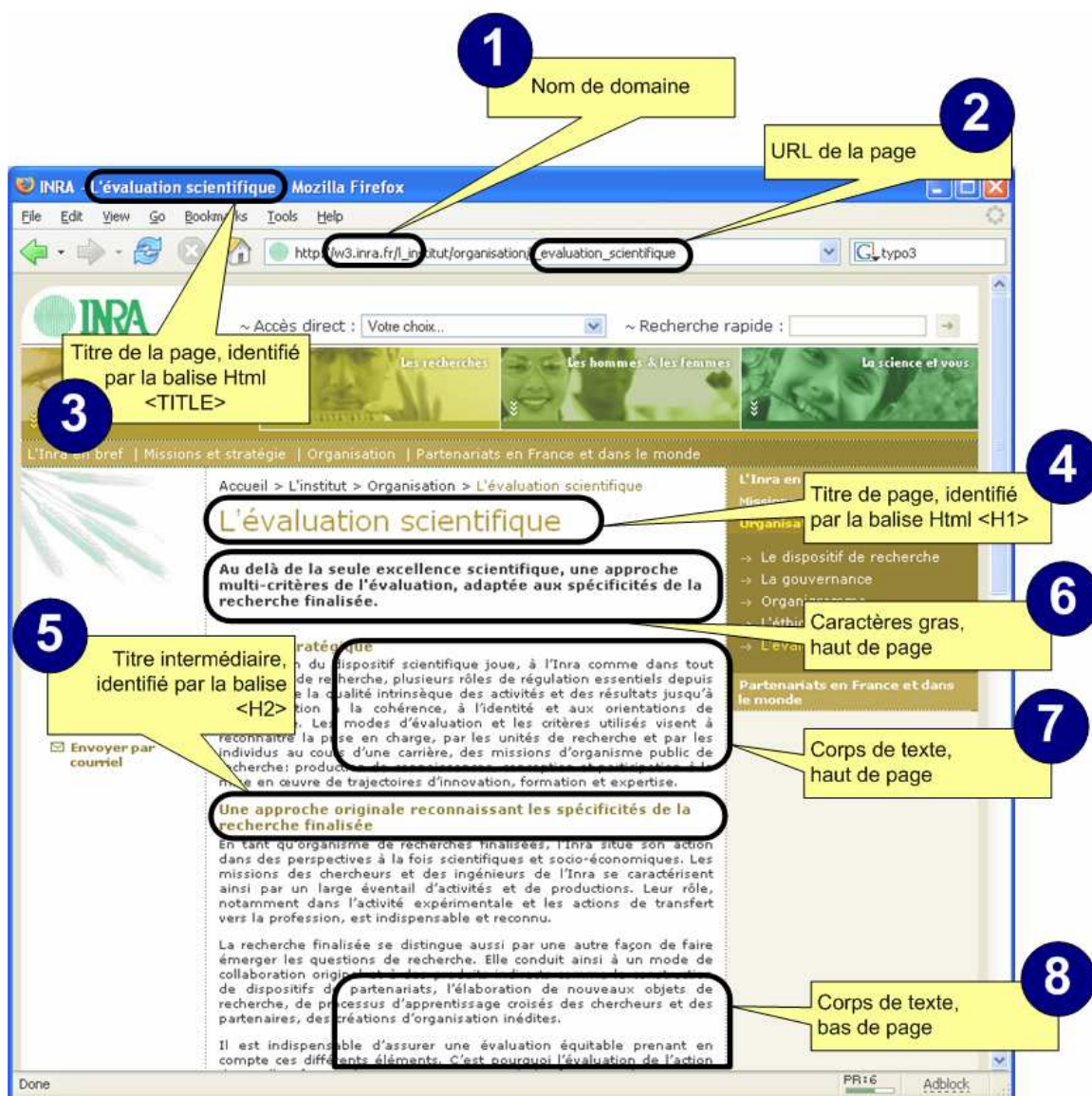


Figure 2 : Les éléments importants d'une page bien référencée

Les URLs

Pour ce qui est du cas des URLs, considérons quelques cas d'école :

[Location appartement Paris location appartements Paris](#)

Trouvez votre **location d'appartement** en région parisienne, Paris.
www.apartment-paris.com/fr/location-appartement-paris/ - 20k - 17 juin 2005 -
[En cache](#) - [Pages similaires](#)

Celui-ci, par exemple, a choisi un nom de domaine correspondant directement aux critères de recherche ciblés. Notez que ce n'est pas une faute d'orthographe, la dénomination est destinée aux anglophones. Le même a également des sous-domaines nommés spécifiquement :

[Location saisonniere Paris appartement meuble semaine](#)

Trouvez votre **location** saisonnière à **Paris**. ... Visitez nos partenaires] LODGIS, **location d'appartement** à **Paris** New-York et dans le monde ...
vacation.apartment-paris.com/fr/ - 20k - 17 juin 2005 - [En cache](#) - [Pages similaires](#)

L'exemple suivant a une URL beaucoup moins parlante, on en conviendra... Un exemple à ne pas suivre.

[Location appartement paris France location de vacances annonces](#)

Location appartement paris region parisienne maison vacances en France ... consulter petites annonces **location appartement** de particulier **paris** france ...
pageperso.aol.fr/_121b_9WPF77UI/MqVDx9XXhUwvrAVW9eUfdX+nVDQPcU9BKVOFxxH69g9yVRKulQX6HqN - 49k -
[En cache](#) - [Pages similaires](#)

Titres

Considérons un autre exemple – à ne pas suivre :



Ici le site corporate de Sagem (2005). Le titre (<TITLE>) de la page est « Homepage » : non seulement il ne porte pas de mots-clés, mais il ne mentionne pas même le nom de l'entreprise. Une occasion manquée pour un bon référencement. De plus, dans une liste de favoris, ce lien apparaîtra comme « Homepage », sans plus d'information.

Une autre conséquence, importante, de cette pondération des mots dans la page est la suivante :

Il faut utiliser les vraies indications de titres du Html (H1, H2, ...) plutôt que des styles spécifiques.

Des styles spécifiques auront peut-être un rendu de titres, mais ne pourront pas être compris comme des titres par le robot d'indexation.

C'est-à-dire qu'il faut définir <H1>Le Référencement</H1> plutôt que par , ou encore <p style=...>. Dans le premier cas on énonce clairement que l'expression « le référencement » a un rôle de titre de chapitre de premier niveau, un rôle important donc. Dans les cas de mise en forme directe, ce n'est pas aussi clair pour le robot.

Bien entendu, on utilisera une feuille de style pour définir la mise en forme associée aux titres H1.

Texte des liens

Les mots intervenant *dans les liens qui pointent vers cette page* ont également une forte pondération. C'est un point souvent méconnu, qu'il est important de souligner car c'est la seule information *extérieure à la page* elle-même, qui influence fortement son référencement.

On suppose que si une page B comporte un lien vers la page A et que ce lien mentionne « Microsoft Word », cela signifie que pour l'éditeur de la page B, la page A était particulièrement pertinente en rapport avec ce thème.

On aurait même pu dire que ce jugement serait d'autant plus valable que la page B appartiendrait à un autre site, un autre nom de domaine, car l'appréciation de pertinence serait plus objective. Mais c'est une chose sur laquelle il est facile de tricher et qui n'est donc sans doute pas prise en compte.

Ainsi, au sein même de votre site, il est important de choisir vos mots pour créer des liens internes.

Le texte des liens pointant vers une page est considéré comme partie intégrante de la page, avec une pondération importante.

Il faut donc éviter les liens du type génériques tels que « voir l'article ».

Par exemple :

En savoir plus sur <u>les lentilles vertes du Puy et la santé</u> [http://monsite.com/lentilles.html]
--

associe le mot « *santé* » aux « *lentilles vertes du Puy* », apportant ce mot comme contenu complémentaire à la page.

Tandis que

...Les lentilles vertes du Puy sont un trésor de santé, (<u>voir l'article</u>)

n'apporte que les mots « voir » et « article » dans l'indexation de la page citée.



Une autre conséquence est qu'il faut éviter de gérer les liens uniquement sur des images, ou du moins de systématiquement accompagner ces images servant de lien d'une balise 'ALT' qui décrit l'image – mais cela est probablement moins pertinent qu'un simple lien texte.

Meta/keywords

Une balise particulière a été prévue pour qu'un document indique lui-même les mots-clés utiles à son référencement : la balise META/KEYWORDS. Mais les tentatives de tromper les moteurs de recherche en remplissant les keywords à l'excès ont fini par avoir raison de son utilité. **Aujourd'hui, on peut bien sûr toujours remplir de la manière la plus pertinente possible la balise keywords, mais ce n'est plus la base d'un bon référencement.**

Soulignons que sur un Intranet, cette balise garde toute son utilité puisqu’il ne s’agit plus d’abuser le moteur, mais bien de construire un référencement pertinent.

Les outils de gestion de contenu

Les sites web modernes s’appuient généralement sur des outils de gestion de contenus, ou *content management systems (CMS)*, et il est donc naturel de s’interroger sur la compatibilité de ces outils avec un bon référencement.

Si vous n’êtes pas déjà familiers des principes de la gestion de contenu et des meilleurs outils en la matière, nous vous recommandons le livre blanc de Smile intitulé « *Content Management : les solutions open source* ».

Dans un site statique, les pages que voit l’internaute sont des fichiers placés dans une arborescence de répertoires. Le chemin d’accès indiqué dans l’URL est le reflet fidèle des répertoires conduisant au fichier.

Dans un site dynamique, et en particulier un site construit au moyen d’un CMS, les pages n’existent pas sur le serveur, elles sont construites au fur et à mesure qu’elles sont demandées. Les ‘contenus’, c’est-à-dire les textes, images ou documents composant le site, sont placés en général dans une base de données, d’où ils sont obtenus pour fabriquer les pages.

Cela étant, le crawler lui ne s’intéresse pas à la manière dont les pages sont fabriquées : il les demande par une requête http, comme le ferait un simple internaute, les obtient et les lit. Bien sûr dans certains cas, en regardant la forme d’une URL on peut deviner de quelle manière la page a été produite.

Mais il faut bien se souvenir du point suivant :

Le crawler ne fait pas de discrimination, les pages dynamiques ne sont pas moins précieuses à ses yeux que les pages statiques.

Il reste malgré tout quelques différences dont il faut se préoccuper :

- l’adresse générée doit permettre d’identifier chaque contenu ; certains CMS utilisent dans ce but une technique appelée *URL rewriting* (ré-écriture d’adresse) permettant d’utiliser le titre des articles et de leur rubrique, comme adresse de la page ;
- On entend dire aussi qu’il faut éviter les paramètres « Id=... » dans l’URL, que Google n’apprécierait pas, car souvent utilisés pour passer des variables de sessions. Utiliser simplement un identifiant plus explicite, tel que « ProductId=... ». On ne sait pas trop, cependant, où est la frontière des identifiants valides !
- le nombre de paramètres figurant dans l’adresse doit être le plus petit possible (il est conseillé de ne pas dépasser 3 paramètres) ;
- les balises META (description, keywords) doivent être rendues variables en fonction de chaque article ; dans le cas contraire, les moteurs de recherche pourraient considérer toutes les pages générées comme étant trop similaires et en conséquence n’en conserver qu’une.

Les contraintes qu’imposent l’utilisation d’un CMS peuvent alors être transformées en avantages, comme par exemple l’augmentation de la variance du contenu des articles.

**URL stables,
signifiantes et
uniques**

Au delà même de la problématique de référencement, la stabilité des URLs est un principe de base du web, mais un principe que certains outils ne respectent pas.

A une URL doit correspondre une page donnée de contenu. La même URL utilisée le lendemain doit fournir la même page.

L’outil de CMS, ou l’application servant les pages, ne doit pas insérer dans l’URL des données techniques variables qui ne

sont pas pertinentes pour référencer la page concernée : ni jeton de session, ni information de contexte.

A l’inverse, le CMS ne doit pas non plus utiliser d’information de contexte implicite (i.e. ne figurant *pas* dans l’URL) pour déterminer la page à présenter.

Une autre exigence simple à satisfaire par le CMS est qu’il doit permettre de définir des URLs significatives, c’est-à-dire intelligibles, c’est-à-dire du type
/www.monsite.com/societe/resultats.html et non
/www.monsite.com/cmstool ?Id=1294.

Certains CMS sauront utiliser directement le *titre* de la page pour constituer l’URL, d’autres permettront d’indiquer soi-même l’URL désirée. Mais ceux qui n’ont que des URLs cryptiques sont à bannir.

Une autre considération, moins connue, est la réciproque de la précédente : **une même page ne doit pas correspondre à plusieurs URLs différentes.** Car dans ce cas, Google flaire la multiplication artificielle des pages. On a vu ainsi des sites qui utilisaient plusieurs noms de domaine, par exemple www.monsite.com et www.monsite.fr, en servant les mêmes pages sous l’un et l’autre. **C’est une chose à ne pas faire, il faut plutôt mettre une instruction REDIRECT de l’un vers l’autre.**

NOTORIETE ET PAGERANK

L'algorithme de PageRank

En 1998, Larry Page et Sergey Brin, étudiants à Stanford University, créent le moteur de recherche Google sur la base de l'algorithme qu'ils ont mis au point : *Page Rank (PR)*.

Le principe du Page Rank, est le suivant. On considère que lorsqu'une page du web contient un lien vers une autre page, cela signifie que l'auteur de la première accordait un peu de valeur à l'auteur de la seconde puisqu'il jugeait pertinent d'y faire référence. Ainsi, si des milliers de sites de l'Internet contiennent des liens vers la page du site Microsoft consacrée à Word, c'est que cette page a quelque intérêt aux yeux de tous ceux qui y ont fait référence.

C'est donc cela qui fait que la page Word du site Microsoft arrivera en tête de votre recherche : des milliers de sites y font référence tandis que beaucoup moins feraient référence à une page du site Smile traitant du même sujet.

De manière plus précise donc :

- L'Internet, « *la toile* », constitue un immense réseau de pages, reliées entre elles par des liens hypertexte.
- Chaque page P_1 qui contient un lien hypertexte vers une page P apporte une voix, un vote, en faveur de cette page.
- Chaque page *répartit* ses votes entre toutes les pages vers lesquelles elle pointe. Si une page porte 10 liens vers 10 autres pages, alors chacun de ces liens n'apporte que un dixième du vote de la page.
- Les votes d'une page sont *pondérés* par le *Page Rank* de cette page. Un lien depuis le site www.cnn.com (PR9) vers votre site lui apporte beaucoup plus qu'un lien depuis le site lalentillevertedupuy.com (PR3).

Revenons sur ce dernier point. Les *Page Rank* de Google sont restitués sur une échelle de 0 à 10. Mais ce *PR* affiché est une représentation logarithmique du *PR* calculé. La base du logarithme n'est pas connue, et varie dans le temps, puisque c'est par définition celle qui permet à la page la plus référencée d'être à la valeur 10. Imaginons que le logarithme soit en base 10. Cela signifie qu'un lien venant d'une page notée *PR5* vaut autant que 10 liens venant d'une page *PR4*, et autant que 100 liens de pages *PR3*.

Une autre manière d'exprimer cela est qu'il faudrait 10^{10} liens de pages sans valeur (*PR0*) pour apporter autant qu'un seul lien depuis la page d'accueil du site W3C (l'un des quelques *happy few* qui ont des pages *PR10*).

Il faut savoir que toute cette mécanique porte sur des *pages* et non des *sites*. Ce n'est pas un site dans sa globalité qui est plus ou moins bien noté, c'est chacune de ses pages. Il peut y avoir une importante disparité de notes entre les pages d'un même site.

Il faut comprendre également que les liens internes à un site sont pris en compte, au même titre que les liens externes. Cela étant, les mécanismes de pondération et de répartition des votes font que les liens internes ne peuvent seuls remonter la notation d'un site dans son ensemble – ou très peu. En revanche, ils ont pour effet soit de concentrer la note sur certaines pages, soit au contraire de répartir la note. Schématiquement, un site comportant beaucoup de liens internes aura tendance à propager et moyenniser ses notes vers l'ensemble de ses pages.

Un critère difficile à tromper

L'un des effets de cette évaluation par vote est qu'elle est difficile à tromper. Certes il est toujours possible de créer des tas de pages qui pointeront vers votre site, mais ces pages elles-mêmes n'auront pas plus de poids que si elles étaient internes à votre site, car ces pages 'trompeuses' ne seront elles-mêmes référencées par personne.

Répetons que dans le calcul de *PageRank*, les pages internes et externes interviennent de la même manière. Donc si le

dispositif est difficile à tromper, ce n’est pas parce qu’il dépend de liens externes que vous ne maîtriserez pas, c’est essentiellement parce que ces liens sont pondérés par leur propre *PageRank* et qu’en conséquence il ne vous est pas possible de fabriquer des pages pointant vers votre site dont l’évaluation soit supérieure à celle du site.

Certes chaque page, même référencée nulle part, possède une petite valeur résiduelle qu’elle peut apporter par son vote. Mais il faut des milliers de pages non référencées pour apporter autant qu’un unique lien d’un site lui-même pertinent.

C’est l’une des voies de tricheries qui reste ouvertes par rapport aux algorithmes de vote : en construisant des dizaines de milliers de pages pointant vers votre accueil, vous apportez effectivement autant de micro-votes, qui finissent par peser. C’est la technique utilisée par certains, type Kelkoo et ses semblables, qui bien souvent polluent les résultats de vos recherches.

Les échanges de liens

Une traduction simple de l’algorithme *PageRank* est qu’il est bon que d’autres sites pointent vers votre site, c’est-à-dire contiennent un ou plusieurs liens hypertexte en direction de vos pages. Et cela d’autant plus que ces sites sont eux-mêmes connus.

Encore une fois, avant d’essayer de *tromper* ce mécanisme en construisant des liens trompeurs, il est largement préférable d’essayer de jouer le jeu, et d’obtenir de vrais liens, partant de vrais sites.

Si le contenu de votre site est intéressant, alors vous verrez que les liens viendront tous seuls, car d’autres trouveront opportun de faire référence à votre site. Si votre site contient un contenu unique sur l’histoire du stylo à bille, alors tous les sites évoquant ce sujet voudront faire référence à cette page.

Ensuite, vous pouvez bien sûr demander à vos partenaires de tous ordres de bien vouloir placer des liens vers votre site. Si

vous commercialisez des produits, alors ce pourra être les sites de vos distributeurs.

Si votre entreprise appartient à un groupe, alors il est intéressant que tous les sites du groupe placent des liens croisés vers les autres sites du groupe.

La limite de cette technique est dans le nombre : trop de liens dilue l'apport de chacun. Aussi là encore la qualité prime sur la quantité : privilégiez ceux avec vos partenaires et/ou des acteurs pertinents de votre domaine.

Quelques astuces de la gestion des liens

De l'algorithme du *PageRank*, il s'ensuit directement quelques astuces :

- Si toutes les pages de votre site ont un lien sur la page d'accueil, comme c'est l'usage, alors elles apportent toutes leur contribution au *ranking* (classement) de la page d'accueil, le plus important a priori ;
- Votre page d'accueil, avec son *PR* élevé, va à son tour apporter un peu de son poids aux pages de premier niveau vers lesquelles elle pointe. Mais comme on l'a vu le poids de chaque lien est le poids de la page divisé par le nombre de lien : plus il y en a, moins ils pèsent ;
- Il serait dommage d'inclure dès votre page d'accueil des liens *sortant de votre site*. Ils amoindriraient le poids des autres liens. Il vaut mieux dans ce cas proposer une page secondaire, dédiée à la proposition de liens externes.

La barre Google

La barre Google (googlebar) est un utilitaire qui se greffe sur Internet Explorer ou Firefox et permet en particulier des recherches rapides. Mais lorsque l'on se préoccupe du référencement, elle a une autre utilité, c'est d'afficher le

PageRank de chaque page visitée. Il est intéressant d'observer les PR des différentes pages de son site. En règle générale, ils diminuent en descendant à partir de la page d'accueil.

Ces extensions aux navigateurs sont la seule manière de connaître le PR des pages de son site. Il est donc impératif, pour celui qui s'intéresse à son référencement, de les installer, et de suivre le PR de ses pages.

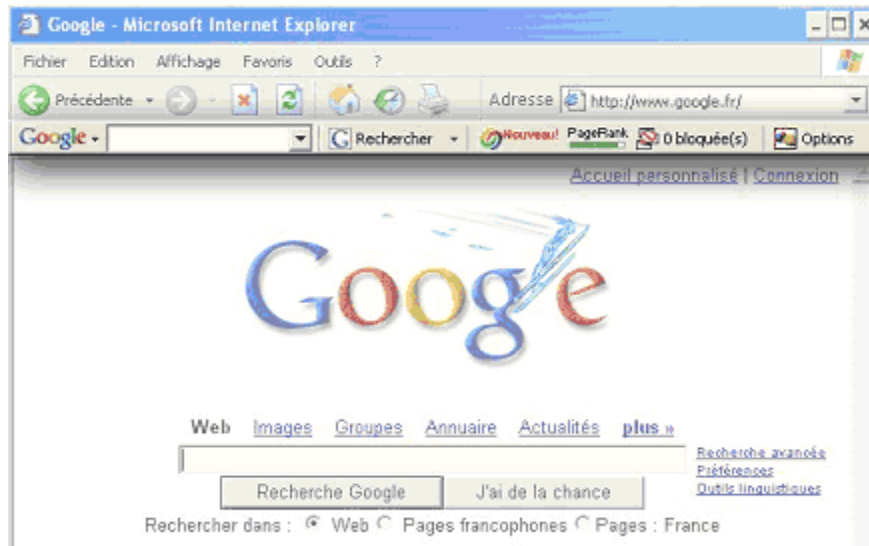


Figure 3 : Le PR de Google.fr (9), indiqué via la GoogleBar

Google SiteMaps

Google propose depuis mi-2005 un nouveau procédé d'interaction de son référencement avec les sites internet, appelé « SiteMaps ».

Google SiteMaps présente un nouveau moyen de demander l'indexation des URLs, puis d'obtenir des rapports détaillés sur la visibilité des pages sur Google.

Son utilisation repose sur la mise à disposition, par les webmasters, d'un fichier XML contenant les adresses des pages du site à référencer, ainsi que quelques infos complémentaires comme la date de dernière mise à jour. Exemple :

```
<?xml version="1.0" encoding="UTF-8"?>
<urlset xmlns="http://www.google.com/schemas/sitemap/0.84">
  <url>
    <loc>http://www.example.com/</loc>
    <lastmod>2005-01-01</lastmod>
    <changefreq>monthly</changefreq>
    <priority>0.8</priority>
  </url>
</urlset>
```

Le bénéfice pour les webmasters est une meilleure maîtrise des pages référencées ainsi qu'une bande passante réduite lors de l'analyse du site par Google (seules les pages modifiées récemment seront ré-indexées).

Le service de communication d'une entreprise sera également intéressé par le rapport détaillé, accessible via une interface, concernant le trafic dirigé par Google.

Google propose un outil gratuit (en Python) permettant aux webmasters de générer automatiquement le fichier contenant le SiteMap.

LA DEMARCHE

La vraie pertinence

L’une des premières choses à retenir pour un bon référencement est la suivante : **avant d’essayer de tromper le moteur, essayez de le satisfaire**. Considérez un peu le référencement comme la séduction : avant d’essayer d’avoir l’air subtil et spirituel et attentionné, soyez-le vraiment !

Ce sera peut-être la meilleure des recettes, et cela pour deux raisons : la première, c’est que vous obtiendrez un bon référencement sans faire des choses compliquées ou tordues, et la seconde c’est que vos visiteurs en profiteront directement puisqu’ils trouveront des informations plus pertinentes sur votre site.

Facile à dire ? Certes, mais pas impossible à faire. La vraie recette tient en peu de mots : placez sur votre site de l’information intéressante et abondante traitant des thèmes correspondant à l’indexation souhaitée. Votre site vend des fournitures de bureau ? Et bien trouvez des choses intelligentes à dire sur les fournitures de bureau. Vous devez en être capables, c’est votre métier après tout, les fournitures ! Citez des marques, des modèles, des catégories, l’histoire du stylo à travers les âges, les qualités de papier, tout est bon. Attention, pas des listes de mots placés côte à côte : non, du contenu, du vrai, non seulement *intelligible*, mais même *intelligent* si possible.

Ensuite, organisez tout ça en sections, sous-sections, ajoutez des liens internes de navigation, et voilà. Sans même tricher, vous avez fait la moitié du travail, et votre référencement est déjà assez bon. Alors imaginez en optimisant un peu !

Il vaut parfois mieux payer quelqu’un à créer du contenu intelligent pour votre site que payer quelqu’un à faire croire que ce contenu est intelligent.

Quels mots pour arriver à mon site ?

C’est toujours l’une des premières questions à se poser : **pour quels ensembles de mots est-ce que je souhaite être bien positionné ?** Si j’ai des choses à vendre, alors que recherchent mes clients ? Et plus précisément, comment mes clients exprimeront-ils leur recherche ?

C’est la première question qu’il faut se poser, et il faut se la poser avant de commencer à écrire pour son site : Comment mes visiteurs exprimeront-ils leur recherche ? Quels mots utiliseront-ils ?

Comme on l’a vu, les internautes savent de plus en plus qu’il leur faut cibler leur recherche en combinant plusieurs mots. C’est donc pour différents *groupes de mots* qu’il conviendra d’apparaître en bonne place.

Le premier exercice est donc de lister ces mots et groupes de mots par écrit, à l’occasion d’une séance de réflexion de type *brainstorming*.

Ensuite, on s’assurera que ces mots sont bien présents dans vos pages. Il arrive couramment que rédaction et référencement soient deux processus disjoints : on essaye a posteriori d’associer des mots-clés à des articles déjà écrits. Mais il est largement préférable que les textes du site utilisent effectivement les ensembles de mots choisis.

Attention également aux synonymes ou variantes. Dans le cas du site Smile par exemple, les visiteurs peuvent saisir « opensource » ou bien « open source » ou encore « logiciel libre », et d’autres équivalents encore. Il est difficile pour nous d’utiliser systématiquement tous ces mots dans un article, et le souci d’un style clair nous amènerait plutôt à choisir une formulation et à nous y tenir. Mais pour la qualité du référencement, il pourra être préférable au contraire de varier les expressions. Varier les expressions à dessein, certes, mais tout en évitant les variantes de pur style, qui au contraire pollueraient la perception.

Soyons clairs toutefois : si le vocabulaire, *pour les thèmes fondamentaux*, doit être étudié avec soin, il ne s’agit surtout

pas d'écrire *pour le référencement*, c'est-à-dire de faire des phrases qui n'auraient pas d'autre finalité que le référencement. Elles gêneraient le lecteur, sans apporter le bénéfice attendu.

Quels mots recherchaient mes visiteurs ?

La réflexion amont, évoquée ci-avant, doit être validée par une analyse en aval : quels mots avaient saisi mes visiteurs lorsqu'ils sont parvenus sur mon site par un moteur de recherche ?

Les outils de suivi d'audience tels que AWStats, WebAlyzer ou WebTrends permettent de connaître les mots-clés qu'avaient saisi les visiteurs de votre site, si c'est au moyen d'un tel moteur que l'internaute est arrivé. En effet, les mots-clés recherchés sont inscrit dans l'URL appelante, ou 'referer'.

Il est important de consulter régulièrement cette liste des mots-clés ayant conduit à votre site, pratiquement dans toute son étendue.

C'est ce qui permettra de valider ou d'ajuster les mots que vous-même utilisez pour votre référencement. Peut-être que vos visiteurs avaient une manière de formuler leur recherche qui n'était pas ce que vous attendiez. Peut-être aussi que certains visiteurs parviennent à votre site par erreur, avec des mots-clés qui ne correspondent pas à la finalité de votre site. A moins que vous ne recherchiez l'audience à tout prix, ces erreurs de routages impliqueront également un réajustement des mots utilisés pour le référencement.

Les mêmes outils, de suivi d'audience, vous donneront une autre information essentielle : la part de vos visiteurs qui sont arrivés sur votre site par l'intermédiaire d'un moteur de recherche. Il est essentiel de la connaître et de la suivre.

Si votre site connaît une chute d'audience par exemple, est-ce dû à un problème dans son référencement ? Il est fondamental de pouvoir répondre à cette question. Bien d'autres facteurs peuvent être considérés : un site concurrent

draine du trafic, un problème en hébergement a ralenti votre site et fait fuir des visiteurs, un site partenaire a retiré un lien qui amenait des visiteurs, ou tout simplement l’intérêt de vos informations a baissé.

Quels liens pointent vers mon site ?

On a vu toute l’importance des liens entrants vers votre site, surtout en provenance de sites eux-mêmes à forte notoriété. Il est donc bien sûr intéressant de connaître ces liens que d’autres ont définis vers vos pages.

Il existe une fonction de Google qui répondra à cette question : il suffit de saisir dans la commande de recherche :

link:www.monsite.com

et Google listera les pages référençant votre site, dont il a connaissance.

Il n’y a rien de particulier à en faire. Si ce n’est être reconnaissant, et savoir quelles pages ont été dignes de l’intérêt des autres. Et également, comme on l’a vu, veiller à ne pas casser ces liens entrants à l’occasion d’une réorganisation malencontreuse. Cela permet enfin de vérifier qu’aucun site « douteux » ne propose de lien vers votre site.

Contenu original et copyright

Si notre premier conseil est d’écrire pour votre site des contenus abondants et pertinents, porteurs de *valeur originale* pour vos visiteurs, un autre conseil – peut-être plus évident – est de les protéger et d’en interdire la copie.

Au delà même du droit d’auteur essentiel, la duplication de contenu est contraire au principe même du web, qui veut que l’on *fasse référence* à un contenu tiers au moyen d’un lien hypertexte permettant au lecteur d’y accéder en un simple clic s’il le souhaite.

Vous vous donnez du mal pour enrichir votre site, ne permettez pas à d’autres de reprendre à leur compte vos contenus. Si vos contenus ont une forte valeur ajoutée, alors d’autres *voudront y faire référence* depuis leurs pages, et cela amènera naturellement des liens entrants et une meilleure appréciation de vos pages. Un contenu répliqué – même s’il mentionne l’origine et l’auteur – ne vous apporte rien en termes de *PageRank* : ce que vous voulez, c’est un lien entrant.

On pourra autoriser, certes, la copie d’un court extrait, en mode *teasing*, à la condition expresse qu’il porte un lien vers le contenu original, ou bien vers votre *home*.

C’est justement le principe de la syndication de contenus par le mécanisme du RSS. Il consiste à mettre à disposition des autres sites une description en XML des contenus offerts par votre site. Ils sont libres – généralement – de reproduire ce petit extrait sur leur site, à la condition d’inclure un lien vers l’article original. Un moyen efficace d’obtenir des liens entrants et d’augmenter son *PageRank*. Le tout étant bien sûr d’avoir des choses à dire qui intéressent le reste du monde.

Le volume compte

Le nombre de pages d’un site est, en soi, un facteur de bon référencement.

On a coutume de privilégier la qualité sur la quantité, et il est clair que pour un visiteur, il serait préférable d’avoir 20 pages synthétiques et pertinentes plutôt que 200 pages diluées et redondantes.

Le service du visiteur et les besoins du référencement seraient-ils, pour une fois, contradictoires ? Pas vraiment : il suffit de mettre en lignes 200 pages toutes synthétiques et pertinentes !

Non, ce n’est pas si simple bien sûr. Mais retenons juste ce principe : le volume compte.

Pour les sites qui présentent, au moyen d’applications spécifiques, des contenus issus d’une base de données, par exemple des petites annonces d’emploi ou d’immobilier, ou bien des produits issus d’un catalogue, il y a une conséquence

toute simple : la totalité des pages de contenus doit être référencée. C’est à dire qu’il faut faire en sorte d’aménager un chemin pour le crawler qui mène vers chacune des pages de détail.

Lorsqu’on est un site d’annonce tel que Cadremploi.fr par exemple, avec 15 000 offres d’emploi en base de données, donner accès à ces 15 000 pages de contenus pertinents pour l’indexation, par rapport aux quelques centaines de pages de contenus éditoriaux, peut faire une énorme différence.

LES RUSES

Les mots invisibles

Bien sûr, l'artisan du référencement souhaite berner le mécanisme d'indexation, mais sans polluer la page que voient ses visiteurs. L'une des manières de procéder, historiquement, a été d'insérer dans la page des mots invisibles, par exemple en blanc sur fond blanc. Ou bien à l'intérieur de layers masquées. Les plus rustiques de ces techniques sont aujourd'hui détectées par les crawlers, qui non seulement n'en tiennent pas compte, mais de plus peuvent punir le fraudeur en lui affectant un référencement zéro pour une période de quarantaine.

Malgré tout, le crawler ne peut pas s'amuser à exécuter tout le code javascript inclus dans une page. Il le pourrait techniquement, mais cela lui prendrait sans doute plus de temps et de ressources qu'il ne peut en consacrer. La technique actuelle consiste donc à enchaîner une succession d'instructions javascript relativement complexes dont l'exécution va produire l'écriture d'une partie de la page Html. Le navigateur d'un vrai internaute va exécuter ce code, tandis que le crawler ne le fera pas.

Ainsi par exemple, l'instruction suivante :

```
<script>_b="nt.write('<st";_e="ility:hid";_a="docume";_g="
le>')";_c="yle>.h{disp";_d="lay:none;visib";_f="den;font-
size:1pt}</sty";eval(_a+_b+_c+_d+_e+_f+_g);</script>
```

produira l'instruction

```
document.write('<style>.h{display:none ;visibility:hidden;
font-size:1pt}</style>')
```

qui définit un style invisible nommé « .h », sans que le crawler ne le sache. Il suffit ensuite d'écrire différentes choses en utilisant ce style : ces mots seront indexés par le crawler sans être visibles des internautes.

Attention toutefois ! D'une part l'effet n'est pas magique, il dépend finalement de la 'qualité' du texte qui sera ainsi destiné au robot, et d'autre part vous faites cela à vos risques

et périls : si un jour Google devient plus rusé que vous, vous pourrez à votre tour être *blacklisté*.

Des réseaux de pages creuses

Le summum du détournement de pertinence est peut être atteint avec un site comme moteur-recherche.net, qui a fabriqué des milliers de pages vides de sens, correspondant aux paires de mots-clés recherchées par les visiteurs. Il suffit qu'un internaute tape « guide tourisme Italie » pour que le site fabrique une page `guide_tourisme_Italie.html`. Cette page contient le résultat d'une recherche sur ces mots-clés, c'est-à-dire un contenu qui semble pertinent, mais n'a en fait aucune valeur ajoutée vraie. Les moteurs de prix, comme Kelkoo et ses semblables procèdent de manière semblable : quels que soient les mots, ils ont toujours des pages à mettre en face. Ainsi, le site soumet à Google des milliers de pages stupides, dont le seul contenu est lui-même issu d'une recherche, peut-être sur Google soi-même ! A quoi sert tout ce vide ? Sans doute à créer un peu d'audience, puisque les pages de moteur-recherche.net ont provisoirement réussi à tromper le moteur de pertinence de Google, et sortent donc fréquemment en haut de classement. Et un peu d'audience, permet un peu de pub. Mais même les publicitaires devraient se méfier de telles pratiques, qui associent leurs marques à une tromperie.

La technique est donc clairement à déconseiller : à la fois très lourde à mettre en place, et assez risquée. Sans compter que fabriquer une telle pollution à grande échelle sur le web est profondément incivique.

Plusieurs prestataires en référencement (Netbooster, septembre 2004) en a d'ailleurs fait les frais : découverts, ils ont eux-mêmes été blacklistés par Google. Belle réussite, pour un expert du référencement !

Les pages spéciales moteur

Comme on l’a dit, les robots indexeurs sont bien élevés : d’une part ils respectent les instructions du fichier *robots.txt*, et d’autre part ils ne cherchent pas à se faire passer pour un internaute quelconque, ils s’identifient clairement, au moyen du paramètre *user-agent* qui est défini dans chacune des requêtes http.

User-agent permet généralement d’identifier le *navigateur*, et certains sites l’utilisent pour adresser des pages différentes selon les possibilités du navigateur cible.

Ainsi, le robot Google s’identifie en indiquant « user-agent=googlebot » dans chacune de ses requêtes.

Il est donc possible d’utiliser ce paramètre pour servir à Google des pages spéciales, différentes de celles qui seront servies aux internautes.

Cette technique a été beaucoup utilisée aux débuts du référencement, pour servir à chaque moteur d’indexation des pages correspondant à ses caractéristiques. Yahoo aimait les keywords, on lui en donnait, ... Altavista voulait des <H1> mais ne supportait pas le bourrage de keywords, on lui donnait satisfaction aussi.

C’est une technique complexe, qui demande un travail considérable, pour des résultats aujourd’hui assez faibles.

Mais elle a encore ses adeptes. Pour voir un exemple, tapez par exemple « louer appartement » sur Google, et demandez la version ‘en cache’ des pages en tête de liste.

Les mots de ma recherche sont ici «louer appartement »

Les pavés «liens commerciaux » ont été achetés à Google

Tous les sites de la première page ont construit des pages spécifiques «louer-appartement.htm». Ces pages ne sont jamais vues par les internautes, elles redirigent sur l'accueil du site

www.smile.fr

La punition des fraudeurs

On l'a dit, le référencement est une guerre sans merci. Mais dans cette guerre, les moteurs disposent de l'arme atomique et pas vous : le déréférencement ou *blacklisting*. Si le moteur de recherche décèle une tentative de tricherie, il peut *black-lister* le site dans son ensemble, c'est-à-dire que plus aucune recherche ne restituera des pages de ce site, pas même en 1000^{ème} position. Le site n'existe plus pour Google.

C'est une punition sévère, qui dure plusieurs mois. Et comme tout cela est régi par des algorithmes, sans intervention humaine, il est très difficile d'aller supplier un retour en grâce. Le cas n'est pas théorique et nombre de prestataires en référencement un peu trop inventifs s'y sont déjà brûlé les

doigts. BMW, Castorama, Ricoh, ou bien même Netbooster, en savent quelque chose.

Bien que depuis quelques temps il semble que Google ne « supprime » plus les sites qui trichent de son index, mais dévalue simplement leur *PR*, cela reste une raison suffisante pour ne pas essayer de s’y risquer.

Mais comme nous l’avons souligné plus haut, la principale raison est ailleurs : viser un meilleur référencement sans tricher, c’est aussi mieux servir vos visiteurs, en leur offrant une vraie pertinence des contenus.

EN CONCLUSION

Après plusieurs années d’expérience des acteurs de ce domaine, et l’observation de l’évolution des moteurs, il apparaît que la **qualité du fond** (richesse de contenu, pertinence, organisation, spécialisation des pages) et **de la forme** (simplicité, respect des normes, application de règles simples d’organisation du contenu) **restent les valeurs sûres** : un site **bien pensé, bien réalisé**, et **bien suivi**, devrait dans la grande majorité des cas obtenir et conserver un bon positionnement.

Du côté des moteurs, l’hégémonie de Google a permis de stimuler le web pour en augmenter la qualité. L’internaute doit toutefois rester vigilant et critique car cela pourrait entraîner des dérives et excès, et après tout, les résultats d’une recherche ne constituent qu’un seul point de vue.

En somme, que l’on soit du côté des webmasters ou du côté des internautes, le plus sûr est de conserver son bon sens.

ANNEXE – FORMULE DE PR

Revenons sur cette formule du PR, qui s'exprime comme ceci :

$$P = 0,15 + 0,85 \times (P_1/C_1 + P_2/C_2 + P_3/C_3 + P_4/C_4 + \dots P_n/C_n)$$

Dans cette formule :

- P est le 'PR' d'une page donnée en valeur brute, non logarithmique
- P₁, P₂, P₃, P₄, ... P_n représentent les 'PR' des pages qui ont un lien vers cette page.
- C₁, C₂, C₃, C₄, ... C_n représentent respectivement le nombre de liens sortants depuis les pages P₁, P₂, P₃, P₄, ... P_n, c'est-à-dire que pour chaque page, comme on l'a vu, le poids du vote est divisé par le nombre de liens : la page *répartit son vote* sur les différentes pages auxquelles elle fait référence.

Les coefficients 0,15 et 0,85 sont des facteurs de pondération qui assurent que toute page possède un PR minimum de 0,15. Cela revient à dire que le PR d'une page provient à 85% des liens entrants, les 15% restant étant une valeur minimale considérée acquise.

Le 'PR' ainsi calculé n'est pas un nombre entre 1 et 10 : il est pratiquement illimité. Supposons que 1000 pages aient un lien vers une page A, et que ces pages elles-mêmes n'aient aucun lien *entrant* (en anglais, *backlink*) et un unique lien sortant, celui qui pointe vers la page A.

Les 1000 pages en question ont donc chacune un PR de 0,15, et le PR de la page A s'exprime ainsi :

$$\begin{aligned}
 P_A &= 0,15 + 0,85 * (0,15 + 0,15 + \dots + 0,15) \\
 &= 0,15 + 0,85 * 0,15 * 1000 \\
 &= 127,65
 \end{aligned}$$

Imaginons maintenant que cette page A, dont le *PR* vaut 127,65 ait un lien vers une page B. Le *PR* de la page B s'écrit alors :

$$P_B = 0,15 + 0,85 * (127,65)$$
$$= 108,65.$$

On constate que ce lien de A vers B a énormément contribué au *PR* de B, d'avantage que le l'auraient fait 700 pages de *PR* égal à 0,15. Autrement dit, les 1000 liens entrants vers la page A ont en quelque sorte propagé leur apport, leur vote, à la page B.

Ces calculs sont en théorie appliqués à l'ensemble du gigantesque réseau de pages et de liens que constituent le web. Certains sites ont des centaines de milliers de liens entrants, par exemple www.apache.org, ou bien www.google.com, et bénéficient donc d'un *PR* maximum. Sa valeur ne nous est pas exactement connue, mais elle se chiffre vraisemblablement en milliards. Par construction, le *PR* maximum est ensuite converti sur une échelle logarithmique allant de 1 à 10. Si l'on suppose par exemple que le site le plus référencé a un *PR* de 10^{10} , c'est-à-dire 10 milliards, alors il y a un facteur 10 entre chaque numéro de *PR*. Autrement dit un site avec un *PR*5 a en fait un *PR* dix fois supérieur à celui qui n'aura que *PR*4.

Une autre manière de traduire cela, par rapport à l'apport que peuvent avoir des liens entrants vers votre site s'exprime ainsi : un lien entrant depuis un site de *PR*6 améliore votre propre *PR* autant que 100 liens entrants depuis 100 pages de *PR*4. C'est la traduction d'une échelle logarithmique, et il est bon de s'en imprégner. C'est aussi ce qui fait qu'il n'est pas facile de tricher : 1000 pages sans intérêt pointant vers votre page d'accueil ne lui apporteront pas plus qu'un seul lien depuis une page *PR*3.