

ÉDITION 2016-2017
Plus de 40 000 lecteurs !

Olivier Andrieu

Réussir son_



référencement web

Stratégies et techniques SEO

EYROLLES

Réussir son_



référencement

web

Une méthodologie infaillible

Écrit par l'un des plus grands spécialistes français du référencement, cet ouvrage de référence fournit toutes les clés pour garantir à un site Internet une visibilité maximale sur les principaux moteurs de recherche. Dédié au référencement naturel, il explique notamment comment optimiser le code HTML des pages web pour qu'elles remplissent au mieux les critères de pertinence de Google, Bing, Qwant et les autres.

Traitant de tous les aspects liés au référencement, ce livre constituera un guide précieux et complet pour tous ceux qui souhaitent renforcer la présence de leur site sur Internet. Il fournit des solutions techniques mais également des conseils pratiques pour mener à bien un tel projet : où trouver un prestataire et sur quels critères le choisir ? combien coûte un référencement ? quelles garanties doit offrir un référenceur à ses clients ? comment gérer le suivi d'un référencement ?

Totalement actualisée, cette nouvelle édition propose un contenu encore plus synthétique et accessible, distinguant les aspects fondamentaux et les notions plus avancées du référencement naturel. S'adressant aux non-spécialistes comme aux plus avertis, cet ouvrage intéressera tous les webmasters et responsables marketing qui veulent connaître les secrets d'un bon référencement sur Internet.

À qui s'adresse ce livre ?

- À tous les acteurs du Web (chefs de projet, webmasters, développeurs...)
- À tous ceux qui veulent améliorer le positionnement de leur site sur les moteurs de recherche

Au sommaire

Les bases du référencement naturel. Référencement versus positionnement • Liens organiques versus liens sponsorisés • Fonctionnement des outils de recherche • Choix des mots-clés • Sur quels moteurs faut-il se référencer ? • **Le SEO en pratique.** Optimisation des pages du site • Balises meta, attributs alt et title • Liens, PageRank et indice de popularité • Référencement multimédia et multisupport • Le SMO • Méthodologie et suivi du référencement • Retour sur investissement • Mise en place de liens de tracking • Utilisation de la longue traîne • Logiciels de suivi du ROI • Internalisation ou sous-traitance ? • Combien coûte un référencement ? • Un référencement gratuit est-il intéressant ? Où trouver et comment choisir un prestataire de référencement ? • Quelles garanties peut offrir un référenceur ? • Chartes de déontologie • **Devenir un as du SEO.** *Rich snippets* et schema.org • Optimisation de l'indexation • Index secondaire et duplicate content • Freins au référencement et solutions • Spam et pénalités, Google Panda et Penguin • Comment ne pas être référencé ? • Conclusion : la règle des 4C • Webographie.



© Catherine Theulin, Contrast photography

Olivier Andrieu est l'un des experts français les plus renommés du référencement et des moteurs de recherche sur Internet. Élu meilleur référenceur français en 2013 par le Journal du Net et fondateur du site Abondance, premier blog SEO dans le monde francophone, il a écrit plus d'une vingtaine 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.

Préface de Laurent Bourelly
Illustrations de Grifil

Réussir son référencement web

Stratégie et techniques SEO

SUR LE MÊME THÈME

O. ANDRIEU. – **SEO zéro euro.**

N°14033, 2014, 224 pages.

I. CANIVET. – **Référencement mobile.**

N°13667, 2013, 456 pages.

I. CANIVET. – **Bien rédiger pour le web.**

N°13750, 3^e édition, 2014, 736 pages.

D. ROCH. – **Optimiser son référencement WordPress.**

N°14182, 2^e édition, 2015, 296 pages.

R. RIMELÉ, R. GOETTER. – **HTML 5 – Une référence pour le développeur web.**

N°13638, 2^e édition, 2013, 752 pages.

E. MARCOTTE. – **Responsive web design.**

N°13331, 2011, 160 pages.

F. DRAILLARD. – **Premiers pas en CSS 3 et HTML 5.**

N°13944, 6^e édition, 2015, 472 pages.

M. KABAB, R. GOETTER. – **Sass et Compass avancé.**

N°13677, 2013, 280 pages.

DANS LA COLLECTION « DESIGN WEB »

A. MARTIN, M. CHARTIER. – **Techniques de référencement web.**

N°14333, 2^e édition, 2016, 410 pages.

S. POLLET-VILLARD. – **Créer un seul site pour toutes les plates-formes.**

N°13986, 2014, 144 pages.

K. DELOUMEAU-PRIGENT. – **CSS maintenables avec Sass et Compass.**

N°13640, 2^e édition, 2014, 252 pages.

J. PATONNIER, R. RIGOT. – **Projet responsive web design.**

N°13713, 2013, 162 pages.

I. CANIVET, J.-M. HARDY. – **La stratégie de contenu en pratique.**

N°13510, 2012, 176 pages.

C. SCHILLINGER. – **Intégration web – Les bonnes pratiques.**

N°13370, 2012, 390 pages.

S. DAUMAL. – **Design d'expérience utilisateur.**

N°13456, 2012, 390 pages.

G. BARRÈRE, E. MAZZONE. – **Card Sorting.**

N°13448, 2012, 128 pages.

Olivier Andrieu
Préface de Laurent Bourelly
Illustrations de Grifil

Réussir son référencement web

Stratégie et techniques SEO

ÉDITION 2016-2017

EYROLLES

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

En application de la loi du 11 mars 1957, il est interdit de reproduire intégralement ou partiellement le présent ouvrage, sur quelque support que ce soit, sans l'autorisation de l'Éditeur ou du Centre Français d'exploitation du droit de copie, 20, rue des Grands Augustins, 75006 Paris.

© Groupe Eyrolles, 2008-2016, ISBN : 978-2-212-1461-4

Préface

Pour commencer par un aveu, j'aurais bien aimé être Olivier Andrieu. C'est même rageant d'observer qu'il a tout pour lui.

Sur l'échelle de la reconnaissance par ses pairs, Olivier était déjà le n° 1 quand j'ai démarré le métier de référenceur et il n'est jamais tombé de son piédestal. Que ce soit le noyau dur de la communauté, les quelques centaines de personnes visibles en surface, ou le puits sans fond de tous ceux qui s'intéressent au référencement, il est impossible de ne pas reconnaître le rôle majeur d'Olivier Andrieu. Personne n'a autant œuvré pour la légitimité du référencement et la promotion du métier.

C'est en grande partie grâce à lui que nous avons une base saine en France, par rapport à l'approche du référencement. Son éthique a tout parié sur la qualité depuis toujours. La vérité est à l'écran et elle raconte qu'il avait raison de fuir les raccourcis et l'appât du gain.

J'admire aussi la manière dont il a mené sa carrière. Olivier trace sa route tranquillement, semblant gérer où il va et se souvenant toujours d'où il vient. Il semble faire seulement ce dont il a envie et à son rythme. C'est un bon mentor à prendre en exemple lorsqu'on veut apprendre le référencement. C'est ainsi que je suis honoré d'avoir été choisi pour rédiger la préface de la dixième édition de son livre, ouvrage de référence sur mon métier et ma passion.

Parlons-en de ce fameux livre, la célèbre bible du référencement. Avec plus de 40 000 ventes au compteur, c'est un succès impressionnant, pointant un repère sur l'intérêt pour notre secteur d'activité.

J'ai repris le sommaire de la première édition pour constater que le squelette de base n'avait presque pas changé. En 2006, les mêmes fondamentaux étaient préconisés, ce qui est rassurant : on est sur la bonne voie quand on suit la méthode Andrieu.

Pour apprendre à construire un mur comme moi ou cuisiner un gâteau comme Olivier, il va falloir emmagasiner du savoir, mais le plus important est de s'entraîner. Lancer des tests en référencement ne fait pas mal. Personne n'a jamais fini à l'hôpital en optimisant un site web. En revanche, je vous conseille de gagner du temps en partant de robustes fondamentaux.

Le livre d'Olivier est le garant d'une approche performante et pérenne. En plus d'être un excellent guide, cet ouvrage est la mémoire de notre activité. D'autres livres traitent du même sujet, mais aucun ne peut jouer ce rôle de témoin.

Au fil des éditions, l'auteur déroule l'histoire d'une activité passionnante et fascinante. Par exemple, le remaniement complet de l'approche du volet « off site » nous rappelle que nous sommes aujourd'hui très loin de la simple action d'apport de popularité ou manipulation plus ou moins franche du PageRank. Le profil de liens s'est transformé en un véritable écosystème propogateur d'autorité et de notoriété.

C'est aussi très intéressant de noter que les paramètres d'optimisation de base « on page » et « in site » ont beaucoup moins évolué que cet écosystème gravitant autour du site. Certes, beaucoup de choses sont arrivées sur Internet depuis dix ans. Le référencement a suivi cette évolution et façonné les sites web, en fonction de l'évolution des usages et des moteurs de recherche. Les progrès technologiques sont vertigineux, mais les fondamentaux de l'optimisation d'un site demeurent immuables.

Une preuve qu'Olivier était sur le bon chemin dès la première édition du livre se tient à la page 69 : « Content is KING, Optimized content is EMPEROR ». La formule est trop belle pour être traduite, mais le paragraphe en bon français qui s'en suit est remarquable.

Surtout, j'aime cet avertissement sans appel : « C'est le contenu que vous allez proposer en ligne qui va faire la différence entre trafic efficace, ciblé, et trafic stérile ». Pour remettre dans le contexte de l'époque, c'était trop facile de prendre des raccourcis, et je sais de quoi je parle. Prôner la voie de la qualité pour l'internaute avant tout a toujours été un challenge compliqué. Nous devons saluer ce choix courageux, situé dans une période où il était tentant d'arriver dans les premières positions sur Google en le goinfrant de médiocrité, sans se préoccuper de l'intérêt véritable pour l'internaute. Manipuler le moteur de recherche était trop facile pour ne pas avoir envie d'en profiter. Sauf pour Olivier Andrieu, qui avait déjà choisi sa ligne de conduite, entraînant tout un courant derrière lui.

Vous avez donc compris que vous êtes entre de bonnes mains. Suivez les conseils prodigués dans cet ouvrage et vous obtiendrez immanquablement des résultats en référencement. Grâce à cette fameuse première édition en 2006, j'ai fait de bons choix qui sont encore payants dix ans plus tard. Alors merci Monsieur Andrieu pour tout ça, et plus encore.

Laurent Bourelly

Consultant SEO et initiateur, entre autres, du concept de « cocon sémantique »
<http://www.laurentbourelly.com>

Remerciements



Je tiens à remercier ici (une nouvelle fois) toutes les personnes qui m'aident depuis plus de vingt ans (ça ne me rajeunit pas...) à suivre le « petit » monde si passionnant des outils de recherche et du référencement. Que de changements dans ce « court » laps de temps et quelques ouvrages écrits au cours des années. Et je ne parle pas de *Internet guide de connexion*, paru en 1994, déjà chez Eyrolles, avec... une disquette permettant de se connecter une heure gratuitement à Internet ! C'était Byzance et Broadway réunis. Mais je vous parle d'un temps que les moins de vingt ans ne peuvent pas connaître... Alors, si je commence à évoquer le Minitel, un domaine qui a également occupé une bonne partie de ma vie dans les années 1980, je risque de faire fuir quelques-uns d'entre vous.

Merci également à ceux qui ont collaboré à la rédaction de cet ouvrage (pour la plupart, rédacteurs de ma lettre professionnelle « Recherche et Référencement », dont certains articles ont été repris et adaptés dans ce livre).

- Jean-Noël Anderruthy, webmaster spécialisé dans les technologies Google (<http://google-xxl.blogspot.com/>), pour les sections sur les rich snippets (chapitre 11), l'indexation rapide (chapitre 12) et la vitesse de chargement des pages (chapitre 14).
- Philippe Yonnet, directeur de l'agence Search-Foresight/Groupe MyMedia, pour ses informations sur Mayday, Caffeine et Jazz (chapitre 2), les critères temporels (chapitre 5), le PageRank (chapitre 6), AMP (chapitre 7), le chiffrement HTTPS et les balises hreflang (chapitre 14).
- Christophe Deschamps, consultant et formateur en gestion de l'information, responsable du blog Outils Froids (<http://www.outilsfroids.net/>), pour son encadré sur Google+ (chapitre 8).
- Daniel Roch, consultant WordPress, référencement et webmarketing chez SeoMix (<http://www.seomix.fr/>), pour le responsive design et App Indexing (chapitre 7) et le référencement des sites en Ajax (chapitre 13).
- Guillaume Thavaud, pour ses informations sur Bing (chapitre 2), le TrustRank (chapitre 6), Google Adresses et les commentaires des internautes (chapitre 7), ainsi que sur le Knowledge Graph et Schema.org (chapitre 11).
- Sébastien Joncheray, de la société Raynette (<http://www.raynette.fr/>) pour l'exemple d'intégration des rich snippets et de Schema.org (chapitre 11).
- François Houste, LSFe-tourisme Associate Director chez LSF Interactive (<http://www.lsfinteractive.fr/>) et auteur du blog eTourisme Feng Shui (<http://www.etourisme-feng-shui.com/>), pour son aide au sujet du travail de SEO avec une agence externe (chapitre 10).
- Alexandre Diehl, avocat à la Cour, cabinet Lawint (<http://www.lawint.com/>), pour ses indications sur le statut juridique des liens hypertextes (chapitre 6).
- Emmanuel Fraysse, consultant indépendant (<http://lewebsocial.com/>), pour sa collaboration à la section consacrée aux réseaux sociaux (chapitre 8).
- Olivier Duffez (<http://www.webrankinfo.com/>), qui a contribué aux sections sur l'*URL rewriting* et les redirections (chapitre 14).

- Philippe Laine, expert SEO chez Manageo (<http://www.manageo.fr/>), pour les informations sur les Quality Updates de Google (chapitre 15).
- Antoine Mussard, de la société VRDCI (<http://www.vrdci.com/>) et Damien Henckes (consultant indépendant).

Merci à toutes les personnes proches qui sont mes « fournisseurs officiels en énergie », elles se reconnaîtront.

Merci à mes filles, Lorène et Manon, pour leur soutien. Je vous aime.

Et enfin un merci tout spécifique à mon ami Régis, ou plutôt Grifil (<http://www.grifil.com/>). Ta belle amitié m'est chère depuis plus de 15 ans. Merci d'avoir accepté, avec ta gentillesse habituelle, d'illustrer chaque partie et chapitre de ce livre depuis deux ans et l'édition de 2014. Promis, on s'appelle pour voir si on a réussi et je te dis quoi (les Ch'tis comprendront).

Olivier Andrieu

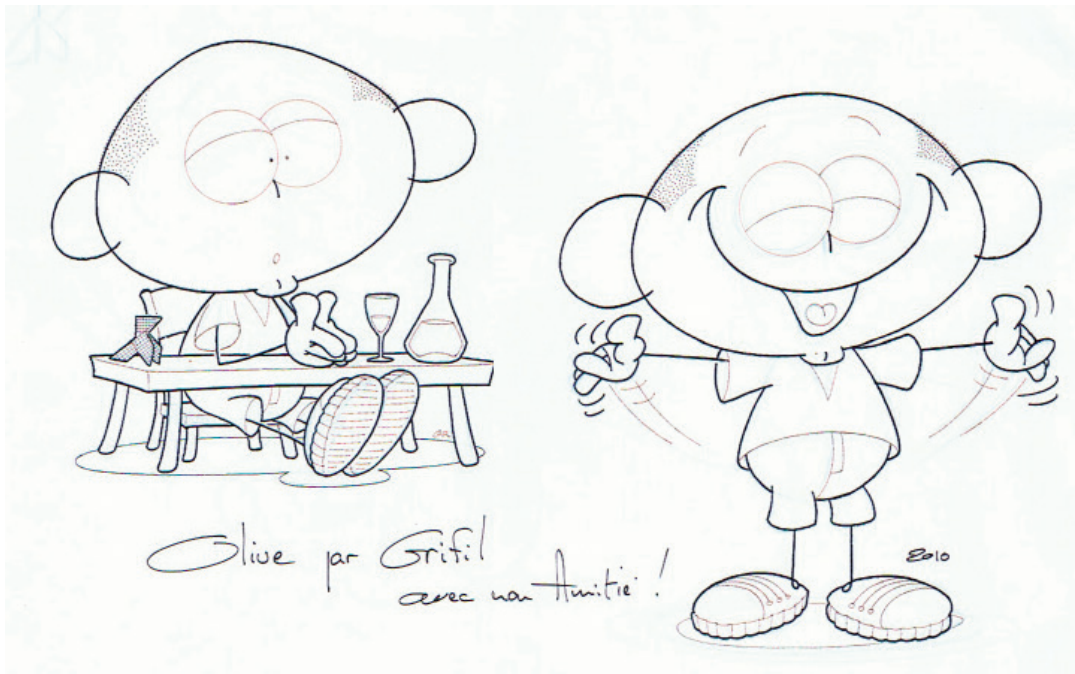


Table des matières

Avant-propos	XXV
PARTIE A	I
Les bases du référencement naturel (SEO)	1
CHAPITRE 1	
Le référencement aujourd'hui : généralités.....	3
Référencement versus positionnement.....	4
Liens organiques versus liens sponsorisés.....	6
Les trois étapes à respecter lors d'un référencement sur un moteur de recherche.....	10
Positionnement, oui, mais où ?	12
Référencement et course à pied.....	22
Deux écoles : optimisation loyale versus spamdexing	25
Le SEO, c'est comme une recette de gâteau !.....	27
CHAPITRE 2	
Fonctionnement des outils de recherche.....	31
Comment fonctionne un moteur de recherche ?	32
Technologies utilisées par les principaux portails de recherche	32
Principe de fonctionnement d'un moteur de recherche	34
Les crawlers ou spiders.....	34

CHAPITRE 3

Préparation du référencement	55
Méthodologie à adopter	56
Choix des mots-clés	56
Le concept de « longue traîne »	57
Comment trouver vos mots-clés ?	64
Utilisez Google Suggest pour trouver les meilleurs mots-clés	67
Fautes de frappe et d'orthographe	72
Intérêt d'un mot-clé	75
La prise en compte du nombre de résultats sur Google	79
Méthodologie de choix des mots-clés	80
Un arbitrage entre intérêt et faisabilité	85
Le concept de cocon sémantique	85
Le référencement prédictif	86
Google Trends, un outil indispensable	87
Étape 1 – Définir un univers sémantique simple	87
Étape 2 – Effectuer une recherche prédictive	89
Étape 3 – Détecter le début du pic des requêtes et commencer à proposer du contenu un mois avant	91
Sur quels moteurs faut-il se référencer ?	93

PARTIE B 97

Le SEO en pratique 97

CHAPITRE 4

Optimisation – Les critères in page : balises HTML et URL ..	99
Regardez vos pages avec l'œil du spider !	100
Le cache de Google	100
Les simulateurs de spider	102
Autres possibilités	103
Zone chaude 1 : la balise <title>	105
Un titre pour chaque page !	106
Zone chaude 2 : la structuration du texte en balises <h1>	106
Zone chaude 3 : la mise en gras	110

Et si le gras est dans une feuille de styles ?	110
Zone chaude 4 : les liens internes	111
Les balises meta	112
Moins d'importance aujourd'hui	112
Seules comptent les balises meta description et robots	113
Zone chaude 5 : la balise meta description, à ne pas négliger pour mieux présenter vos pages !	113
Longueur : environ 200 caractères	115
Zone chaude 6 : la balise meta keywords	116
Keywords : n'y passez pas trop de temps !	117
La balise news_keywords pour Google Actualités	118
Zone chaude 7 : les attributs alt et title	120
Zone chaude 8 : le nom de domaine	120
Quelle extension choisir ?	122
L'hébergement est-il important ?	122
L'ancienneté du domaine est-elle importante ?	124
Noms composés : avec ou sans tirets ?	124
Nom de domaine : que choisir ?	125
Stratégie de référencement et nom de domaine	126
Des minisites valent mieux qu'un grand portail	127
Les sous-domaines	127
Référencement des sites multilingues	128
Zone chaude 9 : les intitulés d'URL	131

CHAPITRE 5

Optimisation – Les critères in page : contenu textuel	135
Le contenu optimisé est capital !	137
La notion de texte visible	137
La taille d'un texte	138
Faut-il souvent répéter un mot ?	139
Les différentes formes d'un mot	140
La notion de requête principale (RP)	141
La casse des lettres	142
L'ordre et l'éloignement des mots	142

Une thématique unique par page	142
Langue du texte	143
Localisation du texte	143
L'optimisation SEO d'un texte	143
Requêtes principale et secondaires.	143
Structuration en balises <h1> à <h6>	144
Mots en gras (balise).	147
Crosslinking	148
Liens externes.	152
Quelques exemples	152
Balise <title>	156
Titres multilingues	162
Insérer des codes ASCII dans le titre : bonne ou mauvaise idée ?	163
Balise meta description	164
La fraîcheur de mise à jour des informations.	167
La problématique de l'âge et de la fraîcheur	167
Quelles sont les performances des moteurs en matière de fraîcheur de l'index ?	168
Les obstacles à la détermination de l'âge d'une page ou d'un lien.	169
Quelles pages favoriser dans l'algorithme : les pages anciennes ou les pages récentes ?	170
L'analyse temporelle des liens	170
Les autres critères temporels	171
Un exemple d'analyse temporelle des flux de requêtes : les requêtes QDF .	172
L'analyse des tendances	173
La temporalité : un élément à intégrer dans le référencement	174
Plan du site et pages de contenu : deux armes pour le référencement	174

CHAPITRE 6

Optimisation – Les critères off page	177
Liens internes et réputation	178
Réputation d'une page distante.	178
Soignez les libellés de vos liens	180
À éviter le plus possible : images, JavaScript et Flash	180

Les liens sortants présents dans vos pages	182
Liens externes, PageRank et indice de popularité	183
Comment le PageRank est-il calculé ?	183
Mode de calcul du PageRank	184
Le PageRank en images	187
Le PageRank seul ne suffit pas.	188
Mise à jour du PageRank	189
Le netlinking ou comment améliorer son indice de popularité	190
Conseils d'ordre général	191
Évitez l'« échange de liens » massif	193
Des liens triangulaires plutôt que réciproques	194
Visez la qualité plutôt que la quantité	194
Utilisez la fonction « sites similaires »	196
Prenez en compte la valeur du PageRank du site distant	197
Paid linking : bonne ou mauvaise idée ?	198
Attention aux pages des sites distants et de votre site	201
Créez une charte de liens	202
Suivez vos liens	202
Privilégiez le lien naturel en soignant la qualité de votre site	203
Spamdexing ou non ?	203
Le linkbaiting ou comment attirer les liens grâce à votre contenu	205
Link ninja : de la recherche de liens classiques	209
Donnez du sens à vos liens !	209
La sculpture de PageRank	212
Le statut juridique des liens hypertextes	216
La nature et le statut du lien hypertexte	217
Le régime juridique des liens	218
Le TrustRank ou indice de confiance	221
Définition du TrustRank	221
Le TrustRank sous toutes ses formes	224
Le TrustRank en 2016	224
Les autres critères	226
 CHAPITRE 7	
Référencement multimédia, multisupport	227
Référencement des images	228

Désindexer ses images	235
L'avenir : reconnaissance de formes et de couleurs	235
Référencement des vidéos.	236
Des recherches incontournables sur les outils dédiés.	238
Différents types de moteurs de recherche	238
Comment les moteurs trouvent-ils les vidéos ?	238
L'optimisation des vidéos	239
L'optimisation de l'environnement de la vidéo.	240
Deux stratégies (plus une) de référencement de vidéos	245
Privilégier HTML5	247
Le référencement de fichiers PDF et Word	247
Prise en compte de ces fichiers par les moteurs	247
Zones reconnues par les moteurs de recherche	249
Contenu des snippets	250
Référencement sur l'actualité et sur Google Actualités	253
Comment se faire référencer sur Google Actualités ?	253
Comment assurer une indexation régulière des articles ?	256
Un Sitemap pour Google Actualités.	257
Comment apparaître sur la page d'accueil de Google Actualités ?	259
Comment faire apparaître une image ?	261
Comment mieux positionner un article dans les résultats ?	263
Contrôler l'indexation des pages.	263
Le référencement local (Google Maps)	266
Google My Business.	268
Référencement sur les mobiles	277
Concevoir un site « mobile friendly »	278
Optimiser un site mobile	279
Soumettre son site dans les moteurs mobiles	281
Aspects techniques et gestion des spiders	283
SEO et responsive design	288
AMP (Accelerated Mobile Pages)	295
Le référencement des Apps dans les Stores	299
App Indexing : un mélange entre le moteur de recherche et les Apps	302
Le référencement audio	305
Blinkx, autre technologie de recherche majeure	308
Voxalead News	309

L'avenir du référencement audio	310
L'internaute aura-t-il le dernier mot ?	313
CHAPITRE 8	
Le SMO	315
Quels réseaux sociaux utiliser pour son référencement ?	316
Twitter, Facebook et Google+ : indispensables au SEO ?	318
Corrélation ou causalité ?	318
De nouveaux critères de pertinence	321
Le SMO, concurrent du SEO ?	324
Google+, le réseau à suivre.	324
CHAPITRE 9	
Suivi du référencement	327
Le retour sur investissement : une notion essentielle	328
Différents types de calculs du retour sur investissement	331
La mise en place de liens de tracking.	333
Mesure d'audience : configuration du logiciel	333
Logiciels de suivi du ROI	334
Exemple de tableau de bord SEO sous Analytics.	334
Google Analytics et les tableaux de bord personnalisés	334
Premier tableau de bord : l'analyse du trafic	339
Sur quels leviers agir ?	345
Conclusion	349
Le « not provided », fléau du webmarketeur	349
Les outils pour webmasters fournis par les moteurs.	351
Conclusion	353
CHAPITRE 10	
Internalisation ou sous-traitance ?	355
Faut-il internaliser ou sous-traiter un référencement ?	356
Audit et formation préalable	358
Élaboration du cahier des charges	360
Définition des mots-clés	360

Mise en œuvre technique du référencement	361
Suivi du référencement	362
Coûts.	363
Préconisations.	367
Conclusion	368
Réussir l'externalisation de votre SEO	369
L'importance de l'interlocuteur unique.	373
Les points à vérifier avant de signer	374
Quelles garanties un référenceur peut-il proposer ?	380
Combien coûte un référencement ?	380
Un référencement gratuit est-il intéressant ?	381
Où trouver une liste de prestataires de référencement ?	381
Chartes de déontologie	382
Charte de déontologie du métier de référenceur.	382
 PARTIE C	 385
 Devenir un as du SEO...	 385
 CHAPITRE 11	
Comment obtenir plus de visibilité dans les résultats des moteurs ?	387
Authorship, Author Rank ou la confiance apportée aux auteurs de contenus	388
Les rich snippets : l'avenir des balises meta ?	391
Une implémentation simple	393
Schema.org, un nouveau standard de rich snippets	403
Un exemple d'intégration des rich snippets et de Schema.org	410
Le Knowledge Graph, de la sémantique dans les SERP	415
Knowledge Graph et SEO	418
Les SiteLinks (liens de sites) de Google	425
 CHAPITRE 12	
Optimisation de l'indexation	429
Le formulaire de soumission proposé par le moteur	430

Le lien depuis une page populaire	433
Les fichiers Sitemaps	434
Le concept des Sitemaps	435
Formats du fichier à fournir à l'applicatif.	436
Format des fichiers Sitemaps	436
Exemples de fichiers.	439
Travail sur plusieurs fichiers.	440
Cas particulier des sous-domaines.	441
Différents types de Sitemaps	446
La prise en compte par d'autres robots que ceux crawlant le Web. ...	447
Le référencement payant (paid inclusion, trusted feed)	447
L'indexation en temps réel : PubSubHubbub et Ping	448
Les avantages de PubSubHubbub.	449
Optimisez le temps d'indexation d'un nouveau site	450
Mettez en ligne une version provisoire du site	450
Profitez de cette version provisoire	452
Proposez du contenu dès le départ	453
Faites des mises à jour fréquentes de la version provisoire	453
Générez les premiers liens	454
Inscrivez votre site sur certains annuaires dès sa sortie.	454
Créez des liens le plus vite possible	455
Présentez votre site sur les forums et blogs	455
 CHAPITRE 13	
Index secondaire et duplicate content	457
Google : index principal et secondaire	458
Les deux index cohabitent pourtant encore... ..	459
Comment vérifier dans quel index sont vos pages ?	461
Conclusion sur les index de Google	463
Duplicate content : un mal récurrent....	463
Problème 1 – Contenu dupliqué sur des sites partenaires	466
Problème 2 – Contenu dupliqué sur des sites « pirates »	470
Problème 3 – Duplicate content intrasite	471
Problème 4 – Duplicate content par similarité de balises	474
Problème 5 – DUST : même code source accessible <i>via</i> des URL différentes	477

Duplicate content : l'évangile selon saint Google.....	484
CHAPITRE 14	
Freins au référencement et solutions possibles.....	487
Site 100 % Flash	488
Des « rustines » pour mieux indexer le Flash ?	490
Langages JavaScript, Ajax et Web 2.0.....	495
Comment faire du JavaScript « spider compatible » ?	496
Créer des menus autrement qu'en JavaScript	497
La problématique des sites en Ajax ou de style Web 2.0	502
Menus déroulants et formulaires	508
Sites dynamiques et URL « exotiques ».....	509
Format d'une URL de site dynamique	510
Pourquoi les moteurs de recherche n'indexent-ils pas, ou mal, les sites dynamiques ?	511
Quels formats sont rédhibitoires ?	512
Le cloaking	513
Création de pages de contenu	514
L'URL Rewriting	514
Balises multilingues et multipays	521
Le problème de la détection de la langue sur les pages multilingues.....	522
La solution théorique :	
déclarer la langue du contenu dans le code de la page.....	522
Mais on ne peut pas faire confiance aux webmasters !	523
Le problème des variantes locales	524
Des quasi-doublons dus aux différentes versions linguistiques	524
L'annotation hreflang à la rescousse	525
L'indication par défaut : x-default-hreflang	527
Conclusion : dans quels cas utiliser ces annotations hreflang ?	527
Identifiants de session	528
Cookies	528
Accès par mot de passe	529
Tests en entrée de site	529
Redirections	530
Hébergement sécurisé	533

L'https comme critère de pertinence ?	534
Qu'est-ce que le protocole SSL/TLS ?	535
Pourquoi Google veut-il utiliser ce critère dans son algorithme de classement ?	536
Et qui risque d'être gêné par ce changement ?	536
Quelle est l'importance du bonus accordé par Google aux sites sécurisés via SSL/TLS ?	537
Quels sont les avantages et les inconvénients d'un site en https:// ?	537
L'importance du bon choix du type de clé et du bon fournisseur de certificat	539
Quel fournisseur de certificat choisir ?	539
Comment basculer un site en https:// ?	540
Quel sont les risques liés à la bascule entre les URL en http:// et https:// ?	541
Dans quels délais le « bonus https:// » deviendra-t-il significatif ?	542
Google vous pousse à le faire : préparez-vous	542
Les widgets pour créer des liens	543
Widgets et popularité	544
Matt Cutts et les widgets	545
Informez les internautes	546
Évitez le spam dans les widgets	546
Privilégiez les liens éditoriaux	546
Compatibilité W3C : un réel impact ?	548
Le W3C : pour ou contre	551
Temps de chargement des pages, temps de réaction du serveur	552
Quels problèmes pour quelles solutions ?	553
Les outils de test	553
Compressez pour diminuer le nombre de Ko téléchargés	556
Activez le cache du navigateur	556
Activez le préchargement des pages	557
Utilisez des serveurs tierce partie	557
Installez le code Google Analytics asynchrone	558
Les feuilles de styles CSS	558
Les Sprites CSS pour optimiser le chargement d'images	560
Les fichiers JavaScript	562
Optimisation des images	562
Boostez votre référencement en boostant vos pages web !	563
Les frames	564

Optimisation de la page mère	567
Optimisation des pages filles	569
Conclusion	570

CHAPITRE 15

Spam et pénalités, Panda et Penguin	571
Quelques pistes de réflexion au sujet des pénalités	573
La délation	573
Les pénalités infligées par Google	578
Techniques à ne pas employer	579
Pénalité numéro 1 – Le mythe de la sandbox	580
Pénalité numéro 2 – Le déclassement	581
Pénalité numéro 3 – La baisse de PageRank dans la Google Toolbar	582
Pénalité numéro 4 – La liste noire	582
Que faire si vous êtes pénalisé ?	583
Les filtres de nettoyage : Panda, Penguin, EMD...	592
Rappel du fonctionnement des filtres de nettoyage de Google	592
Google Panda	593
Google Penguin	616
Google EMD et Page Layout	627
Les Quality Updates Phantom	629
Conclusion	634

CHAPITRE 16

Comment ne pas être référencé ?	635
Pourquoi déréférencer un contenu ?	636
Les risques de la désindexation	637
Fichier robots.txt	638
Balise meta robots	641
Directive X-Robots-Tag	642
Quel type de désindexation utiliser ?	644
Fonctions spécifiques de Google	646
Balise meta robots spécifique	646
Suppression des extraits textuels (snippet)	646
Suppression des extraits issus de l'Open Directory	647

Suppression de contenu inutile	647
Suppression des pages en cache	648
Suppression d'images	649
Conclusion	651
La règle des « 4C » : Contenu, Code, Conception et Célébrité	652
Contenu éditorial : tout part de là !	653
Code HTML : les grands classiques	655
Conception : indexabilité sans faille	658
Célébrité : popularité, réputation et confiance	662
Conclusion	662
Les 12 phrases clés du référencement	663
 ANNEXE	
Webographie	665
La trousse à outils du référenceur	666
Add-ons pour Firefox	666
Audit de liens	667
Analyse du header http	667
Sites web d'audit SEO	667
Positionnement	668
Les musts de la recherche d'informations et du référencement	669
En français	669
En anglais	669
Blogs officiels des moteurs de recherche	670
Les forums de la recherche d'informations et du référencement	670
Forums en français sur les outils de recherche et le référencement	670
Forums en anglais sur les outils de recherche et le référencement	670
Les associations de référenceurs	670
Les baromètres du référencement	671
Baromètres français	671
Baromètres anglophones	671
Lexiques sur les moteurs de recherche et le référencement	671
Index	673

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 ?

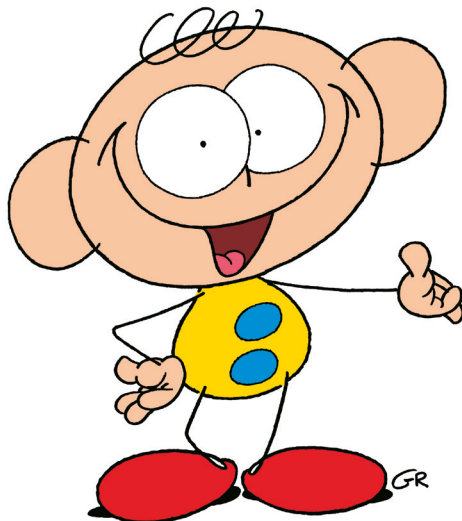
– Bien sûr, dit le renard. Tu n'es encore pour moi qu'un petit garçon tout semblable à cent mille petits garçons. Et je n'ai pas besoin de toi. Et tu n'as pas besoin de moi non plus. Je ne suis pour toi qu'un renard semblable à cent mille renards. Mais, si tu m'apprivoises, nous aurons besoin l'un de 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 du *Petit Prince* d'Antoine de Saint-Exupéry (1943)



Ce livre vous apportera de nombreuses et précieuses informations pour mieux optimiser, de façon « loyale » et honnête, les pages de votre site afin d'acquérir une meilleure visibilité dans les résultats des moteurs de recherche en général et, bien sûr, de Google en particulier. Au fil des années, cet ouvrage a en effet comblé un vide assez important dans ce domaine. En l'an 2000 paraissait la dernière version de mon livre *Créer du trafic sur son site web*, aux éditions Eyrolles. Un autre vendu directement sur le Web au format PDF et intitulé *Référencement 2.0* lui a succédé. Conscient qu'il fallait actualiser les informations que ces livres proposaient, j'avais entrepris l'écriture de *Réussir son référencement web*, dont la première version parut en janvier 2008, la deuxième fin 2009, la troisième fin 2010, la quatrième en janvier 2012, la cinquième en avril 2013, la sixième (pour 2014) en décembre de cette même année et la septième en février 2015. Ces différentes éditions ont connu, selon les dires de l'éditeur, un grand succès. Tant mieux, cela signifie que la demande est forte sur ce créneau. En 2011, il avait été décidé de proposer chaque année une nouvelle mouture de cette édition. Vous avez donc entre les mains le dernier-né pour 2016 (il paraîtrait que certains en font la collection et ont une étagère bien garnie :-)).

L'édition 2014 fêtait également un anniversaire, puisque cela faisait vingt ans que je travaillais (et que je travaille encore) dans le domaine du référencement. J'ai découvert le Web en 1993 au travers d'une démonstration de Mosaic 1.0 (l'ancêtre des actuels navigateurs Chrome ou Firefox) à l'université Louis Pasteur de Strasbourg. Sortant de plusieurs années de Minitel, je me suis immédiatement pris de passion pour Internet et je suis, comme on dit en dialecte « astérixien », « tombé dans la marmite ». Les moteurs de recherche (et les annuaires, à cette époque) ont tout de suite attiré mon attention et la visibilité d'un site web sur ces outils a très tôt été au centre de mes préoccupations. J'ai alors créé ma société, Abondance (car, en 1996, il fallait être bien placé par ordre alphabétique dans les annuaires !), avec une offre de référencement sur Yahoo! et AltaVista. Une autre époque !

Cela fait donc un peu plus de deux décennies que je navigue dans le microcosme parfois tortueux du référencement naturel (qui s'appelle de plus en plus SEO pour *Search Engine Optimization*) et que j'essaie de m'adapter au quotidien à de nouvelles stratégies que nous proposons (ou nous imposent) l'actualité et bien sûr Google depuis plus de quinze ans.

Mon activité de consultant m'a amené à auditer de très nombreux sites, à rencontrer des problématiques parfois complexes, à examiner et analyser les stratégies de mes clients. Cette expérience, j'ai essayé de la résumer dans ce livre.

J'espère sincèrement que les versions précédentes de cet ouvrage ont 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 leur parution et des témoignages qui me sont rapportés lors de mes conférences et formations, il semblerait que oui et je ne peux que remercier ici toutes les personnes qui m'ont fait part de leur expérience depuis.

Un site web dédié au livre

Vous pouvez obtenir plus d'informations au sujet de cet ouvrage (et des autres que j'ai « commis » également) sur le site qui lui est dédié :

- <http://www.livre-referencement.com>.

Et notamment les témoignages des lecteurs des précédentes éditions :

- <http://www.livre-referencement.com/temoignages.html>.

Comme je l'ai dit, il s'agit de la huitième édition de cet ouvrage. Au fil des années, il n'a eu de cesse de s'enrichir d'informations supplémentaires. Mais cela avait également généré, de façon presque obligatoire, des redites dans certains chapitres. L'édition 2013 avait en partie corrigé cela. Mais il fallait, à un moment donné, revoir en profondeur le contenu de l'ouvrage et c'est ce que j'avais décidé de faire pour l'édition 2014, qui a été remaniée pour proposer des informations précises, mises à jour et correspondant à une vision actuelle du SEO. La version que vous avez entre les mains, qui est celle pour 2016, n'est qu'une mise à jour de la mouture totalement remaniée deux ans auparavant et déjà actualisée en 2015.

Il existe aujourd'hui de nombreux ouvrages sur le webmarketing, 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.

Le contenu de ce livre est donc centré sur le référencement naturel et ses notions connexes : positionnement, optimisation de site, analyse de l'efficacité d'une stratégie de référencement, etc. Et Dieu sait s'il y a des choses à dire ! Vous vous en apercevrez dans les pages qui suivent...

L'articulation de cette édition 2016 est divisée en trois grandes parties.

- La première (*Présentation du SEO, définitions et généralités*) a pour but d'expliquer ce qu'est le domaine du webmarketing et de vous présenter en détail les différentes stratégies possibles, le mode de fonctionnement des moteurs de recherche, etc.
- La deuxième (*Les fondamentaux du SEO*) présente la partie technique (les différentes balises à mettre en œuvre), la rédaction (écrire pour les internautes en pensant aux moteurs) et le *netlinking* (les liens entrants sur un site, si importants aux yeux de Google). D'autres domaines comme le SMO (réseaux sociaux), le référencement des images, des vidéos, etc., sont également abordés. Le suivi et la mesure d'un bon référencement, phases essentielles d'une stratégie SEO, ainsi que la question de l'internalisation ou de la sous-traitance des actions à mettre en place sont traités. En un mot, tout ce qu'il faut savoir sur le sujet si vous n'êtes pas un spécialiste de ce domaine.
- La troisième partie (*Pour aller plus loin*) aborde des sujets plus techniques, parfois plus complexes, comme le *duplicate content*, les pénalités, les filtres Panda et Penguin de Google, la réécriture d'URL, la désindexation, etc. En d'autres termes, de nombreux points, plus « avancés », que vous pourrez étudier une fois que vous vous serez familiarisé avec les chapitres précédents.

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 (et si le message est reproduit sur le site web du livre, vous y gagnerez un lien au passage !).

Pour clore cet avant-propos, il me semble important d'indiquer une chose : le référencement n'est pas une science exacte. Vous pourrez être d'accord ou pas avec certaines affirmations proposées par cet ouvrage. C'est normal. Nous œuvrons tous dans un domaine empirique où seuls les tests et l'expérience nous font progresser. Le dialogue, 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 qu'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 toujours le caractère principal de certains de ses acteurs).

Enfin, rappelons qu'à la demande de nombreux lecteurs, les URL sont pour la plupart citées dans cet ouvrage grâce au raccourcisseur d'adresses (à tout seigneur tout honneur) goo.gl. Elles sont ainsi beaucoup plus faciles à saisir à la main lorsqu'elles sont très longues ! Ceci dit, lisez-les et déchiffrez-les attentivement : un « I » (i majuscule) est parfois proche d'un « 1 » ou d'un « l » (L minuscule), un « 0 » (zéro) peut être confondu avec un « O » (lettre majuscule), etc. Toutes les adresses raccourcies ont en tout cas été (re) vérifiées pour cette nouvelle édition. Êtes-vous certain de les avoir bien saisies en cas de problème ?

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

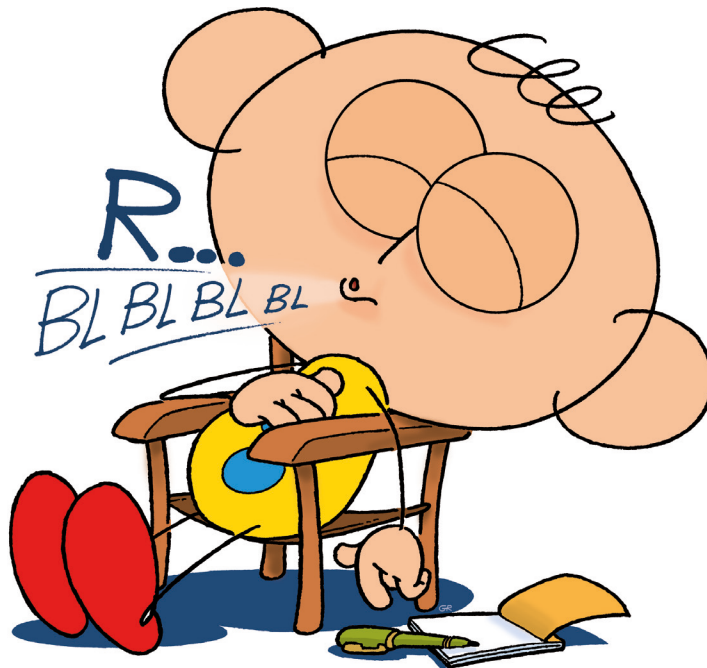
Partie A

Les bases du référencement naturel (SEO)



1

Le référencement aujourd'hui : généralités



« Je ne sais qu'une chose, c'est que je ne sais rien. »

Socrate

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 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.

Ainsi, vous trouverez dans ce chapitre 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... vous connaissez la suite).

Commençons avec le terme de « référencement » et 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.

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, en tête de gondole, ou encore à proximité des caisses de paiement, ce qui les rend plus « visibles ». Certains produits sont alors mis en avant à des endroits stratégiques, beaucoup plus facilement repérables par les clients potentiels. Ils sont ainsi bien « positionnés »... Vous voyez où nous voulons en venir ?

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 (qu'on appelle « index ») d'un moteur, on dira qu'il est « référencé » ou « indexé ». 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. De même, un produit sans intérêt pour le consommateur ou obsolète (date de conservation dépassée) pourra être retiré des rayons. Il sera donc « déréféré » ou « désindexé ». Là encore, il en sera de même pour votre site : vous ne devrez montrer à Google que les pages les plus intéressantes, celles qui ont le meilleur potentiel pour l'internaute. Nous y reviendrons...



Figure 1-1

Dans les grandes surfaces, les produits sont également « référencés »...

Source : D.R.

Grande surface et galerie marchande

Pour poursuivre l'analogie précédente, on peut estimer que la grande surface représente les résultats « naturels » du moteur de recherche, alors qu'une galerie marchande propose 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, comme dans la vie réelle.

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

Enfin, une troisième étape, malheureusement souvent négligée, sera nécessaire pour vérifier *in fine* que le positionnement a porté ses fruits, en évaluant le trafic engendré par vos efforts d'optimisation. Croyez-vous que les responsables de supermarchés ne vérifient pas si leurs produits se vendent mieux – ou pas – 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 le rayon 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... Nous reviendrons bien sûr amplement sur ces notions dans les chapitres suivants.

Le référencement, une réalité aux sens multiples

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 engendré, etc. alors qu'en réalité, il s'agit uniquement de définir par ce terme le processus d'indexation des pages d'un site web par le moteur de recherche. Et donc la simple connaissance de ces pages par Google et consorts. 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.

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 ou financier 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. On les appelle également parfois (notamment chez Microsoft/Bing) les « liens bleus ».

SEM = SEO + SEA

En termes de stratégie de visibilité sur les moteurs de recherche, on parle la plupart du temps de SEM (*Search Engine Marketing*). Ce terme regroupe le SEO (*Search Engine Optimization*), ou référencement naturel, et le SEA (*Search Engine Advertising*), ou gestion des liens publicitaires, souvent appelé improprement « référencement payant » alors que le SEA n'a rien à voir avec une quelconque stratégie de référencement !

La figure 1-2 illustre notre propos 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 facturation au clic. Les résultats sont affichés selon une formule complexe appelée *Quality Score*. Si ces zones ne font pas spécifiquement partie intégrante d'une stratégie de référencement dit « naturel », « organique » ou « traditionnel » (SEO), le système du lien sponsorisé (baptisé SEA pour *Search Engine Advertising*) en est complémentaire. Nous ne l'aborderons cependant pas – ou très peu – dans cet ouvrage.

- La zone 3 représente ce qu'on appelle les liens organiques ou naturels, qui sont fournis par l'algorithme mathématique de pertinence (la formule de classement) du moteur de recherche. Ils n'ont rien à voir avec la publicité affichée dans les zones situées au-dessus. C'est donc dans cet espace que nous allons essayer de vous aider à positionner vos pages.

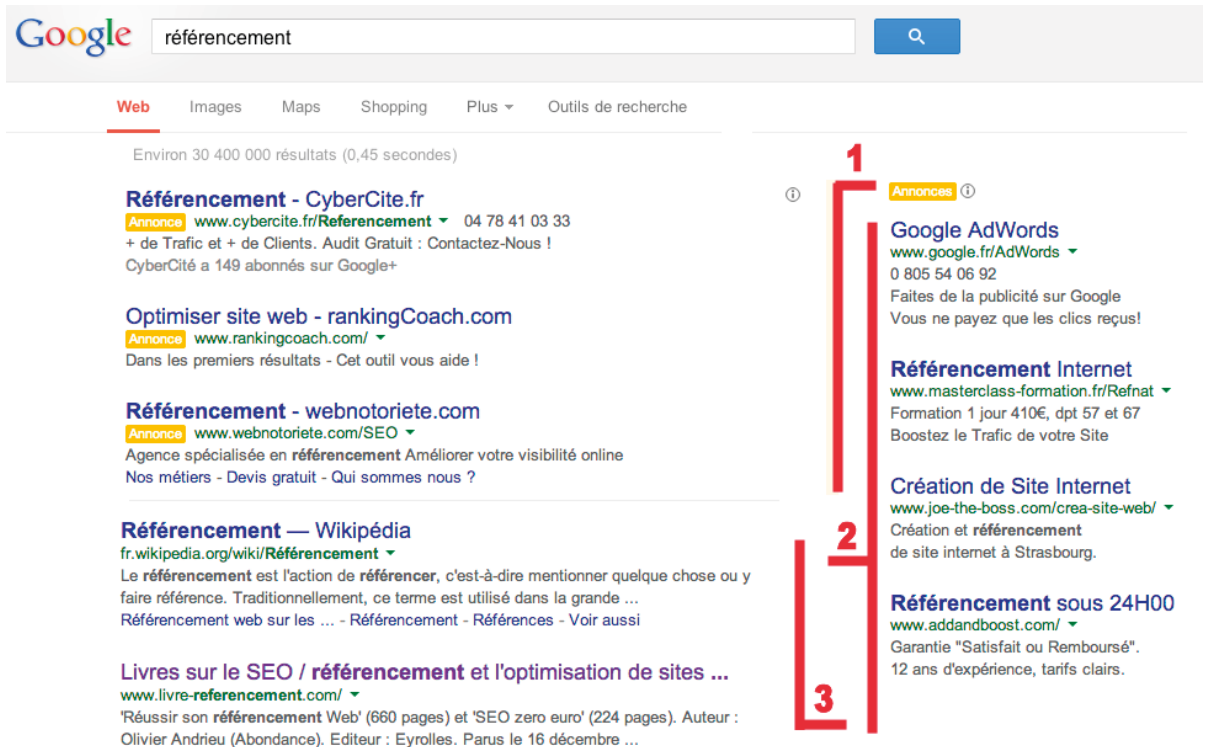


Figure 1-2

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

Tous les moteurs de recherche majeurs proposent sur leurs pages de résultats cette dualité liens sponsorisés/liens organiques.

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 l'une de l'autre. Être un gros annonceur sur Google ou Bing n'influe donc pas 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 outil de recherche est de présenter des résultats objectifs et indépendants des budgets publicitaires. Que de « serpents de mer » n'ont-ils pas été imaginés à ce sujet depuis de nombreuses années...

Ceci dit, si par défaut liens naturels et sponsorisés sont indépendants, cela ne signifie pas qu'ils n'ont pas d'influence l'un sur l'autre. Le fait d'arrêter une importante campagne de liens sponsorisés, par exemple, influera sur le trafic engendré sur le site, et donc, la notion de trafic étant aujourd'hui prise en compte par Google dans son algorithme de pertinence – contrairement au passé –, les positionnements en liens naturels pourront en pâtir. On a vu de nombreuses situations de ce genre se produire sur le moteur Google : l'arrêt d'importantes campagnes AdWords peut influencer négativement les positions d'un site en SEO, qu'il regagnera lorsque de nouvelles campagnes seront mises en œuvre. Ceci est logique et représente l'impact indirect d'un gain ou d'une perte de trafic sur un site. Mais une mention du site dans une émission de télévision très suivie, par exemple, et générant un pic de trafic pourrait avoir le même effet. En revanche, l'impact d'une campagne publicitaire ne sera décelable que si les budgets mis en œuvre sont importants. Quelques dizaines d'euros d'AdWords par jour ne devraient logiquement générer aucune fluctuation en termes de SEO.

Les figures 1-3 et 1-4 illustrent des pages de résultats de recherche sur les moteurs de recherche Yahoo! et Bing, principaux concurrents de Google à l'heure actuelle. Comme on peut le voir, les résultats sont également répartis en trois zones distinctes : les liens sponsorisés figurent dans les zones 1 et 2 et liens organiques dans la zone 3. Il en est ainsi sur la majeure partie des moteurs de recherche actuels.

The screenshot shows the Yahoo! France search interface. The search bar contains 'référencement' and the 'Rechercher' button is highlighted. Below the search bar, there are tabs for 'WEB', 'IMAGES', 'VIDÉO', 'ACTUALITÉS', 'SHOPPING', and 'PLUS'. The search results are displayed in three main sections:

- Zone 1 (Sponsored Results):** This section is highlighted with a red '1' and contains several sponsored links. The first link is 'Devis Référencement' from Visiplus.com, followed by 'Référer Son Site' from LogicielReferencement.com, 'Une meilleure audience' from BingAds.Microsoft.com, and 'référencement' from Netissime.com. A red vertical line separates this section from the organic results.
- Zone 2 (Sponsored Results):** This section is highlighted with a red '2' and contains more sponsored links. The first link is 'Référencement Pro' from addandboost.com, followed by 'Référencement Pro Gratuit' from seo-gratuit.page-internet.net, 'PUB gratuite en ligne' from pmtch.com, and 'Devis en Référencement' from Companeo.com/Devis/Referencement.
- Zone 3 (Organic Results):** This section is highlighted with a red '3' and contains organic search results. The first result is 'Référencement - Wikipédia' from fr.wikipedia.org/wiki/Référencement, followed by 'Référencement' from Kelformation.com/Alt-web.

On the left side of the page, there are links for 'RECHERCHES ASSOCIÉES' and 'FILTRER PAR DATE'.

Figure 1-3

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

WEB IMAGES VIDÉOS CARTES ACTUALITÉS PLUS

bing référencement

11 100 000 RÉSULTAT(S) Affiner par langue ▼ Affiner par pays ▼

Devis Référencement Annonces
 Visiplus.com/Devis_Référencement
 100 % Made In France ! A partir de 490 € / mois.

Référencer Son Site
 LogicielReferencement.com
 Référencement de Site en 1ère Page! Plus de 2800 références en 2012

Élargissez votre audience
 BingAds.Microsoft.com
 Stimulez et proposez votre activité sur le Yahoo! Bing® Search Network

referencement
 www.Netissime.com/Domaine
 Offre exceptionnelle : Votre .FR pour seulement 3,99 € HT l'année !

Référencement naturel sur les moteurs de recherche
 www.referenceement.com ▼
 Des campagnes de **référencement** naturel pour optimiser le positionnement de votre site web. Nos offres de **référencement** naturel vont aider votre site eCommerce ...

Référencement - Wikipédia
 fr.wikipedia.org/wiki/Référencement ▼
 Référencement web ... · Notes et références · Voir aussi
 Le **référencement** est l'action de référencer, c'est-à-dire mentionner quelque chose ou y faire référence. Traditionnellement, ce terme est utilisé dans la ...

Devis en Référencement Annonces
 Companeo.com/Devis/Referencement
 Obtenez 3 devis en 48h en **Référencement** de Sites Internet.

Référencement à 29 €
 www.addandboost.com
 +10.000 sites déjà référencés par nos services. Satisfait ou remboursé

Référencement Pro Gratuit
 seo-gratuit.page-internet.net
 Votre site dans plus de 200 annuaires gratuitement!

referencement
 ZapMeta.fr/referenceement
 Cherchez **referencement** Regardez **referencement**

PUB gratuite en ligne
 www.pmtch.com
 Référencement pour votre site seulement en quelques clics
 Voir votre annonce ici

Figure 1-4

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

La différence se fera, sur un moteur ou l'autre, au niveau de la clarté de différenciation des différentes zones. Certains outils indiquent sans ambiguïté ce qui est de la publicité et ce qui ne l'est pas, alors que d'autres y sont moins enclins, et ce, essentiellement pour des raisons de profits. En effet, l'internaute, qui pense avoir affaire à un lien organique, va en fait cliquer sur une publicité... Reste à estimer le bienfait pour l'internaute de ce type de pratique. Notons qu'en 2013, le fond pastel utilisé pour les liens sponsorisés en position « premium » (au-dessus des résultats naturels) s'est peu à peu éclairci sur les principaux moteurs (Google, Yahoo! et Bing), et a complètement disparu sur Google en 2014 au « profit » d'un encadré intitulé « Annonces », en blanc sur fond orange. Des méthodes qui peuvent tromper l'internaute et qu'il faut surveiller de près.

Attention aux sociétés peu scrupuleuses !

Parfois, certaines sociétés vous proposeront, sous l'appellation de « référencement naturel » ou « référencement », l'achat de mots-clés en liens sponsorisés sans le mentionner expressément. Dans ce cadre, des 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) !

Une stratégie de référencement 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.

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 (mots-clés dits « transactionnels »). 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, Bing...) 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, comme nous le disions précédemment. Le fournisseur d'accès Internet Free a, par exemple, lancé en 2007 une page de résultats associant liens naturels et liens sponsorisés sans qu'il soit facilement possible de les distinguer visuellement (<http://goo.gl/kh71Q>). Ce type de stratégie peut tuer le marché de la publicité sponsorisée si elle est appliquée en masse et n'est donc clairement pas à encourager. Free est d'ailleurs revenu en arrière quelques mois plus tard, conscient de son erreur stratégique. À surveiller...

Pour résumer, 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 (sic) 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.

L'affichage des résultats par un moteur se décompose donc en trois étapes (figure 1-5).

1. Mise en place par le moteur d'un « index », ou copie du Web à un instant T sur ses disques durs, regroupant des milliards de pages web dans lesquelles il effectuera ses recherches. Cet index devra, bien sûr, être le plus « frais » possible afin de fournir à l'internaute la réponse la plus pertinente à ses questions.
2. Extraction, depuis l'index, des pages répondant à la requête saisie par l'utilisateur.
3. Calcul et classement des résultats par pertinence.
4. Affichage des résultats.

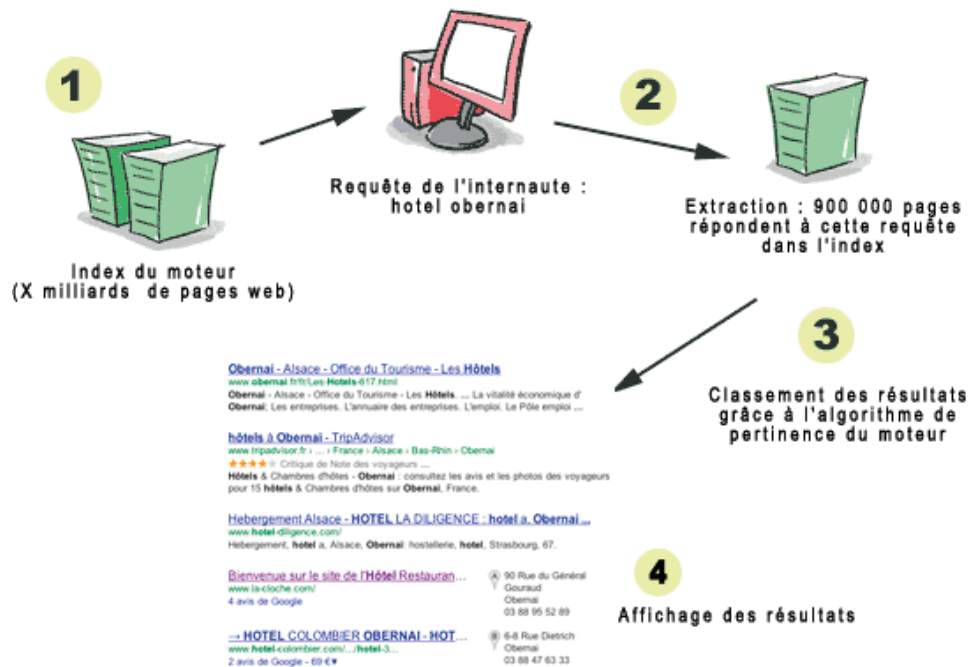


Figure 1-5

Les quatre étapes essentielles d'un processus de référencement
Source des dessins : Google

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 ; les pages de votre site web devront donc y être présentes. Il s'agit de la phase de « référencement » proprement dite. Si votre site propose 100 ou 1 000 pages web, il faudra idéalement qu'elles figurent toutes dans l'index du moteur, même si le filtre Panda de Google a modifié cette approche, comme nous le verrons dans cet ouvrage. 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...

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 dans la partie consacrée au 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. Pour résumer, si vous souhaitez obtenir une bonne visibilité sur l'expression « restaurant caen », 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 l'expression « restaurant caen », Google renvoie plus de 11 millions de résultats... La concurrence est féroce. 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 podium, voir ci-après), il faudra insérer ces termes de recherche dans des « zones chaudes » de vos pages : titre, texte, URL, etc. Et bien sûr, il faudra que votre site ait été conçu pour que ces « zones chaudes » soient bien présentes dans vos pages. Nous étudierons tout cela très bientôt.

Les phases essentielles du référencement

Un processus de référencement s'effectue en trois phases essentielles...

1. Le référencement : votre site doit être « trouvable » (« en rayon ») dans l'index du moteur, de la façon la plus complète possible et pour ses pages les plus pertinentes
2. L'identification : 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. Le 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 les « zones chaudes » des pages.

Et son efficacité doit être évaluée lors d'une quatrième étape.

4. Le suivi : comment estimer et mesurer l'efficacité d'une stratégie de référencement ?

Positionnement, oui, mais où ?

Dans les premières années du référencement, on avait tendance à dire que, après saisie d'un mot-clé, le but d'un bon positionnement était d'apparaître dans les trois premières pages de résultats des outils de recherche, soit entre la 1^{re} et la 30^e position. Il s'agissait effectivement d'une sorte de « contrat » essentiel qu'il ne fallait jamais dépasser à cette époque-là. Être classé après la 30^e position sur un mot-clé donnait équivalait à un trafic quasi nul. En effet, très peu d'internautes dépassaient cette fatidique troisième page de résultats lors de leurs recherches. Au-delà, donc, point de salut.

Aujourd'hui, il n'est plus du tout satisfaisant de se situer dans les trente premiers résultats compte tenu du changement des habitudes des utilisateurs de moteurs de recherche. Auparavant, les internautes saisissaient une requête, puis consultaient les pages de résultats consécutives du moteur s'ils ne trouvaient pas la réponse rapidement (page 1, puis 2, puis 3, etc.). Actuellement, l'internaute saisit une première requête puis, si les dix premiers liens renvoyés ne lui conviennent pas, il modifiera plutôt sa requête initiale : remplacement des mots-clés, ajout de nouveaux termes, etc. Ainsi, l'internaute reste toujours sur la première page de résultats (ou SERP : *Search Engine Result Page*) et ne dépasse jamais les dix premiers liens.

Comment les internautes utilisent-ils les moteurs ?

Les informations qui suivent sont issues d'études menées sur le comportement des internautes sur les moteurs de recherche.

- Pour effectuer des recherches sur le Web, 54 % des personnes interrogées passent toujours un moteur de recherche (contre 50 % en 2011), 32 % utilisent en revanche les réseaux sociaux pour trouver de nouveaux contenus en ligne, contre 18 % en 2010 et 25 % en 2011. La découverte des contenus par les liens passe donc en 3^e place avec 28 % de citations (31 % en 2011).
Source : *Forrester Research* (juin 2013), <http://goo.gl/N4wuDI>.
- Lorsque les internautes américains se posent des questions et recherchent des informations sur la santé, 77 % d'entre eux commencent par un moteur de recherche tel que Google.
Source : *Pew Internet* (janvier 2013), <http://goo.gl/H8WMJP>.
- Le taux de clics dans la première page de résultats naturels des moteurs de recherche est passé de 49 à 53 % depuis 2006. Source : *Ian Howells* (septembre 2011), <http://goo.gl/eYfg0>.
- 56 % des clics sont récoltés par le premier résultat, 15 % par le deuxième et 9 % par le troisième, soit 80 % des clics pour les trois premiers liens ! Source : *Compete* (octobre 2012), <http://goo.gl/QPyYd>.
- La taille moyenne d'une requête oscille entre 4,07 et 4,87 mots.
Source : *Chitika* (janvier 2012), <http://goo.gl/VVFd0>.
- 79 % des internautes interrogés par Yahoo! dans le cadre de la première édition de la Yahoo! Search Academy utilisent plusieurs mots-clés dans la majorité de leurs requêtes. Les internautes sont par ailleurs pressés puisque 61 % d'entre eux ne vont pas plus loin que la 1^{re} page de résultats. Source : *Yahoo! Search Academy : comment les internautes utilisent-ils les moteurs de recherche ?* (mars 2010), <http://goo.gl/PjANR>.
- Moins d'un tiers des internautes ne consultent que la 1^{re} page de résultats et ils sont 45 % à consulter la 2^e et la 3^e page ; ils sont moins d'un tiers à dépasser la 4^e page de résultats.
Source : *Journal du Net : comment les internautes utilisent les moteurs de recherche* (mai 2009), <http://goo.gl/tcTrq>.
- 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 3^e page de résultats. Source : *iProspect: Search Engine User Behavior Study* (avril 2006), <http://goo.gl/RJ4yp>.
- 54 % des internautes ne visualisent que la 1^{re} page de résultats (19 % vont jusqu'à la 2^e et moins de 10 % la 3^e). 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>.

- Dans le domaine de la santé, 77 % des recherches commencent sur un moteur web.
Résumé : <http://goo.gl/snddZi>.

- Le premier lien des SERP représente 32,5 % des clics. <http://goo.gl/7sMfrk>.

Consultez également d'autres études parues début 2011, disponibles à l'adresse suivante :
<http://goo.gl/FUeNH>.

Vous devez donc aujourd'hui être plus exigeant et chercher à apparaître au minimum dans les dix premiers liens affichés, soit dans la première page de résultats. C'est évidemment plus difficile selon les mots-clés choisis. Néanmoins, il est clair que le trafic engendré sera au niveau de votre ambition si vous y arrivez. En 2016, il s'agit de l'objectif de départ que vous vous fixerez. Un adage récent dit d'ailleurs : « Si vous voulez cacher un cadavre, mettez-le en page 2 des résultats de Google, personne ne le trouvera jamais ». Une phrase qui a le mérite d'être on ne peut plus claire !

Il est encore plus difficile – mais bien plus efficace – 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 (assez courante à l'heure actuelle), une page de résultats de Google pour le mot-clé « moteur de recherche » apparaît comme sur la figure 1-6.

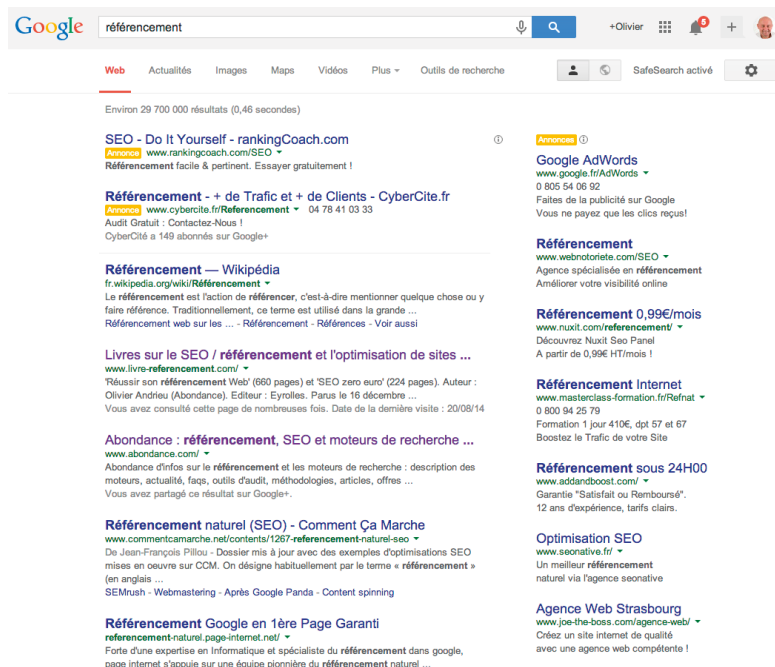


Figure 1-6

Page de résultats de Google en résolution 1 024 × 768

Le navigateur affiche ainsi deux liens sponsorisés (intitulés « Annonces ») au début de la SERP, mais surtout cinq liens organiques (issus de l'index de pages web) du moteur en dessous.

Le pari sera alors, pour que la situation soit encore meilleure, que vous apparaissiez dans ces cinq premiers liens. Attention cependant : le nombre de liens naturels affichés peut grandement varier en fonction de plusieurs facteurs...

- Le type d'informations connexes affichées par le moteur selon la requête (concept de « recherche universelle » : plans (*maps*) ou dépêches d'actualité sur Google, exemples d'images ou de vidéos 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).
- D'une éventuelle proposition de correction orthographique (qui prend plus d'une ligne sur Google).
- D'une « onebox » spécifique proposant des résultats dédiés à une requête particulière (bourse, météo, etc.).

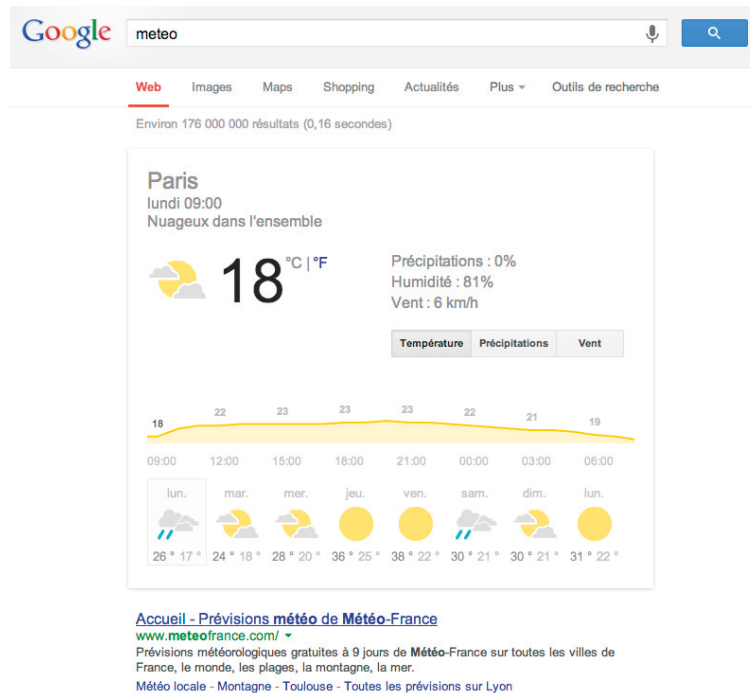


Figure 1-7

Exemple d'une onebox proposée par Google pour la requête « météo » : une zone spécifique apparaît en dessous du formulaire de recherche, proposant des résultats (basés sur la géolocalisation de l'internaute) directement fournis par le moteur de recherche. Les liens naturels sont bien plus bas et beaucoup moins visibles (ici, le site de Météo France).

Google hotel strasbourg

Web Images Maps Shopping Plus Outils de recherche

Environ 13 200 000 résultats (0,25 secondes)

Annonces relatives à **hotel strasbourg**

125 Hôtels à Strasbourg - Meilleur tarif garanti - booking.com
www.booking.com/Strasbourg-Hotels
 ★★★★★ 9 138 avis pour booking.com
 Réservez votre hôtel en ligne.
 +1 de 247 personnes de Paris pour ce contenu

Hôtels Appréciés
 Hôtels Bien Classés
 Hôtels Moins Cher
 Hôtels de Luxe

Hôtels Strasbourg dès 48€ - Accorhotels.com
www.accorhotels.com/Strasbourg
 ★★★★★ 65 avis pour accorhotels.com
 Réservez en toute sécurité dans un de nos 27 Hôtels à Strasbourg !
 Hôtels Économiques - Proche de la Cathédrale

Hôtels à Strasbourg sur Google Lien commercial ⓘ
 Comparez les hôtels en fonction des avis, des prix, des photos, de Street View, etc.

11 août 2013		12 août 2013	
Hôtels à partir de	2 étoiles	3 étoiles	4 étoiles
26 €	35 €	49 €	58 €

Prix par nuit Autres résultats de Google Recherche d'hôtels »

Hôtel Strasbourg : 128 hôtels avec 13 263 avis - TripAdvisor
www.tripadvisor.fr > Europe > France > Alsace > Bas-Rhin > Strasbourg
 ★★★★★ Note : 4,3 - Critique de Note des voyageurs ...
 Hôtels à Strasbourg : consultez 13 263 avis de voyageurs, photos, les meilleures offres et comparez les prix pour 128 hôtels à Strasbourg sur TripAdvisor.

Plan de "hotel strasbourg"

Annonces ⓘ

85 hôtels à Strasbourg
www.hotels.com/Strasbourg-Hôtels
 ★★★★★ 19 669 avis pour hotels.com
 Réservez maintenant votre hôtel à Strasbourg. Promo à saisir!
 481 531 personnes sont abonnées à la page Hotels.com sur Google+.

Hôtel Strasbourg
www.ibis.com/Strasbourg
 Réservez à Strasbourg dès 65€.
 Profitez de notre promotion été!

Hôtels à Strasbourg
www.lastminute.com/Hotel_Strasbourg
 Sélection XXL d'Hôtels Strasbourg!
 Profitez de Nos Meilleures Offres
 556 personnes sont abonnées à la page lastminute.com UK sur Google+.

Figure 1-8

Exemple de la requête « hôtel strasbourg » sur Google : dans ce cas, quasiment toute la visibilité « au-dessus de la ligne de flottaison » est occupée par les liens commerciaux et les résultats issus de Google Maps, l'outil Hotel Finder et Google Adresses (Google+ Local). Pour visualiser les liens naturels, il faut utiliser l'ascenseur et faire défiler la page. Le référencement sur Google Maps devient ici primordial (voir chapitre 7).

Résolutions d'écran

Pour connaître les résolutions d'écran les 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>.

Toutefois, on peut être plus exigeant et tenter de se positionner encore mieux en plaçant un site dans le triangle d'or (voir figure 1-9) des pages de résultats. En effet, selon une célèbre étude menée par les sociétés Enquiro et Dit-It.com, en collaboration avec la société EyeTool spécialisée dans les systèmes de *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 %.

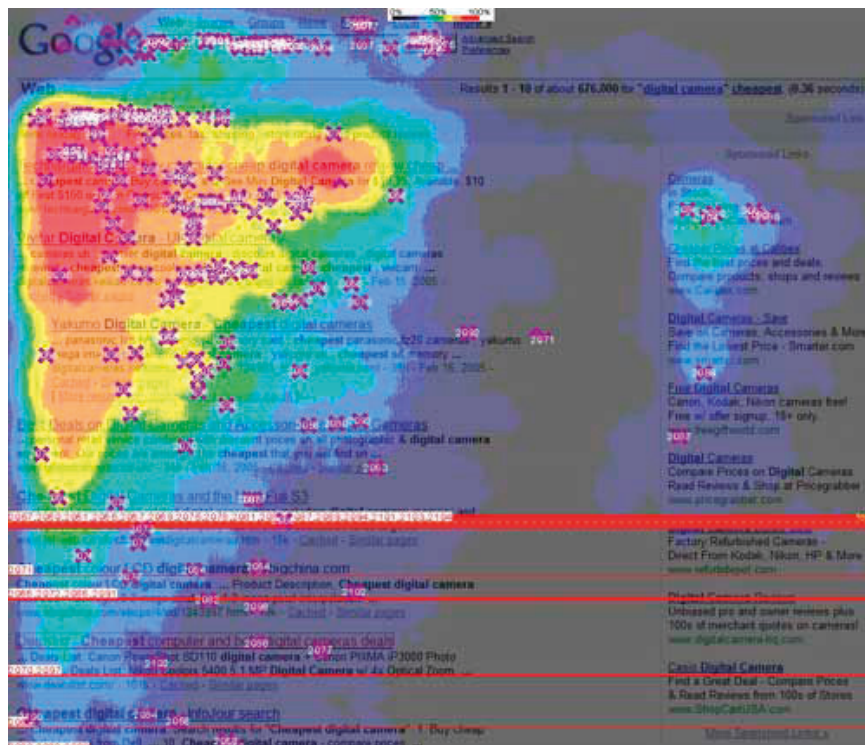


Figure 1-9

Le triangle d'or de la page de résultats de Google : plus le rouge est vif, 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

Vous trouverez davantage d'informations sur le triangle d'or des pages de résultats de Google à l'adresse suivante : <http://goo.gl/VcxOI>.

On peut même se poser une question : ne vaut-il pas mieux être 11^e, donc au-dessus de la ligne de flottaison de la deuxième page de résultats, plutôt que 10^e et en dessous de la ligne de flottaison de la première page ? 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.

Google vers l'infinite scrolling ?

Selon plusieurs sources, Google teste depuis plusieurs années un système de « scrolling infini », comme sur Facebook ou Twitter : lorsque vous descendez avec l'ascenseur dans la page de résultats, le moteur vous propose et rajoute autant de liens que nécessaire de façon automatique et transparente. Cela « tuerait » clairement la notion de pages 1, 2, 3, etc. Mais, début 2016, ce type d'affichage n'était toujours pas mis en œuvre. Une des questions principales serait certainement de savoir où placer les publicités AdWords dans ce cas... Plus d'infos à l'adresse suivante : <http://goo.gl/WEjgG>.

Google a publié en février 2009 quelques résultats de ses propres études menées sur l'eye-tracking (<http://goo.gl/PU3YH>).

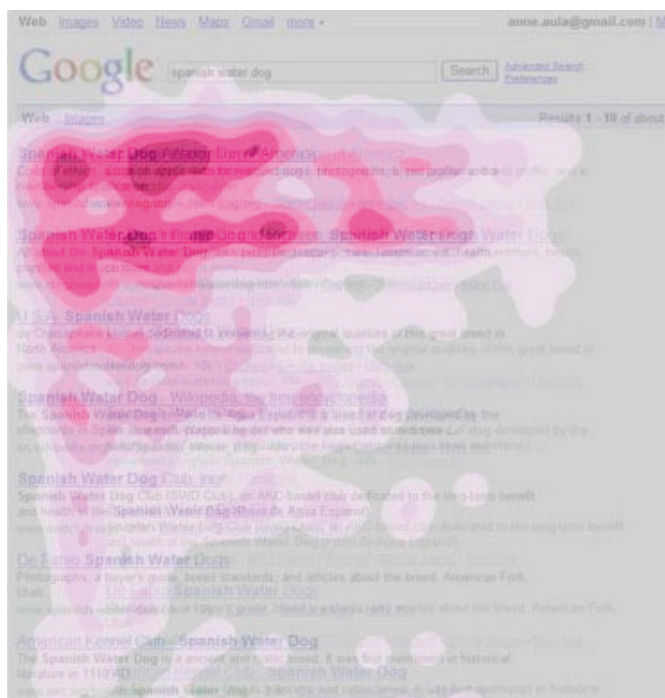


Figure 1-10

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

Enfin, signalons les très intéressantes études, dans le même domaine, effectuées par la société Miratech (<http://goo.gl/81fmM>), notamment au sujet de l'interaction entre les liens naturels et les liens commerciaux des pages de résultats de Google.

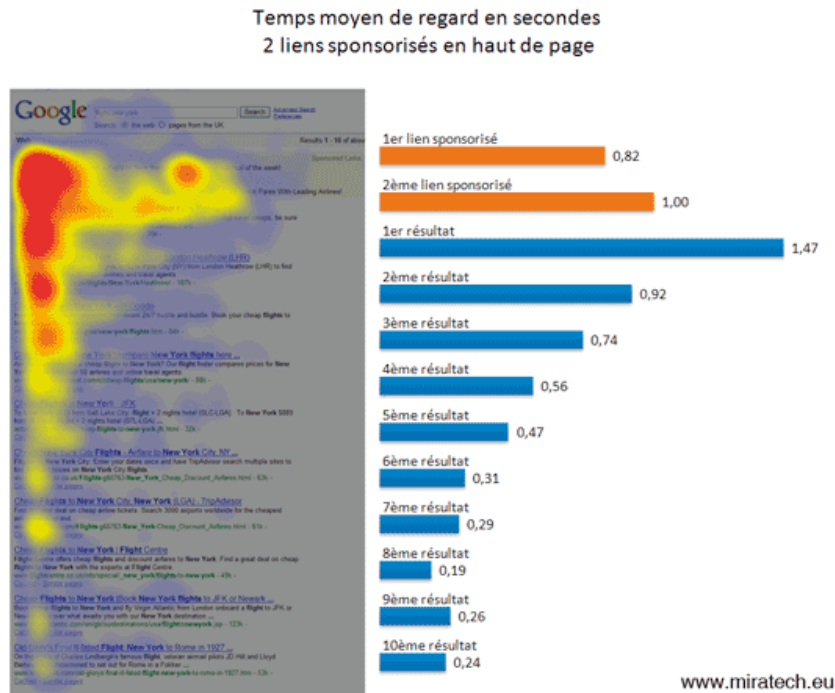


Figure 1-11

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

Visibilité dans les résultats des moteurs

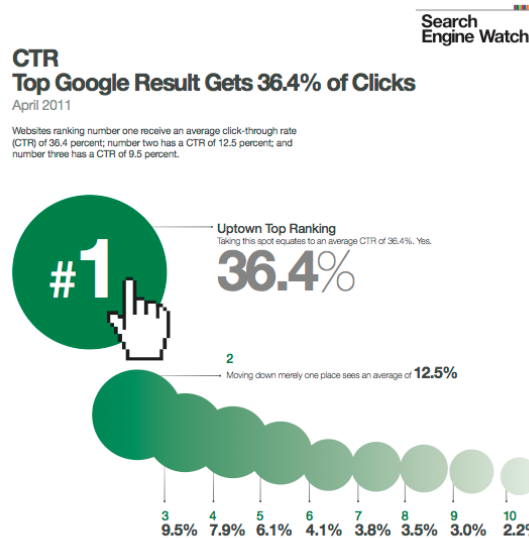
Une étude menée par la société Optify (<http://goo.gl/Ub3Rg>) indique que le taux de clics décroît très vite avec le positionnement du résultat (voir figure 1-12) : le premier lien capte 36,4 % des clics, le deuxième 12,5 % et le troisième 9,5 %, soit 58,4 % en tout pour le podium de résultats, le 10^e lien ne recevant plus que 2,2 % des clics.

D'autres chiffres du même type sont également disponibles ici :

- <http://goo.gl/t9nmnH> ;
- <http://goo.gl/gUWlsM> ;
- <http://goo.gl/p1cnqQ> ;
- <http://goo.gl/u9yqwW> ;
- <http://goo.gl/usuQWS>.

Figure 1-12

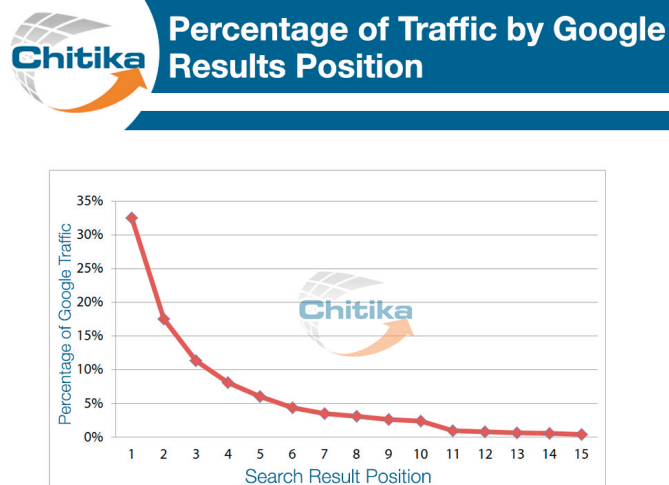
Selon Optify, les trois premiers liens de la page de résultats thésaurisent 58,4 % des clics.



Being number one in Google is the equivalent of all the traffic going to the sites appearing in the second through fifth positions.

Figure 1-13

Autres chiffres par la société Chitika en juin 2013
(<http://goo.gl/u9yqwW>)



La recherche universelle de Google

La donne change encore depuis la mise en place du projet « universal search » par Google (<http://goo.gl/B4hT4>), qui amène le célèbre moteur à mixer de plus en plus, dans ses pages de résultats, des liens issus de ses différents index : web, images, actualité, vidéos, cartes, etc. (figure 1-14). On l'a vu précédemment sur des recherches locales, pour lesquelles Google met en avant Google Maps. Une requête sur le mot-clé « neige » renvoie également en tête de liste des liens vers des images et des actualités qui repoussent les liens web organiques encore plus vers le bas de la page et donc « en dessous de la ligne de flottaison ».

Google neige

Web Images Maps Shopping Actualités Plus ▾ Outils de recherche

Environ 88 300 000 résultats (0,29 secondes)

Actualités correspondant à neige

[Neige : appels à la prudence sur les routes lundi](#)
Le Monde - il y a 3 heures
Avec de nouvelles chutes de **neige** et le retour d'un fort trafic, la situation sur les routes risque d'être difficile lundi matin, prévient le ministère de ...

[Neige et verglas : le point sur le trafic](#)
Challenges.fr - il y a 37 minutes
[Neige et verglas - Une journée de "galère" lundi](#)
Le Point - il y a 4 heures

[Le Figaro - France : Les transports perturbés par les chutes de neige](#)
www.lefigaro.fr/.../01016-20130120ARTFIG00132-les-tran...
Il y a 18 heures – Le trafic aérien est perturbé à Roissy et Orly, les TGV sont ralentis et aucun bus ne circule dans les régions touchées par la **neige**.

Neige - Wikipédia
fr.wikipedia.org/wiki/Neige
La **neige** est une forme de précipitation constituée de glace cristallisée et agglomérée en flocons pouvant être ramifiés d'une infinité de façons. Puisque les ...

Images correspondant à neige - Signaler des images inappropriées

Il y a 5 heures – En Ile-de-France, où le plan **neige**/verglas est intensifié en prévision d'une baisse des températures, les poids lourds sont interdits de 4 heures ...

Figure 1-14

Le concept de recherche universelle par Google consiste à mixer, dans les résultats de recherche, des données issues de différentes bases de données (ici, les images et les vidéos sur une capture d'écran de 2014, ce type de résultats ayant fortement diminué depuis 2015), ce qui décale les liens naturels vers le bas de la page de résultats !

18,5 % de résultats naturels « au-dessus de la ligne de flottaison »

Pour en terminer avec cette notion de visibilité dans les pages de résultats, citons une étude du site JitBit (<http://goo.gl/4uuWQN>) qui indique que les liens naturels « au-dessus de la ligne de flottaison », sur une page Google, ne représentaient fin 2012 plus que 18,5 % de l'espace proposé par le moteur de recherche, contre 53 % il y a quelques années. Le reste est aujourd'hui majoritairement occupé par de la publicité et des services proposés par Google (Maps, Actualité, Images, etc.). Une tendance qui se renforcera certainement encore dans les années qui viennent et qui nous pousse vers un référencement multiservice, multi-support !

Référencement et course à pied...

La situation est plutôt simple (si on peut dire) : 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 justement 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 – et surtout – 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'atteindre 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. On peut ainsi faire une analogie avec une course à pied : plus il y aura de concurrence et donc davantage de candidats motivés sur la ligne de départ, plus la course sera 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. Malgré tout, à cœur vaillant, rien d'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 s'il y a seulement quelques dizaines de participants au départ. En revanche, si plusieurs milliers de personnes participent, la difficulté n'en sera que plus accrue, et encore davantage si parmi elles se trouvent des professionnels de la course.

Pour continuer cette comparaison entre une course et le référencement, on peut mettre en évidence deux éléments.

- Il est tout à fait possible d'obtenir des premières pages sur Google en quelques heures sur des mots-clés non concurrentiels, grâce à une bonne optimisation de vos pages (code HTML et texte de qualité). Et ce, même si le moteur renvoie plusieurs dizaines de millions de résultats, voire plus. Si, si...

- En revanche, dès que la requête devient concurrentielle (plus il y a de « pros » de la course qui participent), le délai s'allonge. 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.

Vous pouvez ainsi 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.

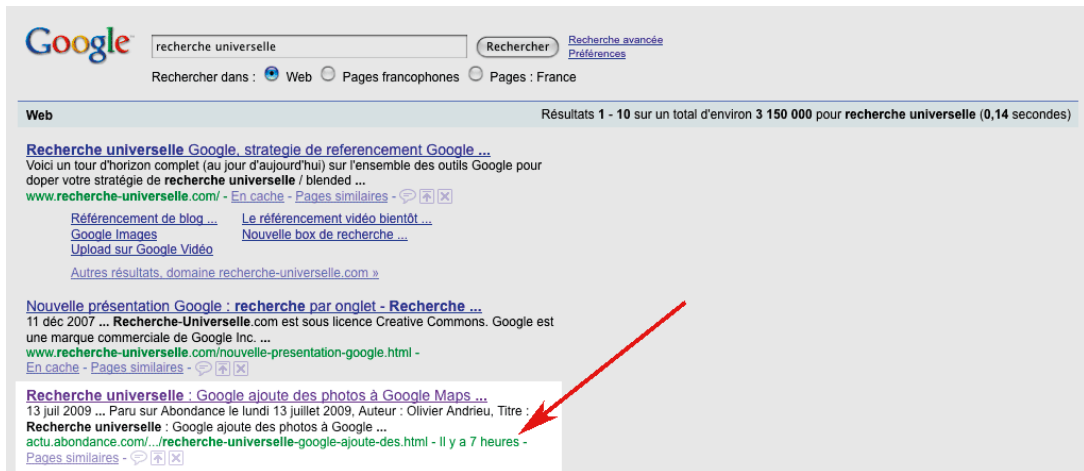


Figure 1-15

Exemple d'une page du site Abondance qui s'est placée en quelques heures sur le podium pour la requête « recherche universelle », qui engendre plus de dix millions de résultats sur Google. Copie d'écran d'époque !

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 stratégies intermédiaires peuvent tout à fait être envisagées : sur le podium pour certains termes, sur la première page pour certaines expressions moins importantes. Toutefois, 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.

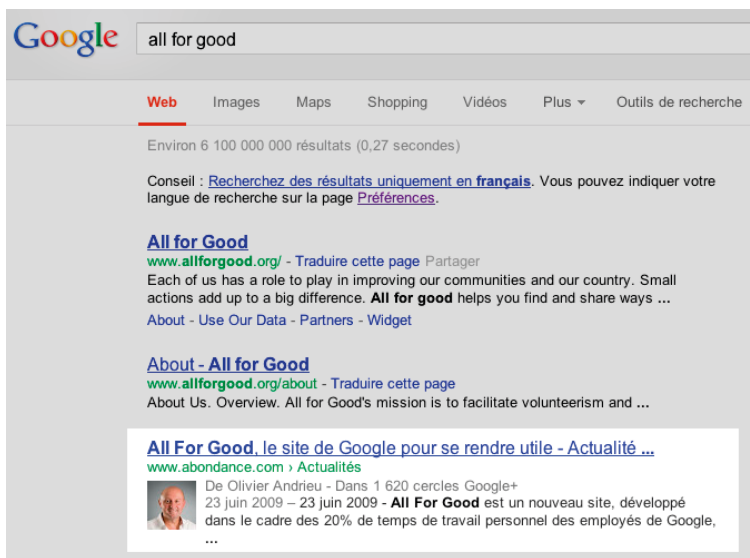


Figure 1-16

Autre exemple d'une page du site Abondance sur le podium de la requête « all for good » (6,1 milliards de résultats !). Positionnement obtenu quelques minutes après la mise en ligne de la page pour une requête certes très peu concurrentielle.

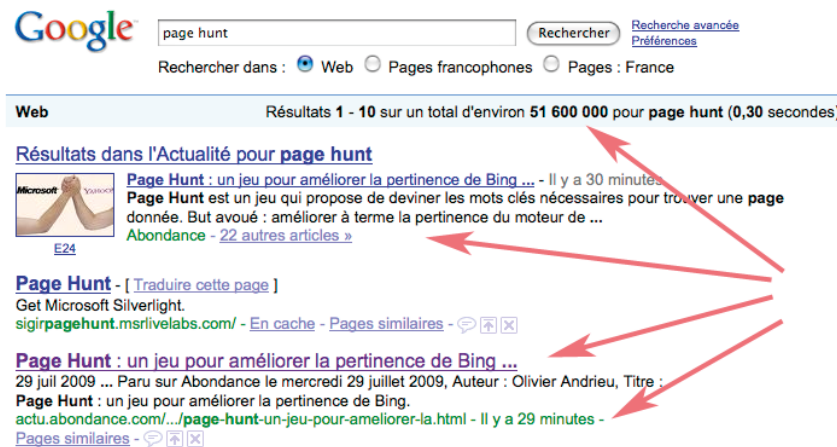


Figure 1-17

Un double positionnement (Google Web et Google Actualités) en moins d'une demi-heure sur la requête « page hunt », qui renvoie plus de 380 millions de résultats... mais qui n'est pas très concurrentielle, encore une fois. La situation aurait été bien différente sur des mots-clés pour lesquels la « bataille » est plus rude. Copie d'écran d'époque. Cette situation (double présence Actualités et Web) n'est plus possible aujourd'hui. Si un lien apparaît en première page de Google dans les actualités, il ne peut plus être affiché dans les résultats web. Il sera alors relégué en deuxième page de résultats.

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), objectif assez obsolète aujourd'hui ;
- dans la première page de résultats (les 10 premiers liens), un minimum à l'heure actuelle ;
- au-dessus de la ligne de flottaison, c'est-à-dire 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 loyale versus spamdexing

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

Deux écoles cohabitent actuellement 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 14). C'est donc une optimisation « à la source » du code HTML des pages du site, sans artifice. Vous devrez ensuite proposer un contenu de qualité, qui va naturellement attirer vers lui les liens d'autres sites. On parle alors de référencement *white hat* ou « chapeau blanc ».
- La deuxième école tente de manipuler et détourner les algorithmes des moteurs en utilisant des techniques interdites dont nous reparlerons au chapitre 15. Dans ce cas, on vise des stratégies *black hat* ou « chapeau noir ».

Notons que de nombreux webmasters tentent de se positionner entre les deux extrêmes en grisonnant plus ou moins légèrement leurs pratiques, ce qui donne du *grey hat* (et ses cinquantes nuances, bien sûr...).

La situation, en termes de *spamdexing*, est vieille comme le monde et le déclin des balises meta keywords, utilisées à l'époque par le moteur AltaVista, en est historiquement un bon exemple. L'abandon de la prise en compte de ces balises s'est déroulé en trois temps.

1. Les balises meta 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.
2. Certains webmasters sont allés trop loin et ont réellement fait n'importe quoi avec ces balises meta, les « truffant » notamment de mots-clés n'ayant aucun rapport avec le contenu du site, indiquant de nombreuses occurrences d'un même terme, proposant la même balise dans chaque page, etc.

3. Que s'est-il passé à l'époque pour ces balises ? 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 les géraient de façon propre en ont fait les frais.

Il en a été de même, quelques années plus tard, pour une méthode mettant en œuvre des pages web spécifiques, appelées « pages alias », « pages satellites », *doorway pages* ou encore « pages fantômes » (toutes ces dénominations désignant le même concept ou presque). Par exemple : une page satellite était construite pour l'expression « voyage Maroc ». Elle était optimisée pour cette requête et contenait une redirection vers la page qui traitait de ce thème sur le site du client. Si cette page satellite était bien positionnée dans les pages de résultats des moteurs, l'internaute cliquait dessus et était donc redirigé vers la « vraie » page du site qui, elle, n'était pas optimisée. Ce type de « rustine » a longtemps été utilisée sur le Web, au moins jusqu'en 2005 ou 2006.

Les moteurs de recherche considèrent aujourd'hui les pages satellites comme du spamdexing, tout comme de nombreuses autres techniques assimilées à du black hat (interdites par Google et ses confrères).

On pourrait multiplier de tels exemples à l'infini ou presque. De nombreuses méthodes qui avaient pourtant certains avantages (les pages satellites permettaient, par exemple, de pallier le fait que certains sites étaient réalisés à 100 % en Flash et donc très difficilement référençables) ont été interdites par les moteurs dans leurs recommandations en ligne. À ce sujet, n'hésitez pas à consulter les conseils techniques de Google aux adresses suivantes :

- <http://goo.gl/FN4pN> ;
- <http://goo.gl/ISEDq>.

Lisez attentivement ces deux pages, elles regorgent de recommandations très intéressantes. Toute démarche de SEO devrait commencer par cette étape préliminaire.

D'autres moteurs que Google proposent également dans leur site des *guidelines* assez précises dans ce domaine. Pour Yahoo!, consultez la page <http://goo.gl/pGSIJ> et pour Bing, la page <http://www.bing.com/toolbox/webmaster/>.

La conclusion nous semble donc évidente : n'utilisez plus de techniques interdites par Google et les autres moteurs ! L'unique bon référencement est bien celui qui repose sur l'optimisation à la source d'un site web de qualité, sans informations cachées dans les pages et sans liens factices pointant vers un site web. Ceux qui seront « blacklistés » ou pénalisés dans un proche avenir pour avoir abusé de techniques prohibées ne pourront pas dire qu'ils n'ont pas été prévenus. Nous en reparlerons au chapitre 15.

Il ne reste plus alors aux webmasters qu'à envisager un référencement basé sur une optimisation « propre ». Cela donne d'ailleurs d'excellents résultats, comme vous allez pouvoir vous en rendre compte en lisant cet ouvrage. Mais certains webmasters inventent presque chaque jour de nouvelles possibilités de contourner les algorithmes des moteurs, la grande tendance actuelle portant sur la création de liens entrants (backlinks) factices. Et le jeu des gendarmes et des voleurs continue. Jusqu'à quand ? Certainement la pénalisation du site.

Ne nous y trompons pas : en 2016, de nombreuses sociétés de référencement en France n'utilisent généralement pas de techniques « spammantes » comme système de référencement/positionnement et basent 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.

Pour cela, cependant, 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 réalisées (moins d'animations Flash, d'images, 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 moins qu'on puisse dire est que, sur ce point, on est loin d'une situation idéale en France. Très très loin...
- 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, sans artifice ni rustine à durée de vie limitée et sans backlinks artificiels. 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.

Ainsi, 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 blacklisting (mise en liste noire, voir chapitre 15) ou la pénalisation de certains sites par les moteurs. Cependant, 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 techniques critiquables sont-elles prêtes à cette révolution qui n'est peut-être d'ailleurs qu'une évolution ? L'avenir le dira.

Vous l'aurez compris, nous sommes clairement en faveur de l'optimisation propre (*white hat*) des pages constituant un site web. Ce sont ces pratiques d'optimisation loyales, aujourd'hui éprouvées, efficaces, et extrêmement pérennes, que nous allons expliciter dans la suite de cet ouvrage. Autant être clair dès le départ : si vous cherchez à savoir comment manipuler les algorithmes de classement de Google, ce n'est pas dans ce livre que vous trouverez l'information.

Le SEO, c'est comme une recette de gâteau !

Pour résumer en termes simples le référencement, nous faisons souvent une analogie, dans les formations que nous animons tout au long de l'année, entre le SEO et une recette de gâteau.

En effet, pour faire un bon gâteau, de quoi avez-vous besoin ?

- D'un moule de qualité, qui n'accroche pas et fait bien cuire le gâteau de façon uniforme.
- D'ingrédients, là aussi de la meilleure qualité possible : farine, œufs, lait, beurre, sucre, etc. (en fonction de la recette).
- D'un four qui permettra de faire cuire et « mijoter » le tout avant dégustation.



Figure 1-18

Trois points essentiels pour réussir un bon gâteau !

Il en est de même avec le SEO, pour lequel trois étapes sont indispensables.

- Le moule est le code HTML, la partie technique du codage qui doit être réalisée en amont de la création du site : présence des balises `title`, `meta description`, `Hn`, `canonical`, compatibilité W3C, etc. En somme, le « récipient » technique 100 % *search engine friendly* (optimisé pour les moteurs de recherche) qui va recevoir les « bons ingrédients ».
- Les ingrédients correspondent au contenu textuel de qualité, en quantité suffisante et avec le bon dosage !
- Enfin, le four est représenté par le netlinking, avec des liens les plus naturels (et les moins artificiels, les fours bas de gamme ça ne cuit pas bien !) possible, qui va « faire gonfler » votre œuvre et la transformer en un vrai festin !

**Figure 1-19**

Et trois points tout aussi capitaux quand on parle de SEO !

Bien sûr, il faut également que vos convives aiment votre gâteau. Dans ce cas, les réseaux sociaux et une bonne stratégie SMO feront l'affaire.

Ce schéma est bien sûr simpliste, mais il illustre bien, par expérience, la difficulté du métier de SEO, qui est à la croisée de trois métiers souvent très différents les uns des autres : la technique (le moule), le rédactionnel, qui demande un minimum de profil littéraire (les ingrédients) et le marketing qui s'occupera du netlinking. Pas toujours facile de faire s'accorder les trois, surtout quand une seule personne assume toutes ces tâches dans une même entreprise (et qu'on y ajoute une pincée de social). Mais le fait que ces trois métiers soient représentés dans une même société n'est pas toujours un gage de réussite programmée. Encore faut-il que « la mayonnaise prenne »... ce qui est un comble pour un gâteau.

Sur ce, bon appétit à tous et lisez bien les chapitres suivants pour élaborer vos propres recettes !

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 dans les pages suivantes que cela se passe.

2

Fonctionnement des outils de recherche



« Conviviale est la société où l'homme contrôle l'outil. »

Ivan Illich

Avant d'y référencer votre site, savez-vous ce que le moteur 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 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. Le fait de bien maîtriser le fonctionnement d'un outil de recherche vous permettra de mieux appréhender le référencement et l'optimisation de votre 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.

1. La collecte d'informations (*crawl*) grâce à des robots (aussi appelés *spiders* ou *crawlers*).
2. L'indexation des données collectées et la constitution d'une base de données de documents nommée « index ».
3. 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.
4. La restitution des résultats identifiés, dans ce qu'on appelle communément des SERP 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 au chapitre précédent, 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 commerciaux, issus d'un système d'enchères.

Nous allons nous concentrer ici sur les techniques utilisées par les moteurs 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 deux leaders du marché (Google et Bing), de nombreux moteurs n'utilisent pas leurs propres technologies de recherche mais sous-traitent cette partie auprès de grands moteurs. C'est le cas de Yahoo! qui utilise Bing, la technologie de Microsoft, pour

son moteur de recherche. Les deux sociétés ont signé un accord en ce sens fin juillet 2009 (<http://goo.gl/cD2zz>). En fait, il n'existe que peu de « fournisseurs de technologie » sur le marché : Google et Microsoft sont les principaux aux États-Unis, comme sur le plan mondial. En France, les acteurs majeurs sont Exalead (mais il n'est plus vraiment présent sur le marché), Qwant (qui utilise en partie Bing) et Orange/Voila, qui côtoient d'autres noms moins connus, et bien sûr les deux leaders Google et Bing. Voici un récapitulatif des technologies utilisées par les différents portails de recherche en 2016.

Tableau 2-1 Technologies de recherche utilisées par les principaux portails de recherche francophones en 2016

Technologies de recherche	Google	Yahoo!	Bing	Exalead	Qwant	Voila
Google	X					
Yahoo!			X			
Bing			X			
MSN			X			
Orange/Voila						X
Qwant			X		X	
AOL.fr	X					
Free	X					
Neuf/SFR	X					
Bouygues Telecom	X					
Exalead				X		

Tableau 2-2 Technologies de recherche utilisées par les principaux portails de recherche anglophones en 2016

Technologies de recherche	Google	Yahoo!	Bing
Google	X		
Yahoo!			X
Bing			X
MSN			X
Facebook			X
AOL	X		

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 dernières sont ensuite indexées par des moteurs d'indexation, les termes répertoriés enrichissant un index, qui consiste en une base de données des mots contenus dans les pages régulièrement mise à jour. Enfin, une interface de recherche permet de restituer des résultats aux utilisateurs en les classant par ordre de pertinence (phase de *ranking*).

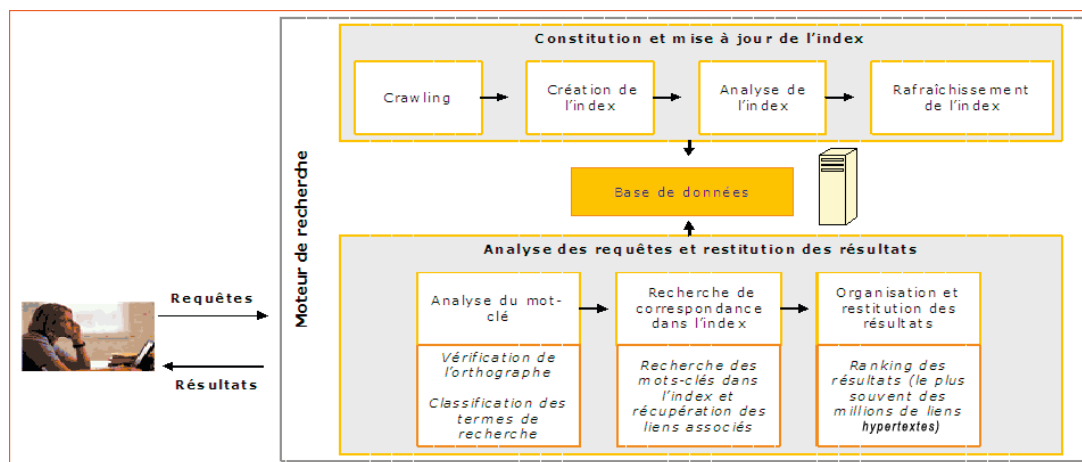


Figure 2-1

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.

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 (équivalent de l'option Enregistrer 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, l'enregistre, détecte les liens qu'elle contient (liens internes et externes), s'y rend, les sauvegarde, y détecte de nouveaux liens, etc. Et cela 24 h/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 est celle d'un internaute fou qui lirait et enregistrerait 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, ou BingBot de Bing.

Figure 2-2

Principe de fonctionnement d'un spider

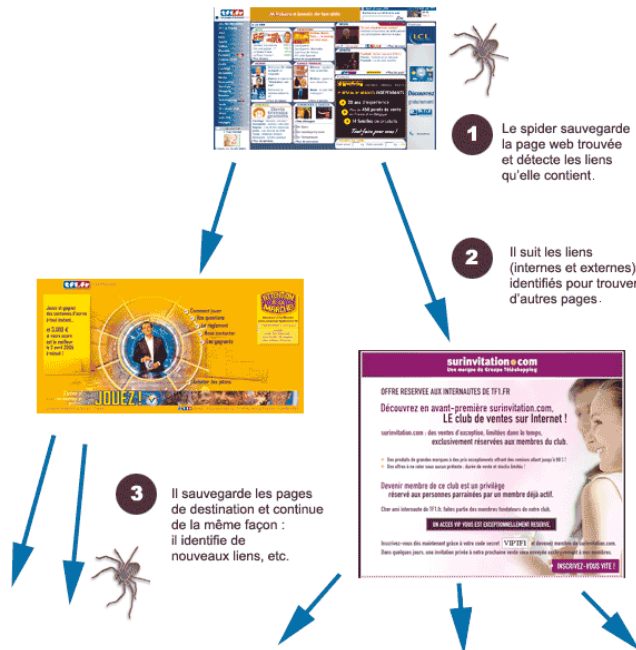
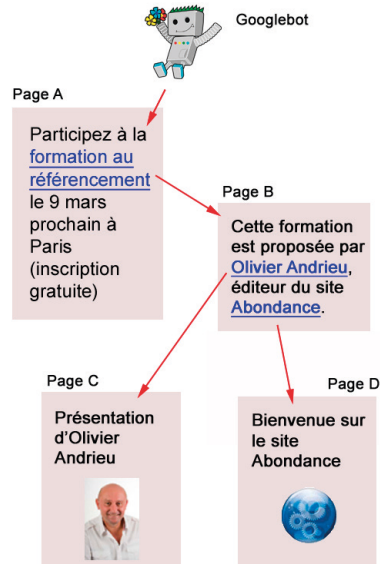


Figure 2-3

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



Le fichier robots.txt

Le fichier `robots.txt` est utilisé par les webmasters pour indiquer aux spiders les pages qu'ils ne souhaitent pas voir crawler (voir chapitre 16).

Cependant, 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 l'a déjà parcourue dans le passé, il contrôle si la version découverte est plus récente que celle qu'il possède. 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 le second type de modification.

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

De la « Google Dance » à l'indexation en quasi temps réel

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, maintenu 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. Pour la plupart, les moteurs gèrent le crawl de manière continue. 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, sur un site d'actualités) sera visitée chaque jour – voire plusieurs fois par jour – par le robot, tandis qu'une page rarement modifiée sera « crawlée » beaucoup moins souvent.

De plus, la mise à jour du document dans l'index du moteur est quasi immédiate. Ainsi, une page souvent mise à jour sera le plus souvent disponible à la recherche sur le moteur quelques heures, voire quelques minutes plus tard. Ces pages récemment crawlées sont par exemple identifiables sur Google qui affiche la date et l'heure d'indexation (voir figure 2-4).

890 changements d'algorithme pour Google en 2013 ...

www.abondance.com/.../20140820-14191-890-changements-dalgorithme...

Il y a 8 heures - 20 août 2014 - Amit Singhal, qui est à la tête du moteur de recherche de ... apportées par l'outil depuis 10 ans et fait u par Actualité Abondance.

Figure 2-4

Affichage par Google de la date d'indexation de la page. Ce délai peut être très rapide, parfois de l'ordre de quelques minutes.

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 à l'adresse suivante : <http://goo.gl/Sr0W0>.

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 qui 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. Mais certaines plates-formes comme WordPress peuvent effectuer un *ping*, c'est-à-dire alerter les moteurs en leur envoyant un signal numérique, dès parution d'un nouveau contenu ;
- 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. Nous y reviendrons au chapitre 14.

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). Les outils statistiques comprennent généralement dans leurs graphiques ou données une rubrique « 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. Voici quelques solutions gratuites :

- RobotStats : <http://www.robotstats.com> ;
- SpyWords : <http://www.spywords.com> ;
- Watussi Box : <http://box.watussi.fr> ;
- Botify : <http://www.botify.com>.

Des « marqueurs » doivent parfois 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.

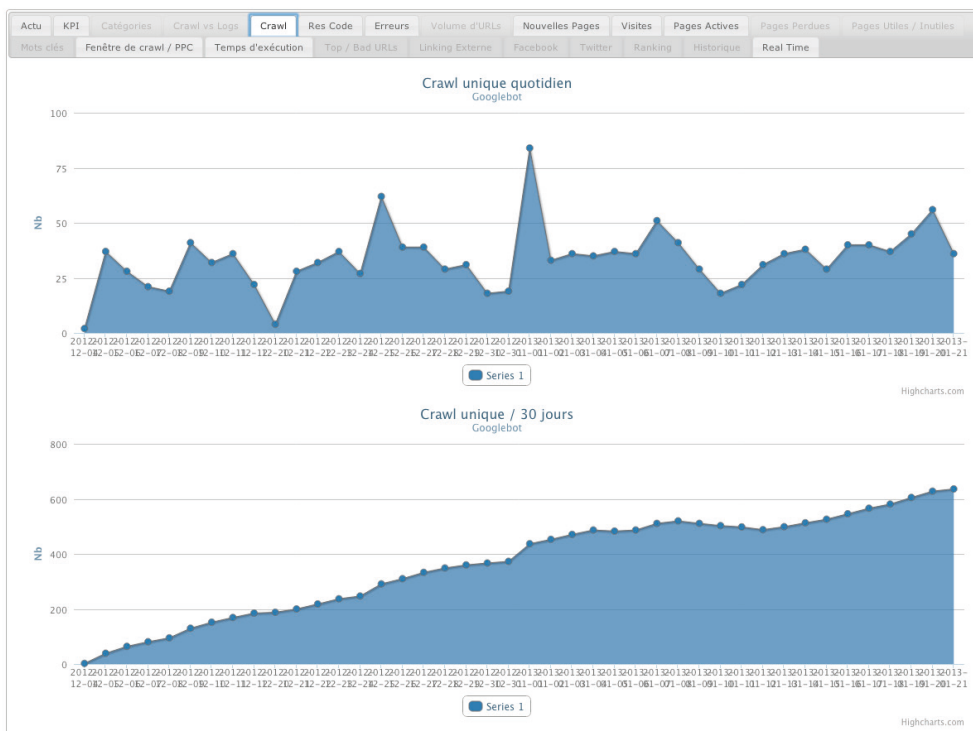


Figure 2-5

Exemple de statistiques fournies par un utilitaire de statistiques en ligne (ici la Watussi Box)

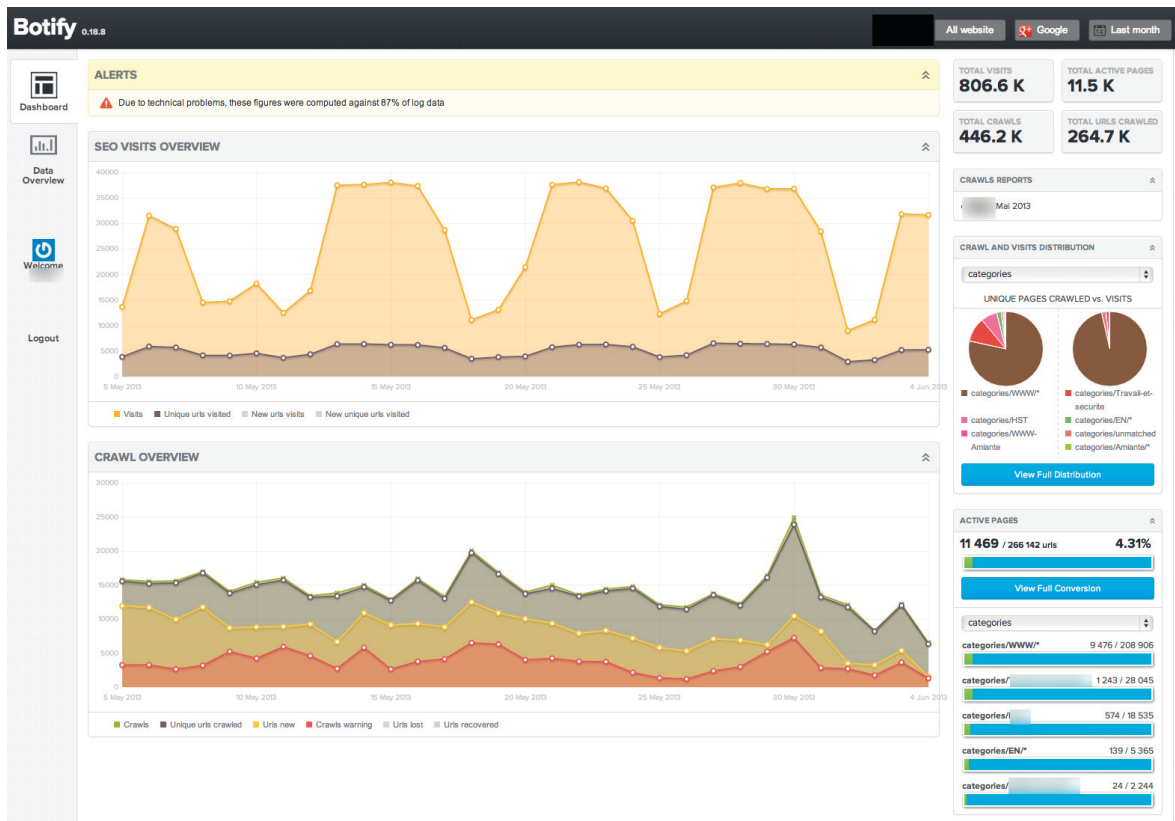


Figure 2-6

Exemple de statistiques fournies par Botify, un service très complet et très professionnel d'analyse du crawl d'un site par Google. Un outil payant à utiliser pour les sites ayant plus de 10 000 pages indexées par Google.

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. L'indexation s'effectue en texte intégral : tous les mots d'une page, et plus globalement son code HTML, sont alors pris en compte.

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, mais elle a laissé dans les esprits l'idée qu'« il ne faut pas proposer plus de 100 liens par page ». En effet, aux débuts de Google, il n'était pas intéressant de dépasser la centaine de liens sortant

d'une page puisque le moteur ne lisait pas tout le code. Mais cette époque est révolue depuis bien longtemps et si vous proposez 500 liens dans une page, Google les lira sans problème ! Mais certaines légendes urbaines ont encore et toujours la vie dure. D'autres moteurs peuvent effectuer une sélection en fonction des formats de document (tableur Excel, présentation PowerPoint, fichier PDF, etc.).

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 alors 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 les moteurs de recherche ignorent les stop words tels que « le », « la », « un », « de », « et », etc. (en anglais, « the », « a », « of », etc.). Ceci est exact, comme mentionné dans l'explication de Google dans son aide en ligne (<http://www.google.fr/intl/fr/help/basics.html>) :

« Google ignore les chaînes de caractères dont le poids sémantique est trop faible (également désignés « mots vides » ou « bruit ») : le, la, les, du, avec, vous, etc., mais aussi des mots spécialisés tels que « http » et « .com » et 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. Mais ça n'est pas le cas ! S'il existe 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 vide 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 ».

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 le stop word « le » 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 une seule façon de le faire : avec les guillemets.

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.

Notons enfin que le signe +, pour demander une orthographe exacte d'une requête sur Google, ne fonctionne plus sur ce moteur de recherche. Consultez la page suivante pour plus d'informations : <http://goo.gl/qwQrb>.

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/>) contenant 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 ».

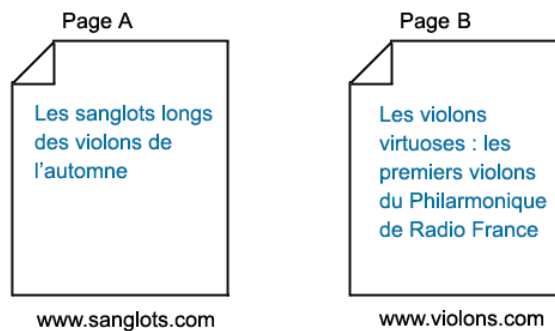


Figure 2-7

Deux pages prêtes à être indexées par un moteur de recherche

Les données du tableau 2-3 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. Retenez 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).

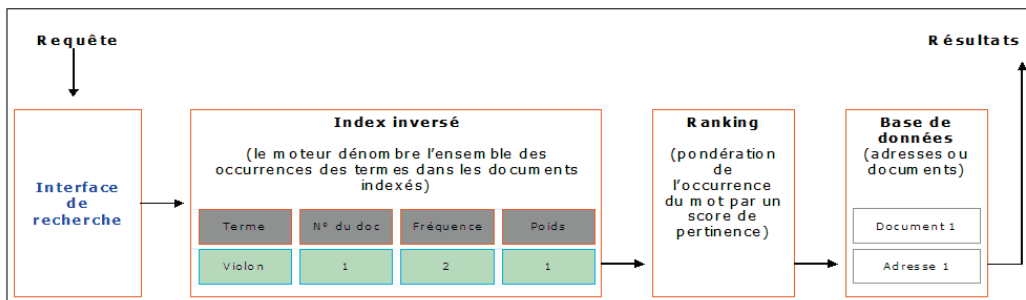


Figure 2-8

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 6 – avec la page pointée (considérant que ces liens renvoyant vers une page fournissent parfois une description plus précise, ou en tout cas plus concise, de la page que le document lui-même). Cependant, un moteur de recherche utilise des dizaines de critères de pertinence différents. Nous y reviendrons là encore.

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 – représentant une copie du Web à un instant T – 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.

Presque tous les 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 avait arrêté de communiquer sur ce point en 2005 alors que sa base d'URL proposait aux environs de 8 milliards de pages après avoir passé la barre du milliard en 2000. Un post sur le blog officiel du moteur (<http://goo.gl/eKZKb>) 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) ou beaucoup de pages totalement inutiles. Ce chiffre représentait donc le nombre de pages web « connues » de Google avant traitement. En 2013, on pouvait estimer la taille de son index à environ 100 milliards de pages au minimum, certainement beaucoup plus en vérité.

En mai 2013, le site Business Insider (<http://goo.gl/vbttDF>) reprenait les informations et les chiffres assez ahurissants donnés à Bloomberg par John Wiley, *lead designer* au sein du moteur de recherche de Google.

- Google arrivait à identifier 30 mille milliards (10^{12}) d'URL uniques sur le Web.
- Chaque jour, Google crawlait 20 milliards de sites web.
- Le moteur de recherche traitait 100 milliards de recherches chaque mois. Parmi celles-ci, 15 % (soit environ 500 millions par jour) étaient totalement inédites et n'avaient jamais été traitées auparavant.

Les syntagmes, prochaine étape des index du futur ?

De nouvelles méthodes d'indexation se mettent en place, autour de la prise en compte des syntagmes ou groupes de mots (contrairement aux mots isolés analysés jusqu'à maintenant), ce qui pourrait profondément changer le paysage du référencement dans les années à venir. Cela signifie que pour améliorer la qualité du moteur de recherche, par exemple dans la phrase « Le chien du voisin a aboyé toute la nuit » on pourra isoler les trois syntagmes suivants : « Le chien du voisin », « a aboyé » et « toute la nuit ». Le moteur devra alors être capable d'identifier que certains groupes de mots sont effectivement liés entre eux, alors que d'autres ne le sont pas. Un tel moteur sera capable de reconnaître que la phrase « certains quartiers avec des architectures modernes ont vu le jour à côté de la vieille ville » ne parle pas de l'« architecture moderne de la vieille ville » mais bien des quartiers qui disposent de ce type d'architecture. On le voit, le défi est énorme et la difficulté non négligeable. Il semble donc intéressant de passer à un système capable d'indexer des groupes de mots (syntagmes) au lieu de simples mots-clés isolés. Pendant très longtemps, les méthodes d'indexation de syntagmes ont buté sur un écueil considérable : lorsqu'on indexe des groupes de mots, au lieu de mots isolés, la taille de l'index explose littéralement. La mémoire nécessaire pour identifier les combinaisons de trois, quatre, cinq mots est également un obstacle.

Par exemple, si on part du principe qu'on veut indexer toutes les combinaisons de cinq mots, et que le corpus (l'ensemble des textes à indexer) contient 200 000 termes différents, on aura dans ce cas $3,2 \times 10^{26}$ syntagmes possibles (soit 3 suivi de 26 zéros). Ce chiffre dépasse les capacités de tout système existant, et même imaginable.

Dans la pratique, toutes les combinaisons de syntagmes ne sont pas utiles dans l'index. Voici un exemple donné par Google, à propos d'une base de textes issus de pages web qu'il met à disposition pour les recherches des linguistes (base Web 1T 5-gram).

- La base de départ contient 1 024 908 267 229 « tokens » (c'est-à-dire de termes, éventuellement présents plusieurs fois et même de très nombreuses fois dans l'ensemble des textes, autrement appelé corpus).
- Le nombre d'unigrammes (termes uniques pris isolément) s'élève à 13 588 391.
- Le nombre de bigrammes (couples de termes présents) est de 314 843 401.
- Le nombre de trigrammes (triplets de termes présents) s'élève à 977 069 902.
- Le nombre de 4-grammes est de 1 313 818 354 !
- etc.

Dans cet exemple, le nombre ne suit pas une logique combinatoire : le nombre de n -grammes identifiés correspond à des groupes de mots constituant des séquences non pas aléatoires mais qui présentent une certaine fréquence d'apparition en commun (fréquence de cooccurrence ou plus précisément fréquence de « collocation »). Cela reste malgré tout intéressant de noter qu'on est obligé de multiplier le nombre de lignes dans l'index par 100 pour parvenir à stocker des expressions contenant jusqu'à 4 mots.

Une autre voie, plus simple et suivie actuellement par les moteurs, est la détection des « entités nommées » comme les noms de personnes, de lieux, d'entreprises, etc., pour fournir des résultats plus précis et plus pertinents aux internautes. C'est sur ce concept que repose le *Knowledge Graph* lancé en France par Google en décembre 2012 (<http://goo.gl/OSv08>) suite à l'acquisition de l'outil Freebase en 2010.

Mais il est clair 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, début 2016)

	Google	Bing	Exalead	Orange*
Taille de l'index	100	40	16	1

* Moteur spécialisé sur le Web francophone.

Les moteurs ne communiquent plus sur ces chiffres et rendent très complexe la mise à jour de ce type d'information.

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 recherche. Pour cela, les moteurs élaborent en permanence de nouveaux algorithmes (des formules mathématiques utilisées pour classer les documents), qui représentent un véritable facteur différenciant. Ces algorithmes ne sont donc que très rarement rendus publics. Ils sont dans la plupart des cas protégés par des brevets et font parfois l'objet de « secrets défense » 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 plusieurs facteurs appliqués aux termes de la recherche (toutes ces notions seront abordées en détail aux chapitres 4 et 5) :
 - localisation d'un mot dans le document (exemple : le poids est maximal si le mot apparaît dans la balise <title>, la balise <h1> 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), titre éditorial (balises <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).

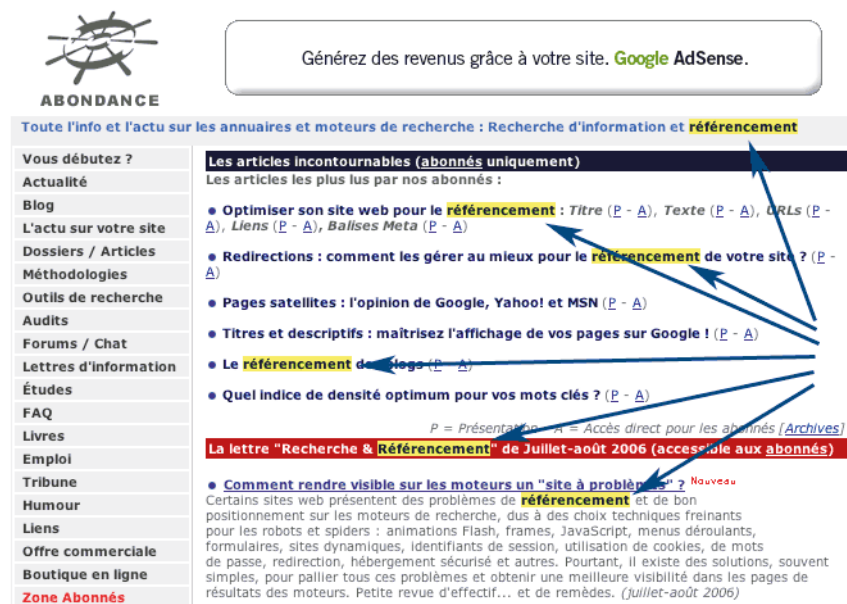


Figure 2-9

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

Tous ces critères sont basés sur la présence des mots-clés de la requête dans une ou plusieurs zone(s) chaude(s) de la page web.

- **Le tri par popularité (indice de popularité).** Popularisé – mais pas inventé – par Google en 1998 avec son PageRank (pour contrer, entre autres, les abus possibles des méthodes de tri par pertinence), 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. Ce type de tri est *a priori* indépendant du contenu. Il s'agit en fait de l'analyse des liens entrants (backlinks) pointant sur une page donnée.

Ainsi, Google classe les documents principalement en fonction de leur PageRank (nombre et qualité des liens pointant vers ces documents, nous y reviendrons en détail au chapitre 6). Le moteur analyse alors les pages contenant les liens. Plus une page est pointée par des liens émanant de pages elles-mêmes populaires, plus sa popularité (son PageRank) 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, *SERP Bounce* ou *Pogosticking*).** 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 aujourd'hui par certains moteurs dont Google.

- **Le tri par catégorie 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, par le Français Exalead (<http://www.exalead.com/>) et l'Américain Vivisimo (<http://vivisimo.com/>), racheté depuis par IBM, ainsi que sur la version américaine de Bing (<http://www.bing.com/>) grâce à la technologie de la société Powerset, entreprise rachetée par Microsoft en 2008.

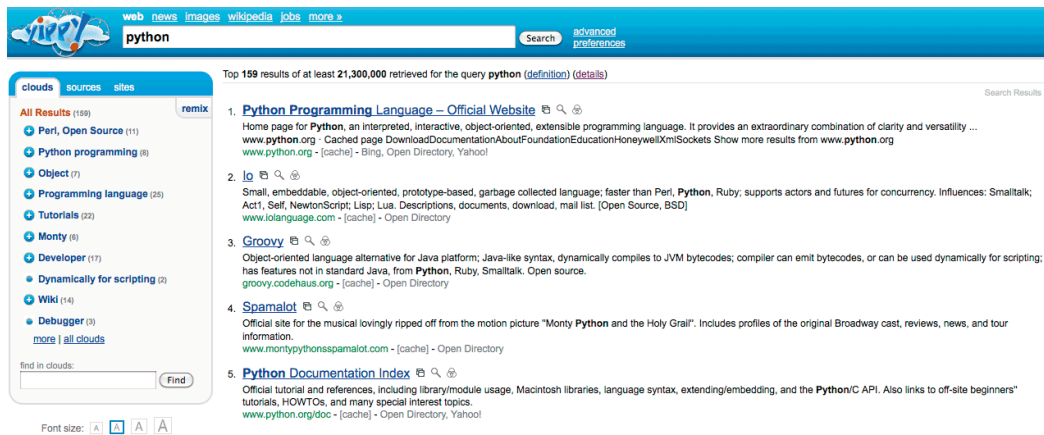


Figure 2-10

Le moteur de recherche Yippy (<http://yippy.com/>) « clusterise » ses résultats : il propose, sur la gauche de l'écran, des dossiers thématiques qui regroupent les résultats par grands domaines.

Les méthodes de tri

Pour plus d'informations sur les techniques de tri existantes, consultez l'article de Jean-Pierre Lardy intitulé « Méthodes de tri des résultats des moteurs de recherche » et disponible à l'adresse suivante : <http://goo.gl/eFumFf>.

Bien entendu, plusieurs de ces méthodes peuvent être utilisées simultanément par un moteur. C'est le cas aujourd'hui de Google, Bing et des principaux outils de recherche qui associent tris par pertinence, par popularité, etc., pour obtenir les meilleurs résultats possibles.

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 généralement offerts. À chaque question, une requête est créée dans l'index et une page web dynamique restitue les résultats, souvent sous forme de listes ou de cartes de résultats (figure 2-11).

YAHOO! SEARCH [Yahoo! - Yahoo! Search - Aide](#)

Recherche avancée

Avec les différentes options sur cette page, vous pouvez formuler une requête très précise. Il vous suffit de remplir les champs dont vous avez besoin pour votre recherche en cours. [Yahoo! Search](#)

Faire une recherche sur

tous ces mots [sur la page](#)

la phrase exacte [sur la page](#)

au moins l'un de ces mots [sur la page](#)

aucun de ces mots [sur la page](#)

Astuce : Utilisez ces options pour rechercher une phrase exacte ou pour exclure des pages contenant certains mots. Vous pouvez aussi limiter votre recherche à certaines parties des pages.

Site/Domaine

☒ n'importe quel domaine

☐ domaines en .com uniquement ☐ domaines en .edu uniquement

☐ domaines en .gov uniquement ☐ domaines en .org uniquement

☐ domaines en .fr uniquement

☐ rechercher seulement dans ce domaine/site :

Astuce : Vous pouvez limiter votre recherche à des sites Web spécifiques (par exemple, yahoo.com) ou à des domaines de premier niveau (par exemple, .com, .org, .gov).

Format de fichiers

Ne donner que des résultats au format : [tous les formats](#)

Filtre adulte

Uniquement pour cette recherche :

☐ **Strict :** Filtrer les résultats à caractère pornographique (pages Web, images et vidéos) - **Filtre activé**

☒ **Modéré :** Filtrer les résultats à caractère pornographique pour les images et vidéos uniquement - **Filtre activé**

☐ **Désactivé :** Ne pas filtrer les résultats web (ils peuvent inclure du contenu à caractère pornographique) - **Filtre désactivé**

Remarque : tout utilisateur ayant ouvert une session sur votre ordinateur en tant que personne majeure pourra modifier ce paramètre. Nous vous recommandons de vérifier régulièrement les paramètres de verrouillage du filtre adulte.

Avertissement : le filtre adulte est conçu pour filtrer le plus grand nombre de contenus à caractère pornographique dans les résultats de Yahoo! Search. Cependant, nous ne pouvons pas garantir l'exclusion de l'intégralité de ce type de contenus de vos résultats de recherche.

[En savoir plus](#) sur la protection des enfants en ligne.

Astuce : si vous souhaitez bloquer les contenus à caractère pornographique pour toutes les recherches, vous pouvez le faire dans [préférences](#). Notez que ce filtre peut ne pas bloquer l'intégralité des contenus que vous pourriez trouver potentiellement choquants.

Pays [Tous les pays](#)

Figure 2-11

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 (*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 (<http://goo.gl/fZdcsP>).

Dévoilée au début des années 2000 (et probablement toujours similaire à l'heure actuelle, même si plusieurs projets, dont les célèbres BigDaddy et Caffeine, l'ont renouvelée), l'architecture de Google (figure 2-12) fait apparaître l'interconnexion de plusieurs composants séparés.

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, etc.).
- L'indexeur distribue les occurrences dans un ensemble de silos (*barrels*), organisés par *docID*.
- Le gestionnaire d'ancres (*anchors*) stocke certaines informations créées par l'indexeur, à savoir les liens hypertextes et les ancres qui leur sont associées (textes des liens).
- Le solveur d'URL (*URL Resolver*) récupère les informations fournies par le gestionnaire d'ancres 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 créer l'index inversé, stocké dans les mêmes barrels.
- La liste des mots fournie par le trieur est comparée avec celle du lexique (*lexicon*) et tout mot ne figurant pas dans ce dernier y est ajouté.
- Enfin, l'interface de recherche (*searcher*) exécute les recherches pour répondre aux requêtes des utilisateurs. Elle 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.

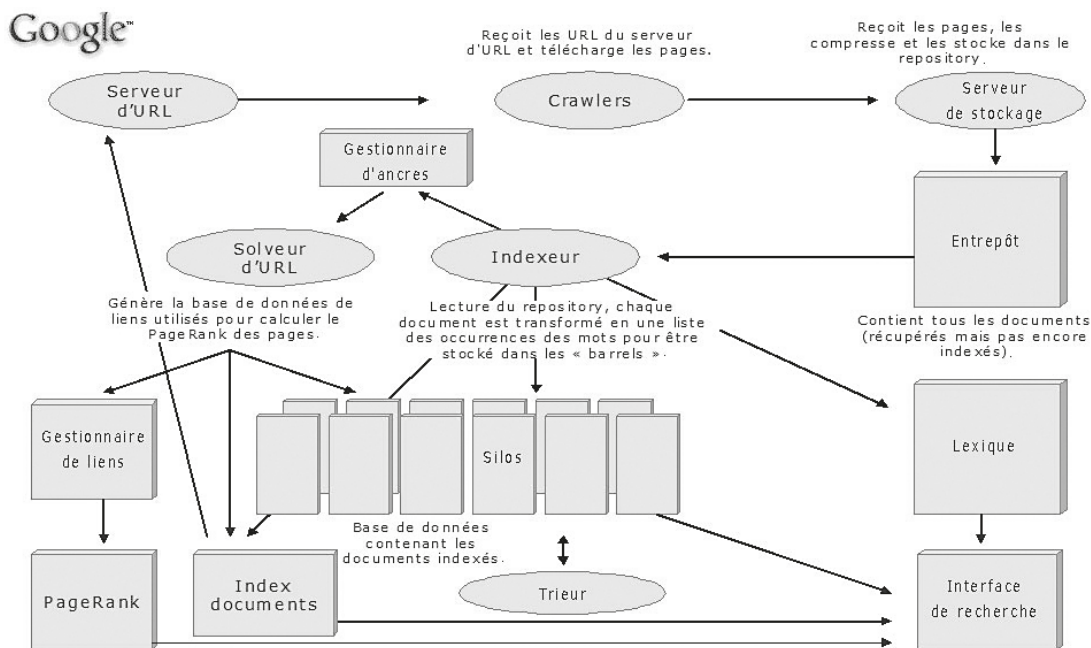


Figure 2-12

Architecture fonctionnelle de Google d'après Sergey Brin et Lawrence Page

(Source : *The Anatomy of a Large-Scale Hypertextual Web Search Engine*, <http://goo.gl/ERo95>)

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 raisonnablement

penser que son mode de fonctionnement global est encore aujourd'hui proche de ce que nous venons de décrire.

Figure 2-13

Autre vision simplifiée du mode de fonctionnement de Google, fournie par les deux cocréateurs du moteur (Source : The Anatomy of a Large-Scale Hypertextual Web Search Engine, <http://goo.gl/ERo95>)

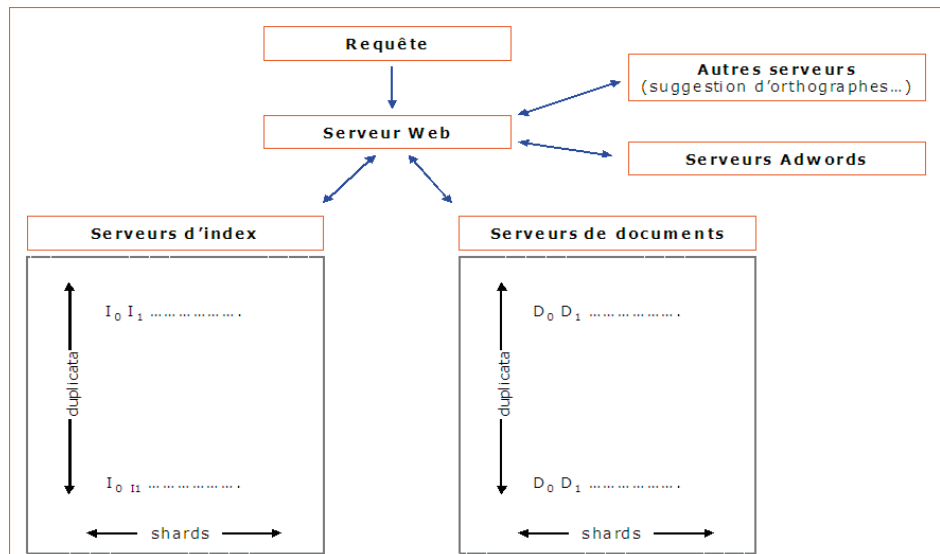
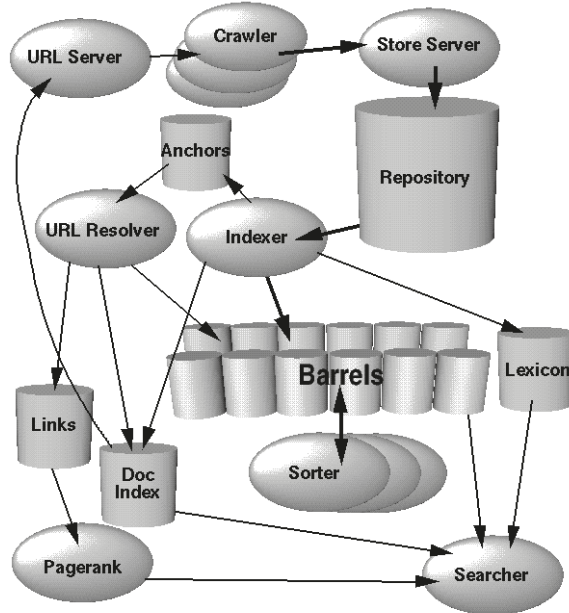


Figure 2-14

Schéma de l'utilisation des serveurs de Google utilisés pour la réponse aux requêtes (Source : <http://goo.gl/ArO7T>)

Mayday, Caffeine, Jazz : quoi de neuf ?

Section rédigée avec la contribution de Philippe Yonnet

En juin 2010, Google annonçait son basculement sur une nouvelle infrastructure technique baptisée Caffeine (<http://goo.gl/9BRcT>).

Caffeine, une nouvelle infrastructure d'indexation

Caffeine est en effet le nom d'une nouvelle infrastructure de Google, dont le déploiement avait été annoncé dès l'été 2009. Il ne s'agit donc pas d'un changement concernant le classement des pages, mais bien dans la façon dont Google explore le Web pour en extraire l'information, dont le moteur analyse, stocke et indexe ces données, et il traite les requêtes. Notons que Caffeine est à l'origine d'un autre séisme intervenu en 2011, sous le nom de Panda, et dont nous parlerons au chapitre 15.

Dans l'annonce officielle publiée sur le blog de Google destiné aux webmasters (<http://goo.gl/Ow6pB>), un petit graphe et un court commentaire révélaient un changement profond dans le comportement de crawl de Google.

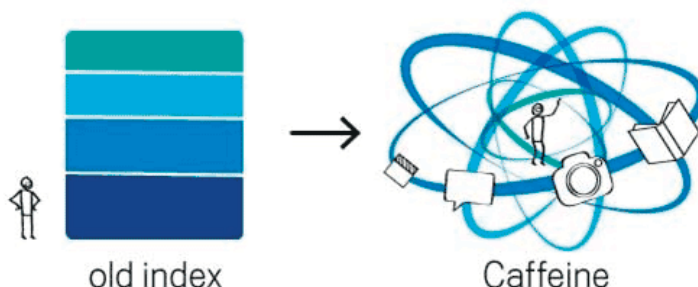


Figure 2-15

Changements impliqués par la nouvelle structure Caffeine (Source : Google)

Selon le billet officiel, l'ancien index (pré-Caffeine) était structuré en différentes couches. Pour construire chacune d'elles, il était nécessaire de crawler l'ensemble du Web à chaque fois. Certaines couches étaient mises à jour plus fréquemment que d'autres et la plus importante était rafraîchie toutes les deux semaines environ.

Dans le comportement de crawl post-Caffeine, le Web est analysé par petits bouts et les mises à jour s'effectuent de manière continue partout dans le monde. Sur certains sites, un nouveau comportement est apparu, Googlebot se mettant à crawler dix fois moins de pages qu'autrefois. Ce nouveau comportement semble très « ciblé » : certaines pages sont crawlées très souvent alors que d'autres moins.

Une étude menée par la société française de référencement Aposition (<http://goo.gl/yMceA>) a démontré que la mise en place de Caffeine ne changeait pas vraiment la fraîcheur des pages indexées.

En revanche, sur la capacité à indexer des pages rapidement sur un sujet d'actualité « chaud », un autre test mené par l'équipe d'Aposition démontrait que Caffeine semble changer la donne et que l'intégration de nouvelles pages s'effectue à un rythme clairement accéléré avec Caffeine pour des thématiques strictement liées à l'actualité (un algorithme baptisé *Freshness Update* semblait confirmer ce type de comportement fin 2011, <http://goo.gl/Q8TgDW>).

On peut donc estimer que ce type de crawl de Google tend à favoriser les sites d'actualités, qui peuvent voir leur contenu (mais aussi leurs images et leurs vidéos) être découverts beaucoup plus rapidement et être indexés, si besoin est, dans les minutes qui suivent leur publication. À l'inverse, certains sites risquent de subir des problèmes d'indexation dus à un crawl plus paresseux et sélectif. Pour ces derniers, en cas de chute du trafic, il sera indispensable d'analyser l'évolution de la liste des pages qui reçoivent du trafic, la quantité de trafic apporté par telle ou telle catégorie de requêtes, ainsi que la liste des pages explorées par Googlebot et la fréquence de *recrawl* de ces dernières. Ces analyses peuvent vous permettre de comprendre quelles sont les pages que privilégie Google sur votre site et de corriger éventuellement le tir si certaines d'entre elles, importantes, sont « oubliées » par le moteur.

Beaucoup de choses ont donc changé dans Google depuis 2010 et les filtres Panda (2011) et Penguin (2012) n'ont fait que confirmer ce phénomène. D'une manière générale, le rythme des changements s'était déjà accéléré dès l'été 2008. De plus, on peut être sûr que les possibilités ouvertes par Caffeine pour traiter « plus de données plus vite » vont rendre possible le déploiement de nouvelles fonctionnalités. Enfin, la nécessité de garder de l'avance sur des concurrents comme Bing semble pousser de plus en plus Google à lancer des nouveautés sur son interface et à essayer de nouvelles approches.

Pour les référenceurs, l'environnement change à un rythme accéléré. Il faut donc s'habituer à l'idée que, d'un mois sur l'autre, il faille imaginer de nouvelles stratégies et savoir réagir à un changement de comportement du moteur. Cela rend toutefois l'exercice encore plus intéressant.

Des SERP à 4 ou 7 liens !

Depuis des années, on a l'habitude de visualiser des pages de résultats (SERP) proposant 10 liens après avoir saisi sa requête. Mais, dès le mois de février 2012, plusieurs sources d'informations remarquaient que Google renvoyait moins de liens sur certaines requêtes (<http://goo.gl/7G6JK>). Ainsi, lorsqu'on saisisait « bbc football » sur Google UK, seuls 3 liens étaient proposés. En août 2012, de nombreuses pages (20 % environ) proposaient moins de 10 résultats (7 la plupart du temps). Et en septembre 2012, des SERP à 4 liens sont même apparus !

Autant dire que si ces affichages se répandaient de façon plus importante qu'actuellement, un site pourrait donc figurer en cinquième position mais apparaître sur la deuxième page de résultats, ce qui, en termes de visibilité, est loin d'être la même chose... Et les outils de mesure du positionnement d'un site web sur certaines requêtes prédéfinies devront, eux aussi, prendre en compte cette nouvelle donne.

Arrivés à la fin de ce chapitre, vous devez avoir une bonne vision du fonctionnement des moteurs. Suffisante, en tout cas, pour bien appréhender leur organisation et la façon dont ils récupèrent, analysent et classent les données glanées sur le Web. Il est donc temps de mettre en place une stratégie de référencement efficace et bien organisée, ce que nous allons voir au chapitre suivant.

Quelques liens sur le fonctionnement des moteurs de recherche

Sur les robots :

- The Web Robots Pages : <http://www.robotstxt.org/>.
Site fournissant une liste des robots actifs.

Sur Google :

- <http://goo.gl/VwQgL>
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://goo.gl/x6b8e>
Article de l'IEEE Computer Society – Web Search For A Planet, *The Google Cluster Architecture*.

Lexiques sur les moteurs de recherche :

- <http://www.idf.net/mdr/seta.html> ;
- <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.