



HAL
open science

Référencement naturel, entre éthique et spamdexing

Antoine Brisset

► **To cite this version:**

Antoine Brisset. Référencement naturel, entre éthique et spamdexing . Sciences de l'information et de la communication. 2010. dumas-01690365

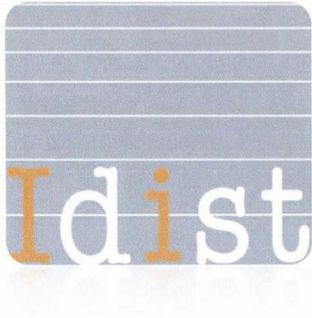
HAL Id: dumas-01690365

<https://dumas.ccsd.cnrs.fr/dumas-01690365>

Submitted on 23 Jan 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Antoine BRISSET

Master 2, Mention ICD, Parcours IDEMM
(Spécialité : Sciences de l'Information et du Document)

MEMOIRE DE STAGE

Mission effectuée du 12 avril au 24 septembre 2010

à

Axecibles

Roubaix

Référencement naturel, entre éthique et spamdexing

Sous la direction de :

Mlle M. Vanhoute

Mme M. Despres-Lonnet

Soutenu le 17 septembre 2010 à l'UFR I.D.I.S.T.

Université Charles de Gaulle, Lille 3 (Campus Pont de Bois)

BP 60 149, 59653 Villeneuve d'Ascq Cedex

Année universitaire 2009/2010

Remerciements

Avant d'entamer la rédaction de ce mémoire, je tiens à remercier vivement l'ensemble des employés du service Webmarketing de l'agence Axecibles, pour leur accueil et leur disponibilité. En particulier, Lhossen Ouhbad, pour l'aide et le soutien qu'il m'a fournis au quotidien, tant dans la mission de stage que dans la rédaction du mémoire ; Rachid Talbi, Tony Fernandes, Dany Gandziri, Jérémy Comparato et Julien Charrier pour les connaissances et autres retours d'expérience qu'ils ont partagés.

Je remercie M. Pinto et chacun des services de l'agence, pour leur collaboration active, ainsi que mes proches qui ont œuvré pour que ce mémoire soit réalisé dans de bonnes conditions.

Mon attention se tourne également vers Mme Despres Lonnet, qui a suivi le projet en tant que tutrice universitaire, et vers Mlle Vanhoute, tutrice de stage, qui a fait preuve d'écoute, de conseil et de disponibilité.

TABLE DES MATIERES

Avant-propos	p.5
Introduction	p.6
1. Le SEO, une étape fondamentale de tout projet web	p. 8
1.1. Axecibles, une web agency dynamique	p.8
1.1.1. <i>La société : genèse et croissance</i>	<i>p.8</i>
1.1.2. <i>La structure de l'agence : services et activités</i>	<i>p.10</i>
1.1.3. <i>L'offre, la cible et la stratégie de communication</i>	<i>p.13</i>
1.2. La mission : gestion d'un portefeuille clients	p.18
1.2.1. <i>Productions, refontes et mises à jour</i>	<i>p.18</i>
1.2.2. <i>Analyse de trafic et suivi de positionnement</i>	<i>p.23</i>
1.2.3. <i>Participation à la vie du service</i>	<i>p.24</i>
1.3. Le référencement naturel, au carrefour de plusieurs disciplines	p.25
1.3.1. <i>Qu'est ce que le SEO ?</i>	<i>p.26</i>
1.3.2. <i>Les SEO, une intégration de l'amont à l'aval d'un projet web</i>	<i>p.27</i>
1.3.3. <i>La nécessité d'une veille permanente</i>	<i>p.36</i>
2. Les techniques black hat : pourquoi et comment ?	p.39
2.1. La chasse au spam	p.39
2.1.1. <i>E-business et diktat de la performance : comment être et rester visible</i>	<i>p.39</i>
2.1.2. <i>En quoi consiste le spam de moteur de recherche ?</i>	<i>p.41</i>
2.1.3. <i>Guidelines et consignes aux webmasters</i>	<i>p.43</i>
2.2. Utiliser le « black hat » : dans quelles occasions et à quels risques	p.48
2.2.1. <i>Secteurs concurrentiels</i>	<i>p.48</i>
2.2.2. <i>Sites MFA</i>	<i>p.52</i>
2.2.3. <i>Filtres et pénalités</i>	<i>p.55</i>
2.3. Les techniques avancées « black hat »	p.59
2.3.1. <i>Automatiser les processus</i>	<i>p.59</i>
2.3.2. <i>Techniques et manipulations diverses</i>	<i>p.63</i>
2.3.3. <i>Se protéger</i>	<i>p.65</i>
3. Comment doit se positionner le référenceur	p.68
3.1. Où s'arrêtent les bonnes pratiques et où commence le spam ?	p.68
3.1.1. <i>Ce qu'apportent les chapeaux noirs au SEO</i>	<i>p.68</i>

3.1.2. Bénéfices et limites des actions black hat p.71

3.1.3. Le grey hat, un entre-deux ? p.74

3.2. Google est-il responsable du spamdexing ? p.77

3.2.1. Les failles de l'algorithme du géant de Mountain View p.77

3.2.2. Optimisation d'un site : pour Google ou les internautes ? p.81

3.2.3. Dérives du référencement et negative SEO p.82

Conclusion p.86

Bibliographie p.88

Annexes p.93

Avant propos

Dans le cadre de la deuxième année de master ICD¹ parcours IDEMM², nous avons eu l'opportunité d'effectuer un stage de six mois en entreprise. L'objectif étant de se confronter à un environnement professionnel, mais également de pouvoir appréhender avec un regard différent les thématiques et autres enjeux du web de demain entrevus lors des cours et des conférences auxquelles nous avons assisté lors de la formation.

Après un premier stage chez l'annonceur, qui m'avait permis découvrir le web marketing appliqué au e-commerce, j'ai décidé pour ce stage de fin d'année de postuler en tant que webmarketeur/référenceur chez Axecibles, une web agency roubaisienne. En effet, je souhaitais approfondir ma connaissance du SEO³, discipline vers laquelle tendent mes intérêts professionnels. De plus, le caractère dynamique de l'entreprise et le rapprochement au sein d'un même pôle de plusieurs disciplines (webdesign, développement, référencement, etc.) m'ont semblé un aspect intéressant de l'entreprise, susceptible de participer à mon évolution professionnelle.

J'ai donc été accueilli au sein de l'agence Axecibles, dans le service webmarketing, entre le 12 avril 2010 et le 24 septembre 2010. La responsable m'a alors confié une mission consistant à mener les campagnes de référencement naturel et à assurer le suivi d'un ensemble de sites web conçus pour une clientèle de TPE et de PME.

¹ ICD : Information Communication Documentation

² IDEMM : Ingénierie du Document Edition et Médiation Multimédia

³ SEO : Search Engine Optimisation, c'est-à-dire référencement naturel

Introduction

Le positionnement d'un site Internet sur les moteurs de recherche est aujourd'hui un tel enjeu économique et commercial que la plupart des agences web proposent un service de référencement. Il s'agit, tout d'abord, de faire connaître le site aux moteurs de recherche, mais aussi et surtout de le « positionner » le plus haut possible dans les résultats de recherche, sur un certain nombre de requêtes engendrant un trafic plus ou moins qualifié.

A l'heure actuelle, l'index d'un moteur de recherche tel que Google comporte plus d'un trillion de pages web⁴, soit 1000 milliards d'entrées, et ce chiffre augmente de manière exponentielle chaque année. Dans cette situation, et dans un contexte économique et social où la recherche en ligne est devenue un réflexe quotidien pour les internautes, il apparaît donc crucial de pouvoir discerner le fonctionnement des moteurs de recherche et de rendre compatibles les sites web avec les exigences de ces derniers. D'autant que des études ont démontré que le comportement des internautes devant les résultats des moteurs de recherche est simple : la plupart des usagers ne consultent pas les résultats au-delà de la deuxième ou troisième page de résultats.

A la croisée du webmarketing et de la gestion de projet web, le référencement est une discipline qui exige de multiples compétences, à la fois techniques, rédactionnelles, et communicationnelles. Il repose sur une somme de petites optimisations, qui permettent à un site web de sortir des profondeurs du web pour venir, sur le long terme, se confronter à la concurrence de la première page des SERP⁵. Une concurrence parfois extrême, qui a, depuis quelques années, entraîné l'apparition de pratiques interdites par les moteurs de recherche, qualifiées dans le jargon du SEO de « black hat ». Exploitant les failles dans les algorithmes des moteurs de recherche, ces pratiques vont à l'encontre de la notion de pertinence de l'information, qui est le credo d'un moteur de recherche comme Google. L'objectif étant d'obtenir la meilleure position possible dans les SERP.

⁴ DUFFEZ, Olivier, *Google a répertorié 1000 milliards de pages web*, 25-07-2008.

<http://www.webrankinfo.com/actualites/200807-1000-milliards-de-pages-sur-le-web.htm>

⁵ SERP : Search Engine Results Pages, ou pages de résultats des moteurs de recherche

Néanmoins, à l'heure de l'e-business, la frontière est de plus en plus floue entre un référencement « white hat », soucieux d'être en accord avec les consignes qualité des moteurs de recherche et un référencement « black hat ». En effet, de nombreux référenceurs, portant l'étiquette « white hat » s'inspirent des techniques « black hat » pour améliorer le positionnement de leurs sites internet. Comment, alors, à l'heure actuelle assurer un bon positionnement aux sites internet sans avoir recours au « spam », sans tromper les internautes et les robots ? Comment, par ailleurs, doivent se situer les référenceurs face à l'existence de telles techniques ?

Pour répondre à ces questions, il conviendra tout d'abord de revenir sur la notion de référencement, puis de circonscrire avec précision la mission du stage, et le cadre dans lequel celui-ci a été effectué. Puis, nous évoquerons la question du référencement black hat, en quoi il consiste et ce qui le justifie. Enfin, nous essaierons de mettre en exergue toute la difficulté actuelle à maintenir des pratiques éthiques dans le domaine de référencement, alors que la concurrence s'accroît chaque jour sur les moteurs de recherche.

1. Le SEO, une étape fondamentale de tout projet web

Avant de revenir sur ma mission de stage et de définir précisément les concepts liés au SEO, revenons tout d'abord sur la structure qui m'a accueilli durant le stage : Axecibles.

1.1 Axecibles, une web agency dynamique

Voyons ici comment est née la société Axecibles, quel est son mode de fonctionnement et quels sont les services qu'elle propose.

1.1.1. La société : genèse et croissance

Axecibles est une entreprise créée en 2001, au moment de la « bulle Internet », c'est-à-dire en pleine période de crise du secteur de l'Internet, lorsque de grands acteurs du marché de l'époque se sont écroulés. Elle est dirigée par monsieur Jimmy Pinto, un ancien consultant qui s'est lancé avec conviction dans l'aventure Internet. L'entreprise accompagne les TPE, les PME, les associations ainsi que les artisans (plombiers, couvreurs, etc.) et professions libérales (avocats, ostéopathes, etc.) dans leur développement au travers de la mise en place d'une stratégie de communication sur le web. L'activité de l'agence est donc axée sur la création et le suivi de sites internet.

A l'origine, le groupe comptait uniquement deux salariés. Il est aujourd'hui fort de plus d'une centaine de salariés, et a pour ambition, d'ici 2012, de franchir le cap des 300 collaborateurs. Les équipes d'Axecibles sont réparties entre une partie commerciale, qui est chargée de la prospection de nouveaux clients dans les secteurs où le groupe est implanté, et une partie plus technique, sédentaire, qui s'occupe de la création, du référencement et du suivi des sites internet des clients.

Le groupe Axecibles est plutôt bien implanté sur le territoire français. Il compte en effet pas moins de dix agences, dont une en Belgique. Ces agences sont réparties dans les villes de Lille, Paris I & II, Reims, Lyon, Nantes, Rouen, Marseille, Caen (ouverte début 2010) et Bruxelles. Le siège social, quant à lui, se trouve à Roubaix, au 42, rue du Général Sarrail.

Le groupe Axecibles apparaît comme un groupe rentable, qui réalise un chiffre d'affaires excédentaire et en progression permanente depuis sa création. Voici, résumée en quelques chiffres, la croissance du groupe Axecibles.

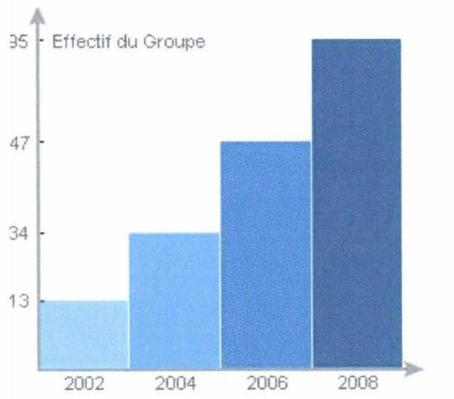


Fig.1 : Evolution du nombre de collaborateurs entre 2002 et 2008



Fig.2 : Agences Axecibles en France et en Belgique

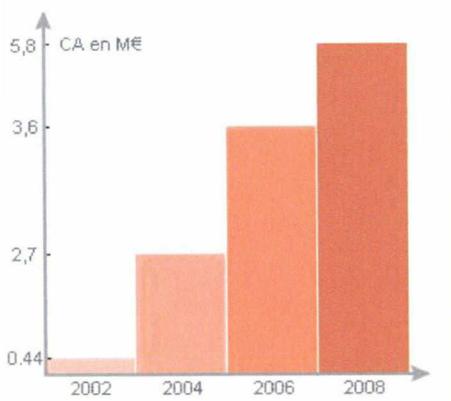


Fig.3 : Evolution du chiffre d'affaires entre 2002 et 2008

Notons que l'ambition du groupe est de réaliser 20 millions d'euros de chiffre d'affaires en 2012 avec 300 salariés et 18 agences sur le territoire. Ces quelques chiffres évocateurs ont permis de dessiner le profil de l'agence. Voyons maintenant comment celle-ci est structurée et comment est organisé chaque service.

1.1.2. La structure de l'agence : services et activités

Le groupe Axecibles dispose d'équipes multi-compétentes qui travaillent en collaboration pour assurer la bonne gestion de chaque projet web. Il convient ici d'identifier le rôle de chacun dans le processus de création du site web.

- **Le service commercial** : le service commercial est placé sous la direction de M. Didier Fiedler et est chargé de la prospection de nouveaux clients pour Axecibles. Les commerciaux sont la clé de voûte de l'organisation puisqu'ils sont à l'origine de la signature des contrats et donc du processus de production. Chaque jour, ils développent un argumentaire précis, en indiquant aux clients qu'ils disposent de solutions clé en main pour développer leur visibilité sur le web, et par là même développer leur chiffre d'affaires. Leur emploi du temps est divisé entre les jours passés sur le « terrain » et les jours passés en agence. Les lundis et mercredis, en agence, la journée se déroule suivant ce schéma :
 - Training : il s'agit de revenir sur les succès et les contre performances de la semaine écoulée
 - Prépa-phoning : recherche des prospects, préparation de l'argumentaire
 - Phoning : contact téléphonique avec le futur client pour décrocher un rendez-vous
 - Réunion pour suivre l'évolution des prises de rendez-vous
 - Préparation des rendez-vous : arguments à mettre en avant, validation du rendez-vous, etc.

- **Le service administration des ventes** : avec à sa tête Laurence D., ce service reçoit le dossier de chaque client lorsque le contrat a été signé par le service commercial. Il assure

également l'accueil téléphonique des clients, et la prise de rendez-vous entre le client et le service cahier des charges. Par ailleurs, il s'occupe de la vérification des informations bancaires de l'entreprise avant transmission au service comptabilité.

- **Le service cahier des charges** : l'équipe du cahier des charges recueille les besoins des clients, leurs souhaits en matière graphique, l'objectif commercial du site, le nom de domaine à réserver, etc. Le service est également chargé de fournir une maquette de la page d'accueil et de l'arborescence du site, documents qui seront ensuite transmis au studio graphique. Une fois la charte créée, ils prennent contact avec les clients pour la livraison du site. Le service est sous la responsabilité de Jimmy Pinto.
- **Le service administration technique** : sous la direction de Stéphanie D., ce service est chargé d'enregistrer les noms de domaine, de procéder aux éventuels transferts, de procéder à la facturation des noms de domaines qui ne sont pas renouvelés.
- **Le service hébergement** : il est placé sous la direction de Sylvain D. et est chargé de fournir aux clients une disponibilité maximale des serveurs pour que leur site web soit accessible à 99,99% du temps. L'agence compte plus de 30 serveurs, avec des serveurs « relais », qui sont utilisés lorsqu'un serveur tombe en panne. Le service est également chargé de gérer l'allocation des ressources pour veiller à maintenir une performance et un coût optimaux.
- **Le service studio** : le studio graphique, sous la direction de David O., est composé d'une équipe de webdesigners, qui, à partir des informations recueillies dans le cahier des charges, vont procéder à la création de la charte graphique. L'objectif est de créer une réelle identité visuelle, qui réponde aux besoins des clients et qui soit conforme à l'ergonomie et aux standards du web. Chaque charte graphique doit être validée par le responsable avant de pouvoir passer à la phase d'intégration, qui consiste au montage du site et à son codage en HTML et CSS.
- **Le service recherche et développement** : placé sous la responsabilité conjointe de Matthias C. et de Joseph D., le service de « R&D » est chargé de réaliser les développements spécifiques demandés par les clients. Le service a développé le back-

office des sites des clients depuis lesquels ces derniers peuvent interagir avec leur base de données, pour, par exemple, ajouter des fiches produits, rédiger une actualité, etc. Par ailleurs, le service recherche s'occupe du développement du réseau intranet de l'agence et développe régulièrement de nouvelles fonctionnalités pour améliorer les performances de chaque service. Par exemple, pour le pôle webmarketing, un outil de rapport de positionnement semi automatisé a été créé.

- **Le service webmarketing** : sous la direction de Marion Vanhoute, le service webmarketing se place en bout de chaîne dans le processus de production. C'est dans ce service que j'ai effectué mon stage. L'équipe webmarketing effectue différentes missions :
 - **Le référencement naturel** : il s'agit de l'activité principale du service. Les chargés de référencement pilotent la stratégie de référencement naturel des sites web des clients, en veillant à ce qu'elle réponde aux objectifs commerciaux mentionnés dans le cahier des charges. Il faut faire en sorte que le trafic soit qualifié et qu'il génère des leads⁶, voire des ventes.
 - **La gestion de campagnes de liens sponsorisés** : le service est susceptible de proposer aux clients ou de répondre à leurs demandes concernant des campagnes de liens sponsorisés. La tâche consiste alors, principalement, en l'achat de mots-clés, en l'optimisation des enchères et du taux de conversion sur chaque mot-clé.
 - **L'affiliation** : les membres de l'équipe peuvent être amenés, dans des cas plutôt rares, à lancer une campagne d'affiliation et à l'optimiser.
 - **Le conseil** : chaque référenceur peut également être amené à donner des conseils aux clients sur la manière dont ils peuvent améliorer leur référencement, que ce soit en travaillant leurs contenus ou en étoffant leur base de données.

- **Le service suivi clients** : il est partagé entre le suivi sédentaire, sous la responsabilité de Déborah H. et le suivi terrain, placé sous l'autorité de Moïse L.

⁶ Lead : Un lead est une action correspondant à un objectif, fixé dans une campagne (contact, demande de devis)

- **Le suivi sédentaire** : il est chargé de recueillir les demandes particulières des clients, de les accompagner dans la création de leurs contenus, et plus généralement de veiller à ce que la relation avec le client soit la meilleure possible. Le suivi sédentaire a également pour tâche de répondre aux demandes des clients concernant des modifications sur leur site et de transmettre l'information aux services concernés. Il doit également faire face aux mécontentements, voire aux contentieux et savoir rassurer les clients quant aux performances de leurs sites. Lorsqu'une fin de contrat survient, après 36 ou 48 mois, ils sont également amenés à convaincre le client de renouveler.
- **Le suivi terrain** : il s'agit d'une équipe qui se déplace dans la France entière chaque semaine pour renouveler des contrats, mais également pour livrer des modifications, proposer de nouveaux services, etc.

A côté de tous ces services, nous aurions pu en citer d'autres qui n'entrent pas directement dans le processus de production, mais qui jouent néanmoins un rôle important :

- **Le service marketing** : il est chargé d'optimiser l'image de l'entreprise, en assurant sa promotion sur le canal web ainsi que dans la presse traditionnelle. Le service marketing se charge également de la rédaction des newsletters envoyées aux clients, et de la communication en interne autour des événements comme les séminaires, les partenariats et tout ce qui touche de près ou de loin à la vie de l'agence.
- **Le service formation** : il est chargé de familiariser les partenaires à l'usage de l'informatique et de l'internet, en leur expliquant comment administrer leur base de données, comment utiliser leur messagerie électronique, comment utiliser les outils qui leur sont fournis comme la solution mailing Performail.

1.1.3. L'offre, la cible, et la stratégie de communication

Axecibles développe pour son parc clients des solutions « clé en main », qui répondent aux exigences de chacun et sont le plus fidèles possibles à leur identité commerciale. L'offre globale comprend à la fois la création du site, son hébergement, son référencement, la

rédaction de contenu (optionnel), une solution d'envoi de newsletter ainsi qu'une formation aux outils informatiques, afin de faciliter la tâche à des clients, qui, pour la plupart, sont des novices en matière de nouvelles technologies de l'information et de la communication.

Les solutions développées par Axecibles ont été regroupées en trois catégories :

Le pack visibilité : il s'agit ici de la solution développée pour les entreprises qui désirent disposer d'un simple site vitrine, avec quelques pages permettant de présenter leurs activités, ainsi qu'un formulaire de contact pour recevoir les demandes de prospects. Le pack visibilité est adapté aux petites et moyennes entreprises désirent faire leurs premiers pas sur le web, avec, en quelque sorte, une carte de visite « virtuelle ».



Fig.4 : exemple de site dit « vitrine »

Le pack performance : le pack performance offre davantage de souplesse vis-à-vis du client. En effet, il propose l'accès à une console d'administration, un « back office », qui permet aux

clients d'interagir avec une base de données et d'ajouter depuis chez eux de nouveaux produits, de nouvelles photos, des actualités, etc. Dans ce pack sont proposées diverses fonctionnalités comme la gestion d'alertes, l'installation d'un forum, d'un formulaire de devis en ligne, etc. Il s'agit



alors d'un site « catalogue », qui a pour objectifs de mettre certains produits en avant et de générer des demandes de contact par rapport à ces produits.



Fig.5 : exemple de site catalogue

Le pack e-commerce : il s'agit ici de mettre à disposition des clients un site marchand, sur mesure. Il est adapté aux entreprises qui souhaitent se lancer dans la vente en ligne d'une gamme de produits, mais également à toutes les boutiques qui souhaitent doubler leur activité « physique », par une activité de vente sur le net. Les sites orientés e-commerce sont donc pourvus de diverses fonctionnalités :



- Gestion des stocks
- Module de paiement sécurisé
- Gestion du catalogue et des fiches produits
- Suivi des commandes

L'objectif de ce type de ce site est de générer un maximum de ventes en ligne. Notons également que c'est pour ce type de site que les campagnes de liens sponsorisés sont les plus fréquentes.

Fig.6 : exemple de site dit « e-commerce »

Les entreprises qui constituent le cœur de cible de l'entreprise Axecibles sont les Très Petites Entreprises (TPE) ainsi que les Petites et Moyennes Entreprises (PME). Le groupe a d'ailleurs été spécialement créé pour répondre aux besoins de cette catégorie d'entreprises, qui ne disposent pas des moyens et des ressources internes pour mener à bien une stratégie de communication sur le média web. Le positionnement commercial de l'entreprise, univoque, est une des clés de sa réussite et de son développement. Durant mon stage, j'ai ainsi procédé au référencement de :

- environ 10 sites d'avocat
- 1 site de diagnostic immobilier
- 1 site de couvreur

- 1 site de location de voiture avec chauffeur
- 1 site de camping
- 1 site d'assainissement des eaux
- 1 site de salon de coiffure
- 2 sites de métallerie et chaudronnerie
- 1 site de boutique bio
- Etc.

Enfin, il est nécessaire d'évoquer comment l'entreprise façonne son image de marque, tant sur le web, qu'au travers des autres médias (journaux, télévision, etc.). Tout d'abord, Axecibles est à l'origine d'un ensemble d'initiatives importantes :

- Le **sponsoring** : Axecibles s'implique dans le sponsoring et a déjà développé de nombreux partenariats comme avec Carole Montillet et Sindiely Wade, dans le cadre du Rallye des Gazelles 2009 (voir photo), ou encore de manière continue avec Mathias Canci, champion de France de saut en hauteur en 2009.



- Le **E-trophée** : il s'agit d'un concours de soutien à la création et au développement d'entreprises sur Internet. Ce concours est organisé par Axecibles et récompense chaque année à hauteur de 120 000 euros deux lauréats ayant su présenter un projet original et

novateur. Les lauréats bénéficient également de la mise en place d'un site interactif et des prestations connexes (hébergement, formation, référencement, etc.)

De plus, une équipe « marketing » au sein même de l'entreprise, s'occupe à temps plein de la promotion de la société sur différents supports de communications. Elle est chargée de :

- l'animation de différents sites qui ont été lancés autour de l'activité principale d'Axecibles (www.evenementiel-axecibles.fr, www.recrutement.axecibles.com)
- relayer auprès des agences de presse les événements qui constituent l'actualité du groupe
- la diffusion de communiqués de presse

Enfin, le Président Général du groupe, Jimmy Pinto, répond régulièrement aux sollicitations des médias, comme récemment sur la chaîne BFM TV⁷.

1.2. Ma mission au sein de l'agence

Après avoir présenté l'entreprise, il est logique de s'attarder quelque peu sur la mission que j'ai effectuée et d'évoquer plus en détail en les différents projets auxquels j'ai été rattaché.

1.2.1. Productions, refontes et mises à jour

Au sein de l'agence Axecibles, dans le service webmarketing, nous distinguons trois niveaux de travail :

- **La production** : il s'agit du processus de référencement appliqué aux acquisitions de contrat, lorsqu'un nouveau client s'engage auprès d'Axecibles. Expliquons brièvement comment se déroule le processus (certaines notions techniques seront explicitées plus loin dans ce mémoire) :

⁷ Interview du 24 mai 2010, disponible ici <http://www.axecibles.com/bfm-tv-v-5.html>

- La rédaction de l'objectif : il s'agit d'identifier le partenaire, de délimiter sa zone de chalandise, de circonscrire son activité et de déterminer ce qu'il souhaite mettre en avant ainsi que la clientèle qu'il cible. Ces indications vont être précieuses dans la façon dont chaque référenceur va orienter le référencement.
- Le contrôle du site : le référenceur procède d'abord à un contrôle du contenu en s'assurant que celui-ci n'est pas en partie ou complètement dupliqué. Dans son centre d'aide pour les webmasters⁸, Google explique en effet que si le contenu est « délibérément dupliqué », dans l'objectif de « manipuler » les classements, il procédera à des « ajustements » dans le classement du site. C'est pour cette raison, d'ailleurs, qu'il conseille de rédiger des balises « title » et « description » uniques pour chaque page d'un site. Google possède un serveur dédié entièrement à l'analyse du contenu dupliqué, le **DupServer** : il est capable de déterminer quelle est la version canonique d'un document. Cette version sera ainsi privilégiée dans les résultats de recherche tandis que la seconde sera filtrée, déclassée, voire reversée dans l'index secondaire de Google. Il existe plusieurs outils en ligne permettant de contrôler que le contenu n'est pas dupliqué, tels que Copyscape⁹. Lorsque le site d'un client présente une forte proportion de contenu dupliqué, les pages concernées sont désindexées. Si le contenu dupliqué est trop important, la production est bloquée et le site est placé en attente de nouveaux contenus.
- L'audit de mots-clés : l'audit de mots-clés est une phase extrêmement importante. Il s'agit de choisir les mots-clés sur lesquels vont s'appuyer les actions du processus de référencement. Chaque page est ainsi étudiée de façon à faire ressortir les mots-clés présents. Ensuite, nous utilisons le générateur de mots-clés Google¹⁰ qui va nous donner des indications sur le volume de recherche mensuel, ou global, par rapport à chaque mot-clé, et sur la concurrence. Cependant, la concurrence se rapporte aux annonceurs Adwords qui enchérissent

⁸ Voir cette page <http://www.google.com/support/webmasters/bin/answer.py?hl=fr&answer=66359>

⁹ Disponible ici <http://www.copyscape.com/>

¹⁰ Disponible ici :

https://adwords.google.fr/o/Targeting/Explorer?_u=1000000000&_c=1000000000&ideaRequestType=KEYWORD_IDEAS

sur ces mots-clés et non pas au référencement naturel. Il est donc préférable de consulter le nombre de résultats de recherche pour ce mot-clé, et d'analyser si les premiers résultats font ressortir des pages très optimisées ou non. Le choix des mots-clés est donc un compromis entre un volume de recherche important, une concurrence moyenne et une expression pertinente, susceptible d'apporter du trafic qualifié sur le site. Nous reviendrons sur cette étape primordiale dans la suite du mémoire.

- La rédaction des balises title, meta description et meta keywords : à partir des mots-clés sélectionnés lors de la phase d'audit, le référenceur rédige les balises title et meta. La balise title est un élément de haute importance pour le référencement naturel. Bien rédigée, une balise title équivaut à 50% du travail de référencement. Chez Axecibles, nous nous limitons à 70 caractères. La balise meta description doit, elle, ne pas dépasser les 200 caractères tandis que la balise keywords contient au maximum 5 mots-clés. Selon les dires de Google, les balises meta description et keywords ne sont plus prises en compte dans le système de classement des pages. Pour Bing et Yahoo, les conclusions sont sensiblement les mêmes, même s'il semblerait que Yahoo utilise encore les balise meta keywords et description comme facteur de positionnement, si l'on se réfère à un test paru sur le laboratoire d'Oseox en 2009¹¹.
- L'optimisation on-page : il s'agit du travail d'optimisation effectué directement sur le code source du site, au moyen du logiciel Dreamweaver. Après avoir ajouté les balises title et meta, plusieurs petites opérations se succèdent :
 - L'optimisation du footer sur les mots-clés et la géolocalisation retenus
 - La modification des balises <h1> à <hn>. Il est conseillé pour le <h1> de reprendre le contenu de la balise title en l'adaptant quelque peu : usage de synonymes, déclinaisons, etc.
 - Le maillage interne, c'est-à-dire la création au sein même du contenu de liens internes pointant vers les pages profondes du site

¹¹BARDON, Aurélien, *Test de la balise meta description*, 27-08-2009: <http://www.laboratoire-referencement.fr/balise-meta-description.php>

- L'ajout de balises strong sur les mots-clés à mettre en valeur
- L'optimisation des images, avec entre autres le renommage des fichiers de façon à correspondre au référencement du site et le remplissage de l'attribut alt, utilisé pour le positionnement des images dans des moteurs verticaux comme Google images
- Le renommage des pages et les redirections des dites pages via .htaccess si elles ont été déjà indexées dans les moteurs de recherche
- La redirection permanente de la page index vers la racine pour contourner le problème de duplicate content interne.
- La création du fichier robots.txt, placé à la racine du site et qui donne des indications aux robots quant à la façon d'indexer le site.

Après avoir effectué plusieurs contrôles (présence de liens morts conduisant à un en-tête HTTP 404 synonyme de fichier introuvable, site valide W3C, affichage correct dans tous les navigateurs, etc), vient la mise en ligne du site. De nouveaux contrôles sont effectués, notamment au niveau de la densité des mots-clés. Pour chaque page, la règle dans l'agence est que l'indice de densité de mots-clés (IDM) ne doit pas dépasser les 8%. Puis, après avoir créé le fichier sitemap¹², vient la phase de soumission.

- Inscription sur les outils Google : le site est inscrit sur Google Webmasters Tools, un outil dédié aux webmasters pour vérifier l'état d'indexation du site, consulter des statistiques de visite, paramétrer la région ciblée, le domaine favori, etc. Il est également inscrit sur Google Adresses, un service réservé aux entreprises leur permettant de créer une fiche avec leurs coordonnées et diverses informations, et qui sera accessible depuis les résultats de recherche sur certaines requêtes géolocalisées. Google peut également utiliser l'adresse IP du visiteur pour lui proposer des fiches Google Adresses correspondant à sa position géographique.

¹² Sitemap : le fichier sitemap, au format XML, est issu d'un protocole mis en place par Google, puis par Yahoo et Bing afin de faciliter l'indexation des pages web

- Soumissions aux moteurs de recherche : lors de cette étape, le site est soumis aux principaux moteurs de recherche utilisés en France et dans le monde, c'est-à-dire Google, Yahoo et Bing.
- Soumissions aux annuaires : les annuaires permettent, facilement, de créer des liens entrants vers son site. Ils peuvent également parfois s'avérer de bons sites référents. Les annuaires sont choisis en fonction de leur qualité, sur la base de nombreux critères : valeur du PageRank, présence d'un lien en dur vers le site soumis, exigence d'une description unique, etc. Les référenceurs réalisent donc plusieurs soumissions, en variant à chaque fois le texte de description et en choisissant au mieux la catégorie à rattacher au site. Ils soumettent à la fois dans des annuaires généralistes, thématiques et spécialisés. En effet, un lien obtenu depuis un site de même thématique a plus de poids qu'un lien depuis un site sans aucun rapport avec le contenu du site.
- La refonte : la refonte d'un site intervient lorsqu'un contrat est arrivé à échéance et est renouvelé. La charte graphique est donc rafraîchie, de nouveaux contenus et de nouvelles pages sont parfois insérés selon le souhait du client. Dans ce cas, la procédure de référencement est plus ou moins la même que pour une simple production, à la différence près que le référenceur va étudier les statistiques du site avant la refonte et identifier les pages bien positionnées, qui amènent du trafic et qui ne doivent donc pas être ré-optimisées, sous peine de subir une baisse de classement dans les moteurs. A l'inverse les pages qui n'engendrent que peu de visites vont être modifiées.
- La mise à jour : il s'agit ici de la procédure la moins longue, puisqu'il s'agit dans la plupart des cas de référencer une ou plusieurs nouvelles pages, d'acquérir une nouvelle position demandée par le client, ou encore d'effectuer davantage de soumissions pour renforcer le linkbuilding du site.

Un autre aspect de la mission de stage a consisté à effectuer un suivi régulier des sites référencés.

1.2.2. Analyse et suivi de positionnement

Chaque référenceur dispose de son propre portefeuille de sites, pour lesquels il effectue le référencement mais également le suivi de performances. Dès lors que les actions de référencement sont terminées, son rôle est donc de suivre de près les statistiques de visite du site afin de vérifier et d'analyser la part de trafic issue des moteurs de recherche. En un mot, vérifier l'efficacité du référencement naturel. Pour cela le service référencement dispose de l'outil AWStats, un logiciel libre d'analyse de trafic.

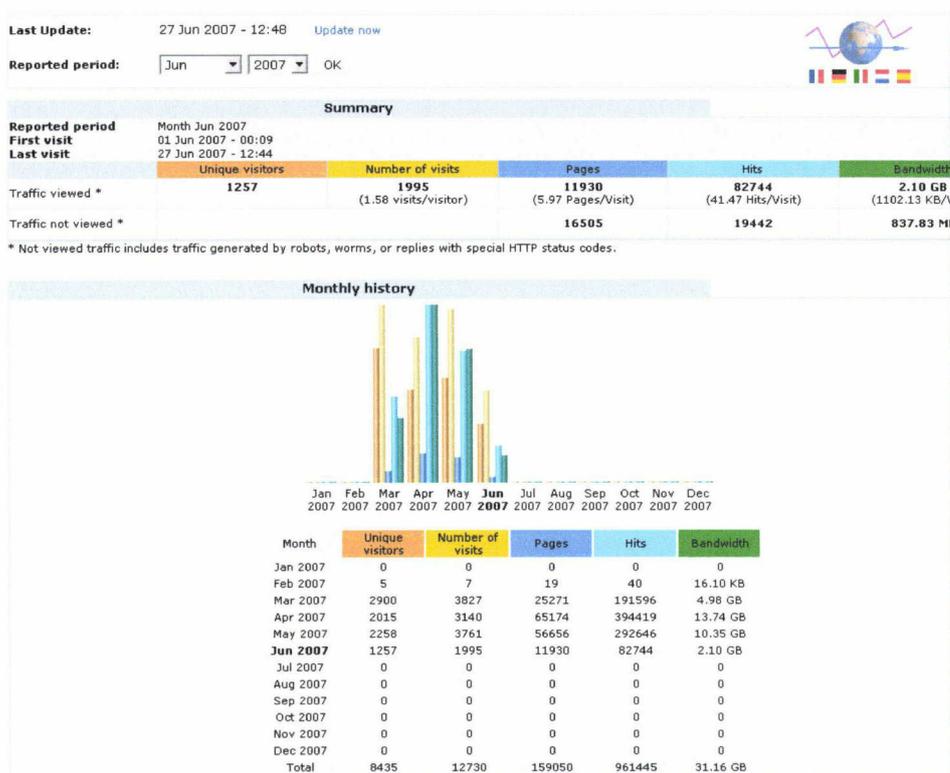


Fig.7: interface de l'outil AWStats

Ce logiciel offre plusieurs opportunités d'analyse au référenceur. Il permet tout d'abord d'avoir une vue sur le nombre de visites et de visiteurs uniques. Rappelons que le nombre de visiteurs uniques représente « le nombre de visiteurs d'un site Web non dupliqués (comptabilisés une seule fois) sur une période de temps donnée ». Cela signifie qu'une fois qu'un visiteur a ouvert une session sur un site, un cookie ou fichier témoin est déposé sur son ordinateur : il est comptabilisé comme visiteur pour la première fois. Tant que

l'internaute n'a pas effacé son cookie, et s'il effectue d'autres visites sur la période donnée, il est donc considéré par l'outil comme le même visiteur. Le nombre de visiteurs uniques nous indique donc si le site commence à se faire connaître (valeur haute), ou si ce sont souvent les mêmes visiteurs qui reviennent (valeur basse).

Par ailleurs, dans AWStats, nous pouvons avoir un aperçu sur les sites référents, c'est-à-dire les sites depuis lesquels les internautes sont arrivés, via un lien hypertexte. Cela permet notamment de voir quelle est la valeur ajoutée de certains annuaires, ou de découvrir si des internautes ont créé spontanément des liens depuis leur site vers le nôtre.

Les pages d'entrée et pages de sortie sont étudiées en particulier lors d'une refonte, pour identifier quelles sont les pages qui sont les mieux référencées sur les moteurs de recherche. La page d'entrée est celle par laquelle l'internaute a commencé sa navigation sur un site. Les pages qui ont cumulé le plus d'entrées sont performantes en termes de positionnement et ne doivent donc pas être modifiées. En règle générale, c'est la page d'index qui observe le plus grand nombre d'entrées.

Enfin, la partie la plus intéressante est relative aux mots et expressions clés saisis dans les moteurs de recherche et qui ont engendré des visites. C'est dans cet encart que le référenceur juge de l'efficacité des mots-clés choisis pour le référencement et peut également découvrir toute la puissance de la longue traîne, au travers des multiples expressions qui ont engendré peu de visites individuellement mais qui, cumulées, en représentent souvent une part très importante, proche des 70%.

1.2.3. Participation à la vie du service

Au-delà du simple travail de référencement, je me suis également investi, dans le service webmarketing, dans des projets plus globaux, visant à améliorer la compétitivité de l'entreprise.

Tout d'abord, j'ai assisté aux réunions de progrès, lesquelles se tiennent environ tous les mois et ont pour objectif de faire le point sur la production, les projets en cours, les relations

avec les autres services, etc. Les réunions de progrès sont également l'occasion d'aborder une actualité du SEO, de présenter les résultats de tests menés en interne, ou tout simplement de présenter des tutoriels à toute l'équipe, de manière à rendre le service plus autonome. Au cours des deux réunions auxquelles j'ai assisté, j'ai effectué deux présentations, l'une sur l'optimisation des balises <h1> et l'autre sur l'utilisation des ancres internes et sur leur intérêt pour le référencement.

Par ailleurs, la responsable m'a proposé, en parallèle de ma mission principale, des missions annexes, qui ont permis de m'intégrer encore plus sérieusement dans la vie du service :

- La formation, de manière accélérée, d'une stagiaire web rédactrice à la pratique du référencement, aux outils et aux stratégies utilisés par Axecibles.
- Le suivi d'un stagiaire chargé de préparer les rapports de positionnement délivrés aux clients. Ce suivi visait notamment à l'orienter sur la manière dont il était possible de conseiller les clients et de les impliquer dans la démarche de référencement.
- La participation à un projet de création d'annuaire par Axecibles, et donc à la mise en place de recommandations techniques avant la rédaction d'un cahier des charges.

Ces quelques informations sur l'entreprise ont permis de préciser le contexte économique et social dans lequel j'ai évolué pendant ce stage. Il convient maintenant d'aborder plus concrètement le thème principal de ce mémoire : le référencement naturel.

1.3. Le référencement naturel, au carrefour de plusieurs disciplines

Le référencement naturel est une discipline qui doit être envisagée au plus tôt dans un processus de création de site Internet. Par ailleurs, elle est souvent appelée à se modifier, du fait des évolutions des algorithmes des moteurs de recherche. Revenons ici sur ces spécificités.

1.3.1. Qu'est-ce que le SEO ?

Le terme de référencement doit être défini avec précision. En effet, le sens commun emploie très souvent le mot « référencement » pour parler de « positionnement » sur les moteurs de recherche. Or les deux termes sont bien à distinguer. Il faut notamment rappeler qu'au sens strict, le terme de « référencement » renvoie à toute action visant à rendre un site web présent dans les bases de données des moteurs de recherche, autrement dit « indexé ». Le référencement naturel consiste donc à soumettre un site aux moteurs de recherche. Néanmoins, aujourd'hui, par extension et/ou par abus de langage, « référencer un site » a une signification beaucoup plus large. Il comprend plusieurs étapes complémentaires :

- **l'indexation du site** dans les moteurs de recherche : il s'agit ici d'une des phases les plus importantes. En effet, si le site ne respecte pas les critères d'indexabilité définis par les moteurs, il ne pourra pas être visible dans les résultats desdits moteurs de recherche. Pour bien comprendre cette première phase, il est important de revenir sur le fonctionnement des moteurs de recherche.

Pour classer les documents qui sont stockés sur le web, les moteurs de recherche se focalisent sur un ensemble de critères dits de pertinence qui vont leur permettre de privilégier un résultat plutôt qu'un autre suite à une requête de l'internaute. Comme l'expliquent Sergey Brin et Lawrence Page dans l'article « *The Anatomy of a Large-Scale Hypertextual Web Search Engine* » paru en 1998, le moteur de recherche passe par plusieurs phases pour indexer les documents :

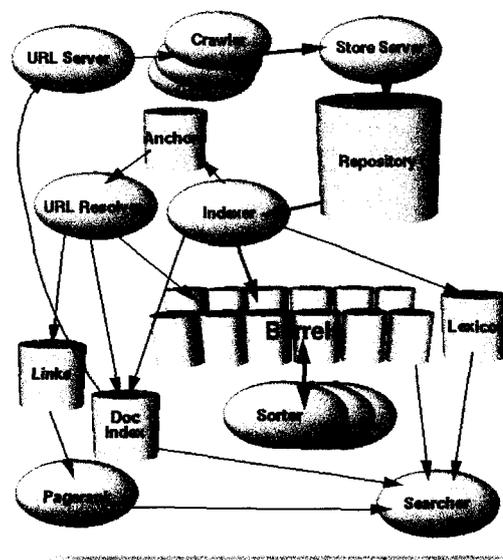


Fig.8 : schéma complexe du fonctionnement du moteur de recherche Google

- tout d'abord, par le biais de robots ou « spiders », de petits programmes qui parcourent le web de lien en lien (on appelle ce parcours le « crawl »), le moteur de recherche indexe et envoie le contenu des pages à un serveur tampon qui va mettre les données en cache
 - un indexeur va ensuite stocker les liens et leurs ancres¹³, et produire en parallèle un lexique avec les termes rencontrés pendant le crawling
 - un résolveur d'URLS va s'occuper de convertir en URL absolues les URL relatives
 - l'ensemble des documents va être versé dans un index principal, celui que les utilisateurs interrogent à travers des requêtes constituées de mots-clés.
- **le positionnement du site** dans les résultats des moteurs de recherche : le positionnement du site, pour Google par exemple, dépend d'un ensemble de plus de 200 critères dont une infime partie seulement a été dévoilée aux webmasters. Lorsqu'un internaute effectue une requête, Google va rassembler l'ensemble des documents, qui, dans son index, sont en relation avec cette requête. Puis c'est seulement après ce premier traitement qu'il va faire intervenir son algorithme, de manière à classer les documents par ordre décroissant de pertinence¹⁴. C'est la phase de ranking.

Cette phase de « ranking », qui précède l'affichage des résultats, est dépendante d'un certain nombre de critères. Ces critères, à la fois de pertinence, de popularité et d'audience vont déterminer si le site web apparaîtra dans les 10 premiers résultats, ou au mieux, dans le triangle d'or, la zone la plus visible en haut à gauche des résultats de Google, qui a été mise en lumière en 2005 par les sociétés Enquiro et Dit-it.com lors d'une étude



¹³ Ancre : Texte que le lien entoure

¹⁴ BOURRELLY, Laurent, *Le guide du référencement*, p.5, 2010

d'eye tracking. Si l'on a l'habitude de considérer que les internautes ne surfent pas au-delà des trois premières pages, il est aujourd'hui même de plus en plus évident que c'est cette zone stratégique, au dessus de la ligne de flottaison, qui est convoitée par tous les webmasters.

Dans le petit monde du SEO, il est de coutume de représenter les « leviers » de ranking dans les SERP selon une pyramide, qui recense les optimisations les plus importantes à prendre en compte lors du travail de référencement.

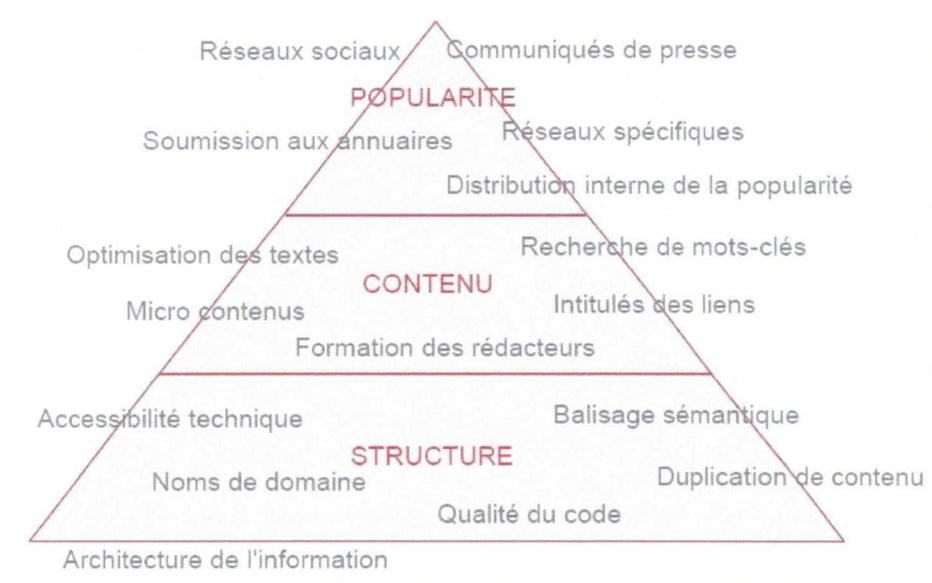


Fig.9 : la pyramide du référencement, vue par Sébastien Billard

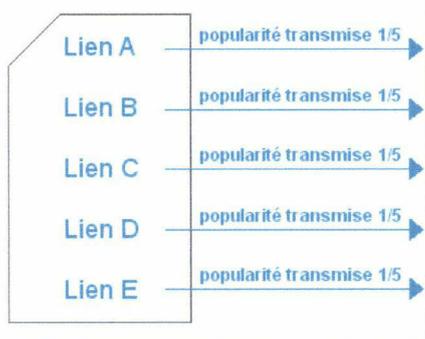
Il existe donc essentiellement 3 axes sur lesquels le référenceur va s'attarder car ils vont favoriser ou, au contraire, pénaliser le référencement et le positionnement du site.

- **La structure** : il s'agit ici tout d'abord, d'un travail sur la qualité du langage de balisage HTML et du langage de présentation CSS. Si un site est « propre » et respecte les standards définis par le W3C, il sera facilement crawlable et indexable. Par ailleurs, il faut que le site soit structuré avec des balises sémantiques, comme les balises <h1> à <hn> qui définissent la titraille du document. De même, le choix du nom de domaine, son ancienneté et la structure des URL sont des éléments auxquels

prêter une attention particulière. Ainsi, le nom de domaine doit de préférence contenir un mot-clé en rapport avec la thématique du site et l'intitulé des URL doit être parsemé de mots-clés, séparés de préférence par des tirets. Enfin, le site doit être accessible, c'est-à-dire qu'il doit ne nécessiter que le minimum de technologies pour être lu (Flash et Javascript, par exemple, sont à utiliser avec parcimonie). L'objectif est que les robots puissent lire correctement le site et interpréter les informations qui s'y trouvent.

- **Le contenu** : l'un des adages du SEO est « *the content is king* » (le contenu est roi). Cela signifie que le contenu est à la base de tout le travail de référencement. Dans le contenu scannable par les robots, il est important de mettre en valeur les mots-clés sur lesquels la page a pour objectif d'être positionnée. Cela passe par une certaine densité de ces mots-clés dans la page, mais également par la mise en valeur de ces mots-clés dans le code HTML (balise <title>, balises meta, balises , etc.)
- **La popularité** : c'est ici le travail le plus fastidieux et qui requiert un maximum d'ingéniosité. Il s'agit d'obtenir un maximum de liens de qualité depuis des sites externes vers son propre site, afin d'accroître sa popularité. Pour créer des liens, il existe plusieurs méthodes. Si le « linkbaiting » consiste à appâter les internautes en leur proposant un contenu original, susceptible d'être « linké », le « linkbuilding » exploite lui d'autres méthodes :
 - La soumission du site dans des annuaires thématiques, généralistes ou localisés
 - La diffusion de communiqués de presse
 - L'inscription du site sur des digg-like
 - L'échange de liens ou netlinking entre sites
 - Etc.

Tous ces liens font l'objet d'un traitement par les moteurs de recherche. Google, par exemple, applique aux documents web un indice de popularité appelé PageRank¹⁵, qui mesure la quantité et la qualité des liens entrants vers une page web. Sébastien Billard définit le PageRank d'une page comme « *la probabilité qu'a un surfeur aléatoire de visiter cette page* ». Le fonctionnement du PageRank est à la fois simple et complexe : à partir d'une page A, la valeur du Pagerank transmise aux pages vers



lesquelles pointent les liens de la page A est divisée par le nombre de liens présents sur cette page (voir le schéma ci-contre). Mais il est plus ou moins possible de manipuler la transmission de PageRank aux pages internes, nous y reviendrons plus tard.

Le nombre et la qualité des backlinks¹⁶ est donc, pour Google, le signe que le site est populaire et apprécié des internautes. A contenu égal et optimisation « on page » égale, un site pourra ainsi faire la différence sur un autre site, dans les SERP, par la qualité de son « linking ».

Mais la notion de popularité d'une page ne s'arrête pas au simple PageRank. La valeur d'un lien est en effet multiple. Tout d'abord, l'ancre d'un lien, c'est-à-dire le texte à l'intérieur du lien¹⁷, transmet à la page ciblée par le lien un indice qu'Olivier Andrieu¹⁸ qualifie de « réputation ». En réalité, le libellé du lien transmet aux moteurs de recherche une information importante quant à la thématique de la page visée. En créant donc des backlinks ou des liens internes avec une ancre optimisée, il est possible de positionner la page cible sur l'expression ou le mot-clé voulu.

¹⁵ PageRank : Algorithme d'analyse des liens inventé par Larry Page et utilisé par Google dans son système de classement

¹⁶ Backlink : lien entrant sur un site

¹⁷ Exemple : `Ancre`

¹⁸ ANDRIEU, Olivier, *Réussir son référencement web*, p.158

De même, la qualité du lien est aujourd'hui primordiale : un lien effectué depuis des sites de « confiance » semble être valorisé par Google dans son algorithme. Bien que cette notion n'ait jamais été clairement rendue officielle par la firme américaine, on parle aujourd'hui de « TrustRank ». Un lien aura ainsi plus de poids si :

- il provient d'un site dont le nom de domaine est ancien et dont les données Whois¹⁹, comme la durée d'enregistrement du nom de domaine, indiquent que le site va perdurer sur la toile
 - il provient d'un site avec beaucoup de pages
 - il provient d'un site sécurisé
 - il provient d'un site noté favorablement par un être humain et jugé incontournable dans une thématique donnée, par exemple Wikipedia, l'annuaire Dmoz...
 - il provient d'un site dont le Top Level Domain²⁰ est .edu, .gov ou d'autres TLD associées à des organismes officiels
 - il provient du site d'une association d'une ONG ou d'une Fédération Internationale
 - etc.
- **le suivi des positions** du site dans les résultats des moteurs de recherche : dernière étape incontournable du travail de référencement, le suivi de positionnement, et de manière plus large l'analyse du trafic, permettent de vérifier si les mots-clés retenus pour le référencement du site sont à l'origine d'un bon positionnement et s'ils apportent du trafic. Chez Axecibles, le service webmarketing dispose de l'outil Yooda SeeURank pour effectuer cette tâche : le référenceur peut y entrer les mots-clés qui l'intéressent puis le logiciel va se charger de contrôler le positionnement du site sur ces requêtes, sur les différents moteurs de recherche sélectionnés. Cela peut être utile lors d'une refonte ou pour apporter des informations aux clients sur l'état de santé de leur site internet. Il est important d'effectuer un suivi régulier, pour traquer d'éventuels problèmes : site mal

¹⁹ Whois : service Internet donnant un ensemble d'informations sur le propriétaire et l'hébergeur d'un site

²⁰ Top Level Domain : l'extension apparaissant à la fin du nom de domaine

indexé, contenu mal optimisé, manque de backlinks, etc. De même, un suivi régulier permet de découvrir les mots-clés de longue traîne qui engendrent de bonnes positions et du trafic, et ainsi de développer du contenu autour de ces nouveaux mots-clés ou tout simplement de modifier quelque peu le référencement sur la base de ces mots-clés.

1.3.2. Le SEO, en amont et en aval d'un projet web

Pour qu'une stratégie SEO soit viable et efficace, il va de soi qu'il faut penser le référencement dès le début du projet. Chez Axecibles, le référencement est omniprésent dans le processus de production et tient une place importante dans chaque service.

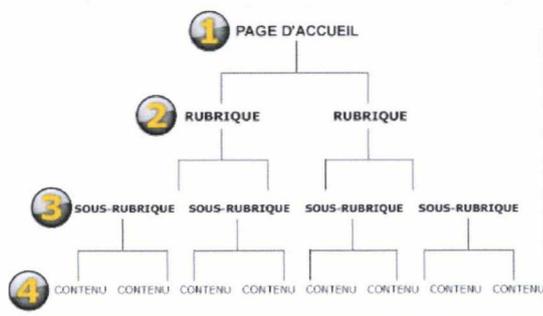
Tout d'abord, comme nous avons vu précédemment, le choix d'un nom de domaine est stratégique. Pour cela, le service webmarketing et le service administration technique peuvent orienter le client sur le choix du nom de domaine. De préférence, le nom de domaine d'un site « *doit contenir un ou plusieurs mots-clés décrivant au mieux ce qu'il propose dans ses pages* »²¹. Voici quelques exemples de noms de domaine choisis dans l'optique d'un bon référencement :

- www.metallerie-metalinox : le site présente l'activité et les produits d'une entreprise de métallerie
- www.avocat-andrieux : le site est celui d'un avocat.

Par ailleurs, Axecibles oriente ses clients vers des TLD en rapport avec la zone de chalandise désirée. Ainsi, comme la majorité des clients s'adressent à une clientèle française, le *.fr* est logiquement le plus adapté. En effet, les versions locales des moteurs de recherche favorisent des extensions locales dans leurs résultats de recherche, même s'ils ne se basent pas que sur ce critère (langue du site, pays dans lequel le site est hébergé, etc. sont également pris en compte). Ainsi une recherche sur Google.fr retournera davantage de noms de domaines en *.fr* qu'une recherche sur Google.com ou Google.ca.

²¹ ANDRIEU, Olivier, *Réussir son référencement web*, p.133

Du côté de la structure, il existe également un réel travail mené par l'équipe cahier des charges sur l'arborescence du site. Tout d'abord, il convient de regrouper au mieux les contenus, afin de mettre en place une arborescence aisément compréhensible par les



moteurs de recherche. Il faut donc cerner ce que le client souhaite mettre en avant sur son site et regrouper ses contenus selon des catégories logiques. En termes de référencement, l'arborescence a plusieurs incidences :

- Les liens des menus de navigation doivent être en « dur », c'est-à-dire codés en HTML, pour pouvoir être suivis correctement par les moteurs de recherche. S'ils sont par exemple codés dans un Javascript complexe ou en Flash, alors les moteurs de recherche ne pourront pas crawler et indexer les pages profondes du site.
- L'intitulé de ces liens doit être explicite. Si ce sont les premiers liens rencontrés par les spiders, alors l'ancre de ces liens doit être choisie judicieusement, car les moteurs de recherche attachent une valeur plus importante aux premiers liens rencontrés, ainsi qu'à leurs ancres. Par contre, nous n'avons pas réellement de réponse quant à savoir si les moteurs sont capables de distinguer les liens de navigation et les liens éditoriaux et effectuer une quelconque pondération.
- Si l'arborescence du site contient trop de pages profondes qui auraient pu être regroupées, alors cela entraîne tout simplement une division plus importante du PageRank entre toutes les pages.

La manière dont est monté le site influe donc directement sur son indexation, mais aussi sur son classement futur. Le rôle du studio graphique est donc de concevoir un site ergonomique, en accord avec les choix graphiques du client mais aussi optimisé pour le référencement. Plusieurs éléments vont donc dans ce sens :

- Les technologies Flash et Javascript sont utilisées au minimum : le fonctionnement du site ne doit pas être dépendant de ces langages, qui sont, globalement, encore incompris par les moteurs de recherche, et représentent donc un obstacle au référencement.
- Le site doit être accessible et valide selon les normes W3C, pour que la phase de crawl soit facilitée et l'indexation optimale
- Le contenu éditorial doit être la première chose vue par les moteurs. Comme les moteurs de recherche « voient » le code source « brut » sans lire la feuille de style CSS, il est possible de placer le contenu avant les autres éléments dans le code source, mais de faire en sorte que la page vue par les visiteurs présente d'abord le menu avant le contenu. Ainsi les moteurs de recherche verront d'abord les éléments éditoriaux les plus importants comme les balises sémantiques <h1> à <hn> ainsi que les mots-clés de la page, avant même les liens de navigation. Cela est possible en utilisant les propriétés de positionnement CSS « absolute » et « relative ». Concrètement voilà ce que cela donne, sur le site <http://www.avocat-cesar.com/> :

A | A+ | A+

Céline César -

Avocat Généraliste dans le Var

L'exercice de ma profession en Cabinet individuel me permet d'avoir une véritable connaissance des dossiers et des clients, en effet j'attache grande importance à l'aspect humain des dossiers en apportant à mes clients une écoute attentive.

" Je jure, comme avocat, d'exercer mes fonctions avec dignité, conscience, indépendance, probité et humanité."

Mes domaines de compétences sont les suivants : [Droit de la famille](#), [Droit des mineurs](#), [Droit pénal](#) et [Droit immobilier](#)

[» Mieux connaître Maître César](#)

Le divorce

Le rôle de l'Avocat est **primordial** dans ce domaine en effet, non seulement la représentation par Avocat est obligatoire mais surtout il peut vous **conseiller** et vous **aider**...

[» Le Divorce](#)

Le droit de la famille

Dans le cadre de ces procédures, **je saurai vous apporter écoute et conseil** afin de régler dans les meilleures conditions possibles les différends relatifs aux problèmes familiaux...

Fig.10 : page d'accueil telle qu'elle est vue par les robots (CSS et Javascripts désactivés)



Fig.11 : page d'accueil telle qu'elle est vue par les internautes

Le service référencement est donc en contact permanent avec le studio graphique pour s'assurer que ces bonnes pratiques sont correctement respectées. De même, lorsqu'il s'agit d'un site catalogue ou e-commerce, où une base de données est présente, c'est le développement du site qui, dès le départ, est orienté dans le sens d'une optimisation du référencement. En effet, les développeurs et administrateurs réseau doivent laisser la possibilité aux référenceurs :

- de mettre en place des balises meta et title dynamiques, en reprenant dynamiquement, par exemple, le nom d'une catégorie ou d'un produit, afin que chaque balise ait un contenu différent. Google recommande en effet de rédiger des balises title et méta uniques pour chaque page d'un site
- d'optimiser l'intitulé des URL grâce à l'URL Rewriting²²
- de configurer le fichier .htaccess à la racine du site de manière à pouvoir gérer les redirections serveurs
- d'améliorer la vitesse de chargement des pages, en optimisant par exemple le temps de réponse aux requêtes HTTP qui sont envoyées au serveur

²² URL Rewriting : réécriture d'URL (on utilise souvent les expressions régulières dans le .htaccess pour faire cela)

Enfin, le référencement accompagne un projet web bien après le processus de production, puisqu'il s'agit d'un travail de suivi constant. Le service suivi clients de l'agence Axecibles est ainsi souvent en contact avec le service webmarketing pour que soient effectués des « audits », c'est-à-dire des bilans complets sur « l'état de santé » et la performance des sites web. Le référenceur en charge de l'audit s'attarde ainsi sur le positionnement du site sur les mots-clés stratégiques, et sur les optimisations on page et off page qui peuvent être améliorées.

1.3.3. Une veille omniprésente

En SEO, la veille est absolument indispensable, tellement les évolutions des moteurs de recherche sont fréquentes. La souscription aux flux RSS de bloggeurs influents est donc une pratique nécessaire et la lecture de ces flux, via un agrégateur tel que netvibes, une activité qui fait partie intégrante du métier. Le SEO dispose d'une communauté assez active dans la blogosphère et la twittosphère, ce qui permet d'être rapidement informé des évolutions d'un algorithme tel que Google.

Pendant le stage, voici les événements qui ont engendré le plus de changements dans la manière d'aborder et de pratiquer le référencement :

- **Caffeine** : il s'agit de la « *nouvelle infrastructure technique de Google* »²³, mise en place début juin 2010 et qui vise à une indexation plus rapide des sites web sur la toile, afin que l'index retourne aux internautes des résultats plus « frais ». La différence tient au fait que Google crawle chaque site web de manière individuelle et l'ajoute immédiatement dans l'index alors qu'auparavant il procédait par groupe de pages.
- 
- **Mayday** : c'est le nom de code qui a été donné, Outre Atlantique, à une mise à jour de l'algorithme de Google qui vise à rendre plus pertinents les résultats des requêtes

²³ DUFFEZ, Olivier, *Google officialise son infrastructure Caffeine*, 09-06-2010.
<http://www.webrankinfo.com/dossiers/indexation/caffeine>

longues, de type « longue traîne »²⁴. A priori le maillage interne (relier les pages entre elles) ne suffit plus, un site doit pouvoir disposer de pages profondes optimisées, avec un contenu de qualité, non dupliqué et un certain nombre de backlinks. Ce sont, apparemment, les sites e-commerce qui ont le plus pâti de ces modifications, avec une baisse de trafic allant jusqu'à 20%. D'autres hypothèses ont été formulées pour expliquer ces baisses de trafic : sanctions vis-à-vis du duplicate content sur des sites dynamiques accessibles via plusieurs URL, plus d'importance donnée aux résultats de la recherche universelle²⁵, pages déclassées à cause de leur temps de chargement, etc.

- **Temps de chargement des pages** : Google a annoncé qu'il prendrait en compte le temps de chargement des pages comme un critère de pertinence supplémentaire dans son algorithme de classement. Cela signifie qu'il faut limiter au maximum tout ce qui peut freiner ou ralentir le temps d'affichage complet d'une page web, notamment en diminuant le nombre de requêtes HTTP nécessaires pour l'affichage d'une page. Cela requiert une collaboration active entre tous les services d'une agence web.

Studio { Compression, appel en bas de page et externalisation des CSS et Javascript²⁶
 Minification des feuilles CSS et des scripts Javascripts
 Utilisation des CSS sprites²⁷ et nettoyage des commentaires dans le code HTML

Administrateurs réseaux { Activer la compression GZIP des fichiers sur le serveur
 Mettre en place un système de cache²⁸
 Dissocier serveurs de pages web et serveurs de BDD²⁹

²⁴ Longue Traîne : Ensemble des expressions-clés, souvent composées de plusieurs mots-clés qui apportent, de manière cumulée, environ 70% du trafic d'un site

²⁵ Recherche universelle : concept qui consiste en l'affichage de plusieurs médias dans les résultats de recherche pour une requête (images, vidéos, actualités, etc.)

²⁶ Les fichiers CSS et JS ne doivent plus être appelés directement dans le code source

²⁷ CSS Sprites : technique permettant de regrouper les images utilisées pour le design du site en une seule image, et d'utiliser les coordonnées x et y pour placer les images dans le flux de la page

²⁸ Le système de cache serveur permet de stocker les pages PHP transformées en HTML et de les envoyer au client (navigateur, moteur de recherche, etc.) qui effectue une requête HTTP

²⁹ BDD : base de données

Comme nous venons de le décrire, le monde du référencement est sans cesse confronté à des modifications des algorithmes des moteurs de recherche. Même si les critères principaux d'optimisation, sont, globalement, les mêmes depuis quelques années, il semblerait que le leader de la recherche d'informations souhaite renforcer son modèle de pertinence et combattre le spam. L'illustration la plus éloquente de ce phénomène est l'introduction par Google, puis par ses concurrents, de l'attribut « nofollow » en 2005. Les liens en nofollow sont utilisés par les webmasters pour signifier à Google qu'ils ne doivent pas être pris en compte dans le calcul de positionnement des pages vers lesquels ils pointent, qu'ils n'ont donc aucun poids. Autrement dit, pas de transfert de PageRank ni de TrustRank, ni-même de transfert de « reputation ». L'objectif était, notamment pour Google, de limiter les commentaires abusifs sur les plateformes de blog, utilisés uniquement dans le but d'obtenir un backlink vers son site. Néanmoins, l'apparition du « nofollow » a eu un effet inattendu pour Google, puisque certains webmasters s'en sont servis, de manière détournée, pour pratiquer le PageRank Sculpting, autrement dit l'optimisation du transfert de PageRank vers les pages internes d'un même site, ce qui a conduit Google à revoir sa définition du nofollow³⁰ ...

Nous sommes alors en mesure de nous demander ce qui motive certains webmasters à utiliser sans retenue les techniques dites de spamdexing. Dans quelle mesure sont-elles nécessaires ? En quoi consistent-t-elles, quels en sont les avantages et les limites ?

³⁰ Dorénavant un lien en « nofollow » ne transfère pas de PageRank mais est bel et bien compté dans la division du PageRank entre tous les liens de la page. Il équivaut donc à une perte pure et simple de PageRank

2. Les techniques « black hat » : pourquoi et comment ?



Le concept de Black Hat n'est pas spécifique au SEO. En effet, selon wikipedia, le « terme **black hat** désigne les hackers qui ont de mauvaises intentions, contrairement aux **white hat** qui sont les hackers aux bonnes intentions ». Concrètement, ceux qui se réclament du black hat utilisent leurs compétences en matière informatique dans un but lucratif, ou pour nuire à des entreprises ou à des organisations diverses. Les dénominations « white hat » et « black hat », seraient des métaphores inspirées des westerns américains, ce qui paraît quelque peu réducteur. Dans le référencement, les « black hat » sont plus précisément ceux qui pratiquent le « spamdexing », ce qu'Olivier Andrieu considère comme une « fraude sur l'index des moteurs »³¹. Voyons donc dans quels contextes les techniques dites black hat sont utilisées, et comment les moteurs de recherche les combattent.

2.1. La chasse au spam

Le spam de moteur de recherche semble s'amplifier dans les domaines où la concurrence est rude et les gains potentiels relativement importants pour vouloir rechercher une rentabilité maximale. Analysons de plus près le contexte qui rend favorable le développement de telles pratiques.

2.1.1. E-business et diktat de la performance : comment être et rester visible ?

Aujourd'hui, obtenir une place de choix parmi les 10 premiers résultats naturels de Google est une quête quasiment vouée à l'échec si le processus de référencement n'est pas considéré avec la plus haute importance. Bien souvent, il faut du temps. Et pour passer devant des concurrents de plus en plus au fait des techniques d'optimisation et du reverse engineering, il faut beaucoup de temps. Le reverse engineering, en matière de SEO, peut être défini comme le processus visant à déterminer comment l'algorithme de classement des moteurs de recherche est construit, en analysant les résultats affichés lors d'une requête et en étudiant les facteurs qui ont pu influencer le positionnement des sites retournés. Les

³¹ ANDRIEU, Olivier, *Réussir son référencement web*, p.347

entrepreneurs ont donc bien saisi l'intérêt d'un outil comme les moteurs de recherche pour augmenter leur chiffre d'affaires et leurs profits. Le web, qui à l'origine n'était qu'un simple réseau destiné au partage de fichiers, est aujourd'hui un canal de promotion devenu incontournable et il est l'objet d'un véritable business model où l'enjeu est d'arriver au plus vite à la performance maximale. Chaque site web un tant soit peu commercial a pour objectif de réaliser le maximum de conversions. Par conversion, nous entendons généralement le processus de transformation d'un simple visiteur d'un site web en acteur, que ce soit pour une inscription à une newsletter, un achat ou un simple contact par mail. Cet aspect purement marketing est directement lié au SEO. En effet, pour pouvoir augmenter son taux de conversion, il faut pouvoir drainer du trafic. Or, une source majeure d'obtention de trafic passe par les moteurs de recherche.

Cependant, les sociétés qui misent sur le web pour développer leur activité n'imaginent pas quelles sont les contraintes actuelles pour positionner un site sur la première page de Google, sur une requête concurrentielle. Beaucoup d'agences de référencement sont ainsi confrontées à des clients impatients, qui souhaitent un positionnement quasi immédiat et qui pensent que le référencement relève parfois de la magie. Sur le blog Axe-Net, l'article « *Crédit, sexe, viagra, poker, soyez patients !* »³² illustre parfaitement cette situation. Comme l'auteur de l'article l'explique, « sur Google, la concurrence va être bien plus importante que dans la vraie vie ». En effet, sur les thématiques susceptibles de rapporter beaucoup d'argent, la concurrence est rude. Dans les résultats de recherche se mêlent des sites d'information, des sites vitrines, de grands portails, les sites plus « officiels » tels que Wikipedia, etc.

Sans un travail continu et sans engranger un maximum de backlinks de manière régulière, prendre la première place est chose compromise. Comment, par exemple, être visible sur une requête aussi concurrentielle que « sac à main », qui retourne sous Google plus de six millions de résultats ? C'est ici qu'intervient le spam de moteur de recherche, encore appelé black hat SEO.

³² PEYRONNET, Sylvain, *Crédit, sexe, viagra, poker, soyez patients !*, 2010. <http://blog.axe-net.fr/credit-viagra-sexe-poker-soyez-patient/>

2.1.2. En quoi consiste le spam de moteur de recherche ?

Le spam de moteur de recherche ou « spamdexing » est, selon Wikipedia, un « ensemble de techniques consistant à tromper les moteurs de recherche sur la qualité d'une page ou d'un site afin d'obtenir, pour un mot-clef donné, un bon classement dans les résultats des moteurs ». Bien souvent, le spamdexing passe par l'emploi de techniques qui visent à présenter aux robots de Google et des autres moteurs un contenu et une structure différentes de ce qui est réellement visible par les internautes. Nous y reviendrons plus loin dans ce mémoire.

En réalité, nous pouvons observer que les techniques de spamdexing ont évolué avec le temps : Google et les autres moteurs de recherche se sont efforcés de réagir de manière efficace face à l'apparition de telles techniques. Olivier Andrieu donne ainsi l'exemple des balises meta³³, dont le poids s'est considérablement amoindri au fil des années, jusqu'à devenir pratiquement nul aujourd'hui. Les moteurs ont en effet sanctionné les abus des webmasters qui utilisaient ces balises et notamment la balise meta keywords pour les truffer de mots-clés, parfois sans aucun rapport avec le site, uniquement dans le but de drainer un trafic important. Il en est de même des pages satellites. Le site www.blackhatseo.fr³⁴ définit



la page satellite comme une page optimisée sur de nombreux mots-clés répétés sur toute la page, et qui, lorsqu'elle est chargée par l'internaute, est automatiquement redirigée vers la « vraie » page d'accueil du site, via un script Javascript ou une balise meta refresh. La page d'accueil ne nécessitant donc aucune optimisation. Par contre, les robots qui arrivent sur cette page satellite, également connue sous le

³³ ANDRIEU, Olivier, *Réussir son référencement web*, p.22

³⁴ Définition tirée du site <http://www.blackhatseo.fr/?page-satellite>

nom de page alias vont bel et bien scanner son contenu. Si, en parallèle, cette page bénéficie d'un certain nombre de liens entrants, alors elle a toutes les chances de se voir bien positionnée dans les résultats des moteurs de recherche.

La position de Google sur les pages satellites s'est longtemps révélée floue. En effet, elle peut être utilisée dans un cadre éthique. Elle peut par exemple constituer une solution pour détecter la langue d'un navigateur et donc proposer un contenu dans cette langue. Cependant, une fois encore, cette pratique ayant engendré une quantité faramineuse d'abus du côté des propriétaires de sites web, Google comme d'autres ont fini par tout simplement supprimer ces pages de leur index.

Revenons à présent sur l'ambiguïté du système judiciaire vis-à-vis de ces pratiques. De fait, ces techniques, interdites par les moteurs de recherche, constituent des infractions juridiques, d'après la loi sur les Systèmes de Traitement Automatique de Données³⁵. Elles sont réprimées par Google, et font parfois l'objet de procédures judiciaires pour plusieurs raisons :

- tout d'abord, le modèle économique de Google est entièrement fondé sur la pertinence des informations qu'il délivre aux internautes. En poussant dans les résultats de recherche des sites ayant utilisé des techniques black hat, et qui parfois ne répondent pas à la recherche de l'internaute, il s'expose à la fuite des internautes vers un moteur de recherche concurrent, et par là même à la remise en question de ses régies de publicité Adwords et Adsense, depuis lesquelles il tire la grande majorité de ses revenus. Il a donc tout intérêt à partir à la « chasse au spam ».
- s'il nuit au modèle économique des leaders de la recherche d'informations en ligne, le spamdexing nuit également aux internautes, qui se voient proposer des sites qui ne correspondent pas forcément à leurs attentes ou auxquels ils attribuent un « faux » crédit, basé sur un classement trompeur dans les moteurs de recherche. Ils peuvent entamer des procédures judiciaires pour « publicité trompeuse », lorsque par exemple,

³⁵ DIMEGLIO, Arnaud, *Le droit du spamdexing*, 27-01-2004.
<http://www.journaldunet.com/juridique/juridique040127.shtml>

un site marchand, bien classé et donc considéré comme pertinent et de confiance par les utilisateurs, a utilisé des techniques de référencement border-line.

- le spamdexing, s'il peut nuire aux internautes, peut également nuire aux propriétaires de sites web, dont les sites passent derrière ceux de sites ayant eu recours à des actions de spamdexing. Il s'agit alors de concurrence déloyale, réprimée par la loi. Le cas le plus typique, qui n'est pas à proprement relatif au SEO, est le typosquatting. Il s'agit d'enregistrer un nom de domaine dont l'orthographe est extrêmement proche de celle d'un site concurrent, le plus souvent avec une faute d'orthographe, ou une lettre en moins. L'objectif est de canaliser vers son site les internautes qui se seraient trompés en saisissant l'URL du site dans la barre d'adresse. A ce moment, deux solutions s'offrent au webmaster qui a utilisé cette technique black hat : soit il crée sous ce nom de domaine un véritable site commercial, ou un site destiné à afficher des liens publicitaires, soit il redirige vers un autre site qu'il a créé et qui aborde la même thématique. Le typosquatting est une technique très répandue dans le milieu du web, mais elle fait aujourd'hui jurisprudence. Cela avait été le cas avec Air France et, plus récemment, avec le groupe les 3 Suisses³⁶.

Comme nous pouvons le remarquer, le spamdexing recouvre un ensemble de techniques très larges que les moteurs de recherche essaient chaque jour de détecter afin de ne pas mettre à mal la qualité de leur index et, par là-même, les fondements de leur système économique.

2.1.3. Guidelines et consignes aux webmasters

Les principaux moteurs de recherche ont mis à disposition des webmasters des « guidelines », c'est-à-dire un ensemble de consignes et de préconisations qui, si elles sont suivies, permettront au site d'être correctement indexés et de jouir d'un bon positionnement. Ces recommandations, sont regroupées, du côté de Google, dans le «

³⁶ CROUZILLACQ, Philippe, *Le groupe 3 Suisses assigne l'Afnic dans une affaire de typosquatting*, 20-07-2007 http://www.lepost.fr/article/2009/10/20/1751299_decouvrez-le-pourcentage-d-internautes-francais-qui-utilisent-google.html

Centre d'aide Outils aux webmasters »³⁷. Plusieurs champs d'action sont abordés par le moteur de recherche américain, qui peuvent être catégorisés en trois points :

- **Le travail sur le contenu et la structure** : il est conseillé de rédiger un contenu hiérarchisé (donc d'utiliser le balisage sémantique) et de créer un menu de navigation en « dur » de façon à ce que Google puisse suivre facilement suivre ses liens. De même, il est stipulé qu'il est important de placer dans le contenu les mots-clés qui, d'après nous, sont recherchés par les internautes. On le voit, Google n'encourage pas l'utilisation de son générateur de mots-clés, qu'il réserve aux utilisateurs de sa plate-forme Adwords. Google recommande également la création d'un plan de site, néanmoins nous pouvons nous interroger sur son utilité depuis que le protocole sitemap est apparu. En outre, Google réaffirme l'importance de créer des blocs de contenu pour les éléments importants d'un site, et d'éviter le surplus d'éléments graphiques. Il semblerait que Google attache de plus en plus de poids à l'application de cette recommandation, en témoigne la prise en compte de la vitesse de chargement des pages dans son algorithme. Cela ne va d'ailleurs pas sans susciter, parfois, des réactions négatives de la part de certains bloggeurs et de webdesigners. Dans un billet intitulé « *Google "dégueulasse"-t-il le web* »³⁸, Sébastien Billard rappelle pourtant que les moteurs de recherche sont conçus pour indexer du texte codé dans un langage simple et accessible, le HTML. C'est pourquoi leurs consignes vont dans le sens d'une construction avant tout éditoriale de chaque site web.
- **Le travail sur le code source et le développement du site** : Google recommande aux webmasters de visualiser leur site à l'aide d'un navigateur texte comme Lynx, ou tout simplement en désactivant images, CSS, Javascript et autres Flash. Ils verront alors leur site tel qu'il est parcouru par les robots d'indexation. C'est une bonne façon de vérifier si un site manque de contenu textuel en « dur ». Par ailleurs, Google insiste sur l'importance de proposer une version unique de chaque page, qui soit accessible via une

³⁷ Disponible ici : <http://www.google.com/support/webmasters/bin/answer.py?hl=fr&answer=35769>

³⁸ BILLARD, Sébastien, *Google dégueulasse-t-il le web?*, 27-07-2010

<http://s.billard.free.fr/referencement/?2010/07/21/616-google-degueulasse-t-il-le-web>

seule URL, dans le but de limiter le duplicate content. Enfin, il conseille d'utiliser le fichier robots.txt de manière appropriée.

- Le dernier point, celui qui nous intéresse le plus ici est relatif à **la qualité** : Google met en garde les webmasters qui « *cherchent en permanence des failles* ». En effet, même si les webmasters ne signent aucune charte de qualité lorsqu'ils demandent l'ajout de leur URL à la base de données du célèbre moteur (voir figuré 11), il n'en demeure pas moins que Google est propriétaire de son modèle et, qu'en tant qu'entreprise privée, il peut plus ou moins fixer ses propres « règles du jeu » et interdire selon son bon vouloir toute pratique qui lui semblerait contraire à son mode de fonctionnement.

Ajouter l'URL de votre site à Google

Indiquez-nous l'adresse de votre site Web.

Chaque fois que nous explorons le Web, nous ajoutons et mettons à jour de nouveaux sites dans notre index et nous vous invitons à utiliser cette page pour nous soumettre l'URL de votre site. Nous n'ajoutons pas dans notre index toutes les URL qui nous parviennent et nous ne pouvons ni prévoir, ni garantir le moment ou l'éventualité de leur inclusion.

Entrez l'adresse URL complète de votre site, y compris le préfixe `http://`. Exemple : `http://www.google.fr/`. Vous pouvez également ajouter des commentaires ou des mots clés décrivant le contenu de votre site. Ils sont uniquement utilisés pour information et n'ont pas d'incidence sur l'indexation ou l'utilisation de votre site par Google.

Remarque : Seule la page de premier niveau d'un hôte étant nécessaire, vous n'avez pas besoin de soumettre chaque page une à une. Notre robot d'exploration, Googlebot, se chargera de trouver le reste. Google mettant régulièrement à jour son index, il n'est pas nécessaire d'indiquer les liens mis à jour ou obsolètes. Les liens désactivés disparaissent de notre index au cours de l'exploration suivante, lorsque nous mettons à jour l'ensemble de l'index.

URL :

Commentaires :

Facultatif : Pour nous aider à distinguer les URL indiquées manuellement de celles soumises automatiquement, entrez le mot tel qu'il apparaît dans la zone ci-dessous.

Fig.11 : Page de soumission d'un site à Google

C'est là un réel paradoxe : les webmasters doivent se conformer à la politique d'une entreprise privée, sur un réseau par essence libre, qui n'appartient à personne, si ce n'est à la communauté des utilisateurs qui contribuent à son développement. A vrai dire, Google n'interdit rien au sens propre du terme, mais ignorer les règles de

fonctionnement de Google, c'est aujourd'hui compromettre énormément sa visibilité, puisqu'à 90% les internautes utilisent ce moteur³⁹. Mieux vaut donc suivre les règles énoncées. Revenons-en donc aux manipulations dépréciées par le moteur de recherche américain.

- **Texte et lien cachés** : il est relativement facile de mettre en place ce genre de techniques, en modifiant les styles CSS d'une page. L'intérêt est, par exemple, de truffier le contenu d'une page de mots-clés qui seront invisibles aux internautes : police minuscule, texte de même couleur que le fond, positionnement négatif d'un block (hors de l'écran), utilisation des propriétés `display:none` ou `visibility:hidden` sur un block, etc. Il en va de même des liens. Pour optimiser le maillage interne d'un site, sans pour autant nuire à l'expérience visuelle de l'internaute, il est possible d'annuler la mise en forme automatique d'un lien pour qu'il ne soit plus possible de le distinguer du contenu. Voici par exemple un code CSS pour pratiquer cette astuce sur un texte de couleur noire :

```
1 a.nom_de_la_classe, a.nom_de_la_classe:hover {  
2   color:black;  
3   text-decoration:none;  
4   font-weight:normal;  
5   cursor:text;  
6 }
```

Fig.13 : manipulation CSS sur un lien de couleur noire

Comme souligné par Olivier Andrieu, ces techniques sont difficilement détectables par les moteurs de recherche⁴⁰, qui auraient besoin de comparer à chaque page parcourue, l'aspect de la page telle qu'elle est affichée par un navigateur et celui de la page « aspirée ». Néanmoins, cette technique est de moins en moins utilisée, du

³⁹ Chiffres avancés par Le Post, dans *Découvrez le pourcentage de Français qui utilisent Google*, 20-10-2009. http://www.lepost.fr/article/2009/10/20/1751299_decouvrez-le-pourcentage-d-internautes-francais-qui-utilisent-google.html

⁴⁰ ANDRIEU, Olivier, *Réussir son référencement web*, p.348

fait des sanctions qui ont été prises sur certains sites et qui ont eu un écho retentissant dans la communauté du web.

- **Pages satellites** : nous l'avons vu dans la partie précédente, il s'agit de construire une page sur-optimisée qui sera redirigée par balise meta ou javascript vers une autre page, plus « propre » et plus agréable visuellement. Google est catégorique sur ce point, il ne faut pas les utiliser. Elles sont de moins en moins en vogue aujourd'hui dans la communauté « black hat ».
- **Cloaking** : il s'agit d'utiliser un langage dynamique tel que PHP et d'utiliser à bon escient l'interaction client/serveur afin de manipuler les fichiers renvoyés en fonction de l'entité qui se présente au serveur : un moteur de recherche ou un utilisateur « humain ».

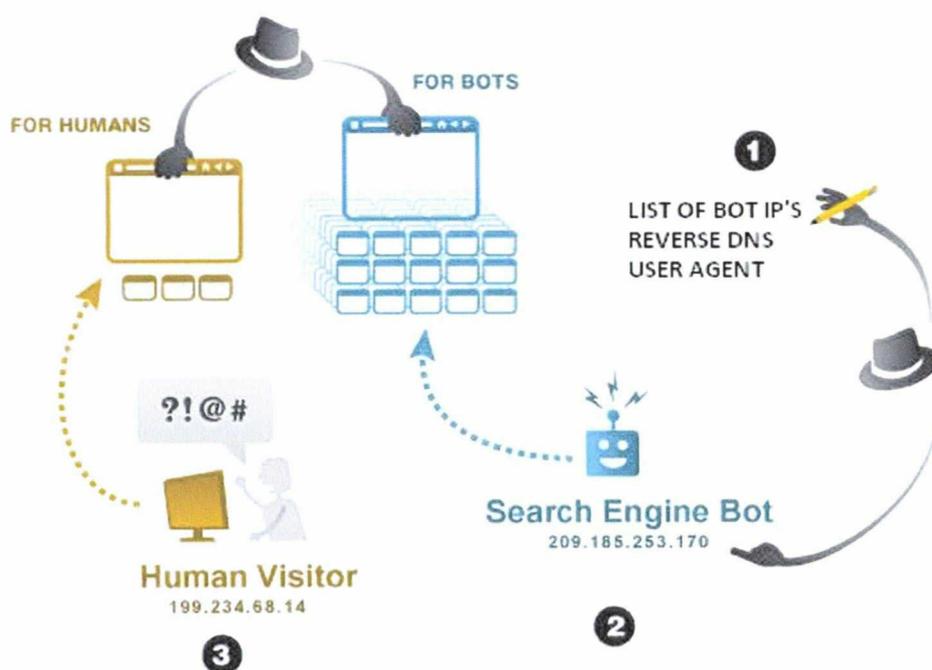


Fig.14 : principe du cloaking

Il semblerait que la technologie de Google appréhende de mieux en mieux les techniques de cloaking. Olivier Andrieu laisse entendre que les spiders de Google utilisent, parfois, d'un côté des IP qui ne sont pas reliés à leurs moteurs et de l'autre,

parallèlement, une IP « classique ». En comparant ainsi les deux versions du crawl d'un même site, ils peuvent détecter les tentatives de spam. Mais il existe d'autres formes de cloaking que le cloaking sur IP et qui sont plus difficiles à repérer. Nous reviendrons plus largement sur ces techniques avancées dans une autre partie du mémoire.

- **Les systèmes de liens artificiels** : c'est ici l'arme favorite des référenceurs « black hat », mais que Google considère avec la plus grande malveillance. Son algorithme serait en effet capable de détecter les liens non naturels (successions de liens, liens qui ne sont pas dans le contenu, réseaux de liens, etc.)

En résumé, l'optimisation d'un site web telle que livrée par Google dans son Outil aux webmasters ne suppose aucun écart de conduite. Néanmoins, les règles édictées par le géant américain sont régulièrement transgressées, en particulier dans certains domaines que nous allons étudier ici.

2.2. Utiliser le « black hat » : dans quelles occasions et à quels risques ?

L'utilisation du black hat, d'après ce que nous pouvons observer sur la toile n'est pas une tendance généralisée, il semble plutôt qu'elle concorde avec des sites où la concurrence est rude, ou encore des sites destinés uniquement à gagner de l'argent par l'affichage de publicités, parfois appelés « MFA »⁴¹.

2.2.1. Secteurs concurrentiels

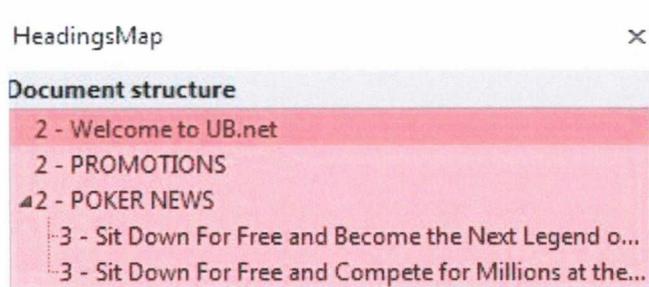
Dans des secteurs ultra concurrentiels, il apparaît que, pour obtenir des résultats rapidement, il est à l'heure actuelle devenu presque incontournable de faire appel à des techniques black hat. Même si les sources sont rares sur ce phénomène, il semble, d'après les témoignages de bloggeurs influents, et en vérifiant soi-même sur Internet, que le secteur des jeux en ligne, par exemple, soit complètement rongé par le spamdexing. Dans un article

⁴¹⁴¹ MFA : Made For AdSense

intitulé « *Online poker too competitive for white hat SEO ?* »⁴² paru sur le célèbre blog Seomoz, l'auteur se demande s'il est effectivement possible de se placer en première page de Google sur des expressions telles que « online gambling » ou « texas hold'em », en recourant à des techniques exclusivement « white hat ». Il semble en effet difficile d'arriver à ses fins sur des expressions aussi concurrentielles, pour lesquelles les premières places sont trustées par des sites volontiers « black hat ».

Prenons l'exemple de la requête « *online poker* » sur Google.com et analysons les résultats. Voici un rapide audit de la page qui ressort en cinquième position, à la date du 18/08/2010 mais que nous n'identifierons pas, par mesure de confidentialité.

- La **densité** de l'expression-clé « online poker » : pour calculer la densité d'un mot ou d'une expression nous pouvons utiliser le site Outiref, disponible à cette adresse : <http://www.outiref.com/>. Nous constatons alors que la densité de l'expression « online poker » est de 2,51%, ce n'est pas excessif voire faible.
- Les **balises sémantiques** : avec l'add-on pour firefox Headings Map, nous pouvons observer que les balises <h1> à <hn> sont utilisées de manière incorrecte : il manque le <h1> avant les <h2> et <h3>



- Les données **Whois** : le Whois est un service Internet qui permet de rechercher des informations sur le titulaire d'un nom de domaine. Grâce à ces données, nous observons que le nom de domaine a été acheté en 1997, ce qui a pu constituer un critère majeur dans la confiance que Google lui a apporté

⁴² RANDFISH, *Online Poker – Too competitive for white hat SEO ?*, 18-02-2007.
<http://www.seomoz.org/blog/online-poker-too-competitive-for-white-hat-seo>

- Le **contenu** : il est très limité mais pertinent et placé en premier dans le code grâce à une manipulation des positions CSS (il se retrouve en bas pour l'utilisateur)
- La balise **<title>** fait figurer les mots-clés sans spammer



- Le fichier **robots.txt** : il est présent et autorise l'indexation de toutes les pages sauf quelques répertoires.
- **PageRank** et liens entrants : grâce à l'extension pour Firefox SearchStatus, nous pouvons avoir un aperçu du PageRank, qui est de 4 sur 10. En contrôlant la page avec Yahoo Site Explorer, sont en effet comptabilisés 735 liens entrants sur la seule page index. Etant donné que la plupart de ces sites référents traitent de la même thématique, il semblerait que ce linkbuilding ne soit pas à base de spam.

A priori, d'après ce diagnostic rapide, on ne décèle aucun abus ni tentative de manipulation. Pour une requête aussi concurrentielle, il semble même assez étrange que le site se place aussi bien dans les SERP. Cependant, lorsque l'on jette un coup d'œil du côté du cloaking, nous pouvons identifier une technique black hat. En effet, lorsqu'un utilisateur lambda se connecte au serveur et demande l'affichage de la page, voici ce qui lui est retourné :

The free online poker site built by those who know poker best. We've been online since 1999 and our focus since day one has been to provide poker fans with a real-world poker room experience unlike anything else online. It's easy, fun, and safe to download and play online.

At UB, you'll find lightning-fast gameplay, great tournaments and ring games, and an amazing poker community to help you hone your game.

Best of all, it's 100% free. But that doesn't mean you have to walk away empty handed. We offer some of the best free tournaments on the planet where you can win a lot more than just a pat on the back. Browse the UltimateBet site to see what's at stake.

Fig 15. Footer de la page avec l'ip d'un utilisateur classique

Par contre, en utilisant sous Google la commande « cache : », suivie de l'URL du site pour afficher la version que les robots de Google ont indexée, voici le résultat :



Fig.16. Footer de la page tel qu'il est « vu » par Googlebot

Nous sommes donc ici en présence d'un cas de manifeste de cloaking : le contenu proposé à Google Bot est truffé de mots-clés avec l'expression « online poker ». D'après le blog Axe-Net, ce genre de procédés, cloaking ou système de liens artificiels notamment, est très largement utilisé dans les secteurs où les intérêts commerciaux sont les plus manifestes, tels que la pornographie, la vente de produits pharmaceutiques comme le viagra ou encore dans le domaine du crédit en ligne. Laurent Bourrelly, dans son article intitulé « *Ne tirez pas sur le référencement black hat* »⁴³ indique également que ce sont des mots-clés qui peuvent rapporter énormément d'argent, étant donné le volume de recherche qui s'y rapporte, et les sommes qui transitent sur les plateformes d'affiliation. Il donne l'exemple d'un webmaster, premier sur Google sur l'expression « viagra » et qui touche jusqu'à quarante mille dollars par jours, soit près de 4 milliards de dollars à l'année. Voici une bande dessinée, publiée sur le blog Ranked Hard, qui résume bien dans quelles circonstances le chapeau noir est le plus souvent de rigueur.

⁴³BOURRELLY, Laurent, *Ne tirez pas sur le référencement black hat*, 18-09-2007. <http://www.laurentbourrelly.com/blog/237.php>

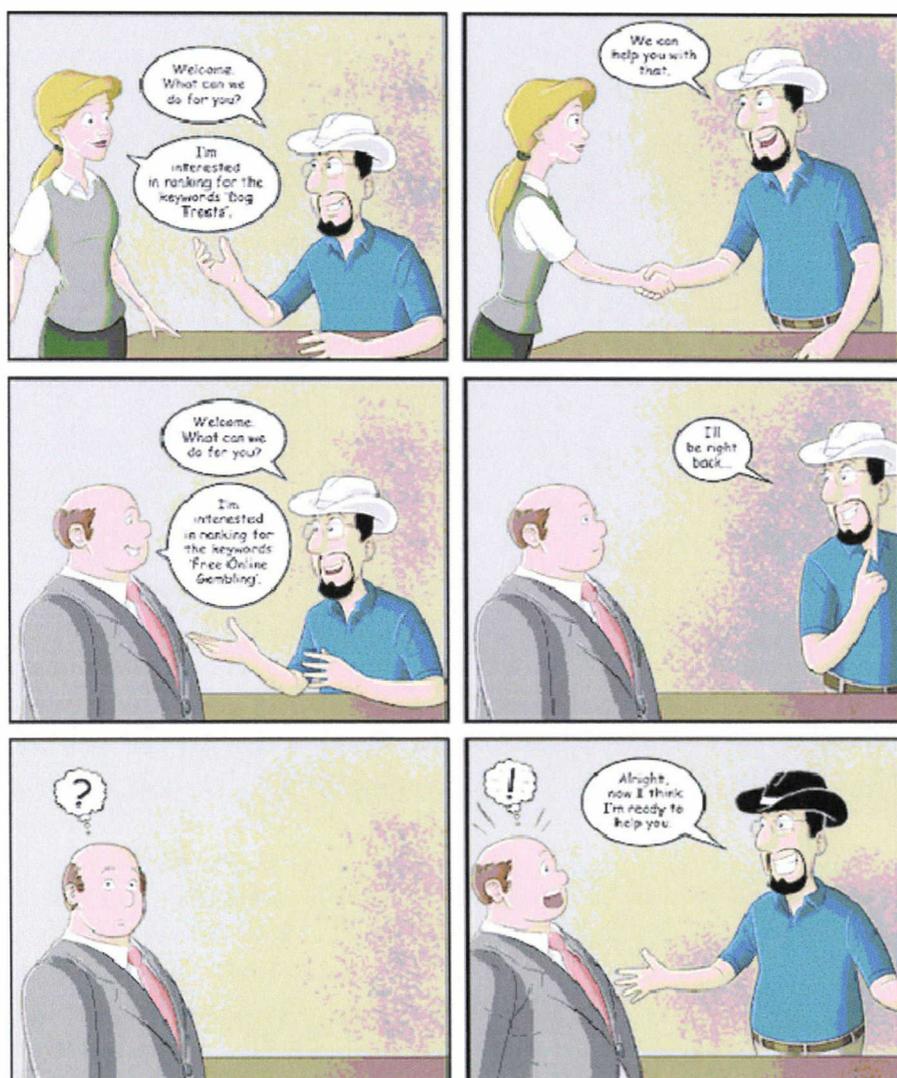


Fig.17 : Illustration black hat vs white hat⁴⁴

Outre ce genre de secteurs extrêmement concurrentiels, il existe un autre profil de sites web, qui utilisent consciemment et massivement les pratiques black hat : les sites « MFA ».

2.2.2. Sites « MFA »

Un site « MFA », « Made for AdSense » est un site qui est conçu exclusivement pour mettre en avant des publicités de la régie de Google AdSense. La régie AdSense utilise les sites web des webmasters qui y souscrivent, pour diffuser des annonces contextuelles. Google identifie en effet la thématique du site puis favorise l’affichage de publicités en rapport avec cette thématique. Le fonctionnement commercial est simple : lorsqu’un internaute clique

⁴⁴Disponible sur le site <http://www.rankedhard.com/>

sur une publicité, le propriétaire du site web touche une partie de la commission versée à Google par l'annonceur, c'est-à-dire celui qui a édité les annonces. Notons d'ailleurs que Google lui-même donne quelques conseils aux webmasters sur l'optimisation de l'affichage des annonces sur leurs sites. Néanmoins, il précise que les webmasters doivent permettre à l'internaute d'effectuer clairement et aisément la distinction entre le contenu du site et les publicités qui sont affichées.

Si de nombreux webmasters utilisent la régie AdSense sur un blog ou un site auquel ils s'efforcent d'ajouter un contenu pertinent et original, en soignant le référencement on page et off page, et ce de manière propre, la démarche de ceux qui produisent des « MFA » est toute autre.

En effet, l'objectif d'un « MFA » n'est pas de proposer un contenu qui soit pertinent et utile à l'internaute mais de l'inciter à cliquer sur les liens ou bandeaux publicitaires, afin de générer le maximum de revenus. Les propriétaires de tels sites sont souvent des coutumiers des pratiques black hat. Comme l'explique Sylvain Peyronnet du blog Axe-Net dans son billet intitulé « *Les MFA | Made For AdSense | sont-ils tous des pollueurs ?* »⁴⁵, certains webmasters vont employer des techniques black hat sophistiquées afin de placer leur site en bonne place sur de nombreuses expressions-clés et ainsi favoriser et optimiser le nombre de clics sur leurs annonces. Voici comment les black hat les plus pointus fonctionnent :

- **1^{ère} étape** : « scraper » un contenu, une liste d'URL ou encore un flux RSS. Sur le blog de Delicious Cadaver, un article intitulé « *Le web scraping ou comment piller les sites à la volée* »⁴⁶, définit ainsi le scrapping comme un moyen de récupérer le contenu textuel d'un site de manière complètement automatisée. Il s'agit donc, en d'autres termes, d'une technique visant à voler le contenu d'un site afin de l'utiliser à son profit. Pour procéder à ce genre de manœuvre, il suffit de lancer un script, développé par exemple en PHP, qui va simuler la visite d'une page, récupérer automatiquement son contenu et le générer à la volée sur son ou ses propres sites. Cette utilisation va bien sûr à

⁴⁵ PEYRONNET Sylvain, *Les MFA | Made For AdSense | sont-ils tous des pollueurs ?*, 04-10-2009. <http://blog.axe-net.fr/les-mfa-made-for-adsense-sont-ils-tous-des-pollueurs/>

⁴⁶ 512Banque, « *Le web scraping ou comment piller les sites à la volée* », 12-06-2009 <http://www.deliciouscadaver.com/le-web-scraping-ou-comment-piller-les-sites-a-la-volee.html>

l'encontre du droit d'auteur mais est pratiquée impunément dans le milieu du black hat. Il est même assez aisé de retrouver des scripts tout prêts, permettant, par exemple de « scrapper » des articles de wikipedia⁴⁷. Il est également possible de récupérer des snippets, des morceaux d'articles et de les agglomérer afin de créer un contenu plus ou moins unique.

- **2^{ème} étape** : le content spinning. Il s'agit, après avoir récupéré le contenu, de lui faire subir un traitement automatique, de manière à ce qu'il soit différent du texte d'origine et donc d'éviter la sanction du duplicate content qui risquerait de pénaliser le site. Pour cela, on ajoute aux mots, groupes de mots et syntagmes de la phrase des synonymes. Une fois passés dans le logiciel, toutes les possibilités sont combinées de manière aléatoire afin de créer un texte unique. Il va ainsi être possible de créer une quantité d'articles, sur la forme différents, mais sur le fond, tout à fait identiques. Ces articles vont venir alimenter des « autoblogs », qui désignent, dans le jargon black hat des blogs qui fonctionnent par eux-mêmes, en publiant des articles régulièrement, sans avoir besoin d'intervention humaine. Voici une démonstration de content spinning, effectuée sur le site <http://www.pagasa.net/spinner.php> :

Le texte à transformer

```
Ce {mémoire|rapport} de stage {traite|analyse} les  
{techniques|procédés} utilisés par les référenceurs  
{black hat|spammeurs}.
```

Le texte transformé

```
Ce mémoire de stage analyse les techniques utilisés  
par les référenceurs black hat.
```

- **3^{ème} étape** : rendre le contenu lisible et engranger quelques backlinks. Une fois que les SEO black hat ont réussi à contourner le risque du duplicate content, le contenu est souvent illisible pour les internautes. Si le robot parviendra facilement à lire le texte et à décrypter les syntagmes utilisés, le confort de lecture de l'internaute, sera lui, mis à mal.

⁴⁷ Voir ici : <http://blackhatseo-blog.com/wikipedia-scraped>

Le risque est alors qu'ils quittent le site sans même avoir cliqué sur les liens de publicité. Pour cela, les méthodes black hat de content spinning sont de plus en plus redoutables. Il existe certains logiciels qui effectuent des traitements complexes, sans toutefois rendre le texte illisible aux yeux des internautes. Une autre technique peut être également d'utiliser le cloaking, en servant aux robots « une bouillie » incompréhensible tandis que les internautes se verront proposer un contenu riche, par exemple un article issu de wikipedia. Par ailleurs, une fois que le site MFA dispose de son contenu, il ne lui reste plus qu'à acquérir quelques liens entrants, de manière à soigner sa popularité et ainsi monter plus facilement dans les pages de résultats. Quelques bons backlinks sont en général suffisants, car les MFA sont pour la plupart centrés sur des marchés de niche, peu concurrentiels. Généralement, les webmasters qui s'occupent de sites MFA puisent leurs liens depuis quelques annuaires sans intérêt pour les internautes, pour ne pas prendre le risque d'être démasqués.

Ceci nous amène donc à faire le point sur les sanctions prises par Google lorsqu'il identifie un site qui n'a pas respecté les consignes qualité évoquées précédemment.

2.2.3. Filtres et pénalités

Dans cette partie, nous nous limiterons à la position de Google vis-à-vis des pénalités imputées aux webmasters ayant pratiqué le spamdexing. Google a mis en place un éventail de filtres et de pénalités selon la gravité de la tentative de spamdexing. A l'heure actuelle, Matt Cutts est la figure emblématique de la politique qualité chez Google. Matt Cutts est en effet le responsable de la « Google Spam ». Il donne aux référenceurs de nombreuses informations et conseils sur la manière de référencer un site pour Google. Il répond d'ailleurs fréquemment aux écueils et aux questions des webmasters sur son blog disponible à cette adresse : <http://www.mattcutts.com/blog/>. Revenons-en à la nature des filtres et des pénalités connues, ou du moins supposées qui sont en vigueur chez Google.

- **La sandbox** : la sandbox, en français, « bac à sable », est un phénomène que Google n'a jamais réellement considéré comme avéré ni pour autant démenti. Cependant, de nombreux webmasters ont remarqué que lorsqu'un site est découvert par Google, le

moteur identifie le nombre de backlinks et le compare avec la moyenne des sites du même âge. S'il considère que ce nombre de liens entrants est trop élevé, alors il semblerait qu'il applique au site l'effet de « sandbox », autrement dit la mise en quarantaine⁴⁸ du site pendant quelques semaines. Le site est bel et bien référencé, mais classé dans les profondeurs des résultats de recherche. C'est d'ailleurs une des raisons pour lesquelles les référenceurs qui pratiquent le black hat sont très actifs dans le domaining, c'est-à-dire l'achat ou la vente de noms de domaine. En effet, en achetant un nom de domaine qui dispose déjà d'une certaine ancienneté ainsi que d'un volume de backlinks non négligeable, il est possible dès le lancement du site de démarrer une stratégie active de linkbuilding et d'obtenir des résultats.

- **Le filtrage** : il ne s'agit pas ici du filtre SafeSearch, qui permet de retirer des résultats de recherche les contenus à caractère « adulte » mais plutôt du filtre de duplicate content, appliqué lorsqu'une page utilise un contenu déjà existant dans l'index de Google. Dans ce cas cette page est placée dans l'index secondaire comme nous l'avons vu précédemment.
- **Le déclassement** : parfois et pour des requêtes très précises, il arrive que Google décline un site de manière brutale. Olivier Andrieu considère qu'il existe trois pénalités de ce type : « minus 30 », « minus 60 » ou « Position 6 penalty ». Elles sembleraient être la conséquence d'une suroptimisation (bourrage de mots-clés, liens cachés, etc .)
- **La baisse de PageRank** : il semblerait que la baisse de PageRank, l'indicateur de popularité auquel nous pouvons notamment avoir accès en utilisant la Google Toolbar, soit une sanction prise à l'encontre des sites pratiquant le commerce et l'achat de liens. De nombreux sites, dont certains très influents sur la toile, en ont fait les frais. Par exemple, le site WebRankInfo, très prisé par les référenceurs qui y dénichent une mine d'informations utiles, a connu en 2007 une baisse de PageRank. Il en a été de même pour le comparateur de prix Pixmania. Néanmoins, le classement de ces deux sites n'en a pas pour autant été affecté. Nous pouvons en déduire que Google essaie, par ce biais, de

⁴⁸ ANDRIEU, Olivier, Réussir son référencement web, p.354

montrer qu'il n'apprécie pas les tentatives de fraude. En effet, il souhaite communiquer sur le fait que pour l'achat de liens, sa plateforme AdSense est la plus adéquate.

- **Le blacklistage** : c'est ici la sanction la plus grave et la plus redoutée des webmasters. Un site blacklisté, ou en liste noire, est complètement désindexé par Google. La commande « site : » suivie du nom du site, permet de s'assurer, sur Google, si un site a, oui ou non été victime de blacklistage, comme nous pouvons le voir ci-dessous.



Pour être à nouveau référencé, il faut alors demander à Google un nouvel examen du site, par le biais des Outils aux Webmasters. Les causes du blacklistage sont multiples : cloaking, pages satellites, fermes de liens (c'est-à-dire réseaux de sites qui pointent les uns vers les autres de manière à gonfler artificiellement leur PageRank).

Dans les faits, nous pouvons remarquer que ces sanctions ne sont que rarement appliquées. Il semble que Google prenne des mesures occasionnelles, afin de « montrer l'exemple » comme ce fut le cas en février 2006, lorsque le site BMW a été blacklisté pour avoir utilisé des pages satellites sur son site. En dehors de ces quelques exemples, les sanctions de Google sont rares, même si l'équipe webspam du moteur de recherche assure qu'elle tente au quotidien d'améliorer son système de détection du spam.

De nombreux référenceurs sur la toile se targuent aujourd'hui d'utiliser impunément des techniques interdites, sans craindre un retour de bâton de Google. Pour eux, la seule vraie

menace est le « spam report ». Il s'agit d'un formulaire mis à disposition des webmasters pour signaler à Google qu'un site a utilisé des techniques illicites.

Rapport de spam

Aidez-nous à préserver la qualité des résultats de recherche Google

Nous travaillons sans relâche afin de vous proposer les résultats les plus pertinents pour chaque recherche que nous exécutons. Dans ce but, nous encourageons les gestionnaires de site à rendre leur contenu simple et compréhensible à la fois pour les utilisateurs et les moteurs de recherche. Malheureusement, tous les sites Web ne se soucient pas de l'intérêt des utilisateurs. Les pratiques consistant à tromper (spam) nos robots d'exploration du Web à l'aide de texte caché, de pages masquées (cloaking) ou de pages satellite (doorway) compromettent la qualité de nos résultats et dégradent la recherche pour tous les utilisateurs.

Nous désapprouvons les pratiques de spam. Par conséquent, nous vous demandons de nous signaler, à l'aide de ce formulaire, tout résultat de recherche Google que vous considérez comme spam. Nous examinons en détail chaque rapport concernant les pratiques trompeuses et prenons les mesures appropriées lorsque nous détectons des cas d'abus avérés. Dans certains cas extrêmes, nous supprimons immédiatement les spammeurs de notre index afin qu'ils ne figurent plus dans nos résultats de recherche. Dans tous les cas, nous utilisons les données de chaque rapport de spam afin d'améliorer notre système de classement de sites et nos algorithmes de filtrage qui, au fil du temps, doivent améliorer la qualité de nos résultats.

Vos commentaires nous aident à améliorer nos services dans l'intérêt des utilisateurs du monde entier et nous vous remercions d'avoir pris le temps de nous contacter. En nous aidant à éliminer le spam, vous permettez à des millions de personnes d'économiser du temps, du travail et de l'énergie.

Page Web ou site concerné :

Requête exacte à l'origine du problème (copiez cette requête à partir du champ de recherche Google) :

Page de résultats Google concernée par le problème (copiez l'URL de cette page) :

Type(s) de problèmes (plusieurs réponses possibles) :

- Texte ou liens masqués
- Termes trompeurs et répétitifs
- La page ne correspond pas à la description fournie par Google

Fig.18 : Rapport de spam Google

Matt Cutts encourageait en mars dernier⁴⁹ l'utilisation de ce formulaire pour dénoncer des sites ayant recours au spam. Il incite notamment les webmasters à être le plus précis possible sur les techniques employées, et à intégrer des mots-clés tels que « keyword stuffing » ou « blog spamming » dans les descriptifs, pour que leurs demandes soient traitées le plus rapidement possible. Les équipes humaines de Google sont en effet inondées de rapports de spam et doivent pouvoir rapidement faire la distinction entre de vrais et de faux rapports destinés à pénaliser un concurrent, puis identifier immédiatement de quoi il retourne.

Après avoir abordé de manière plus concrète l'univers du black hat et surtout dans quels secteurs d'activités il est le plus visible, il convient à présent de revenir sur les techniques avancées des référenceurs des black hat, et sur les moyens qu'ils déploient pour se protéger des éventuelles pénalités que nous venons de présenter.

⁴⁹CUTTS, Matt, *Calling for link spam report*, 03-03-2010. <http://www.mattcutts.com/blog/calling-for-link-spam-reports/>

2.3. Les techniques avancées « black hat »

Parfois, certaines figures du SEO français ou américains clament qu'un black hat SEO et un white hat SEO sont sensiblement identiques, à une différence près : l'automatisation.

2.3.1. Automatiser les processus

Les référenceurs black hat se distinguent essentiellement de leurs homologues white hat par l'aspect automatisé de leurs actions de référencement. En surfant sur les blogs ou les forums, nous pouvons nous apercevoir qu'ils échangent beaucoup autour de scripts, de morceaux de codes en langage PHP ou cURL⁵⁰, en Javascript, et discutent à propos de logiciels qui vont leur permettre d'automatiser toutes leurs tâches.

La pratique qui se prête le mieux à l'automatisation est le spam de commentaires à outrance. En effet, pour spammer un blog, certains utilisent une méthode manuelle, ils visitent des blogs, puis déposent un commentaire, tantôt pertinent, tantôt complètement hors-sujet, avec comme pseudo une ancre, c'est-à-dire une expression-clé. Cette ancre pointant vers le site de leur choix, ils essaient de gagner un maximum de popularité pour ce site sur les mots-clés choisis dans cette ancre. D'autres préfèrent automatiser cette tâche et investissent dans des logiciels plus ou moins performants. Nous pouvons ici faire un état des lieux des principaux logiciels utilisés :

- **Link Farm Evolution** : sous son diminutif LFE, Link Farm Evolution est un logiciel permettant de créer automatiquement un ensemble de splogs, c'est-à-dire des blogs qui n'ont d'autre utilité que de servir au spamdexing. Il s'agit d'utiliser les plateformes gratuites et open source telles que Wordpress MU, Blogger, Tumblr ou encore Pligg. Une simple recherche dans google sur « *wp-signup.php* » permet de se rendre du nombre de sites qui hébergent des scripts de Wordpress MU et donc du nombre potentiel de domaines uniques sur lesquels installer un blog : unblog.fr (voir ci-contre), blogetery.com, etc. Pour contourner l'étape du captcha⁵¹, le logiciel LFE dispose d'un

⁵⁰ cURL : interface en ligne de commande destinée à récupérer le contenu d'une ressource accessible par un réseau informatique (définition wikipedia)

⁵¹ Captcha : Test permettant de différencier un utilisateur humain d'un ordinateur

decaptcher qui va décoder automatiquement le capt'cha. L'utilisateur du logiciel peut alors choisir d'entrer les titres qu'il désire et qui seront utilisés comme sous-domaines des sites créés. Une fois que les blogs ont été créés, le logiciel va automatiquement publier des billets de blogs via la technique du content spinning décrite précédemment. Dans ce contenu seront insérés des liens contextuels vers le site à référencer et la blogroll de chacun des blogs pourra être paramétrée pour accueillir des liens avec une ancre optimisée. L'intérêt de ce logiciel est donc de fournir un ensemble de backlinks vers le site à référencer.

- **XRumer SEO** : ce logiciel russe est destiné aux webmasters qui pratiquent un usage que l'on peut qualifier de « professionnel » du black hat. Il permet de cibler certaines plateformes de blog, de forums, de guestbook, mais également des sites de pétitions en ligne et toute autre page où nous pouvons retrouver une balise <form> dans le code source, pouvant être utilisée par tout internaute pour ajouter du contenu. Le webmaster va alors y poster un commentaire, un avis, un message, en y déposant ses backlinks. Le logiciel permet étape par étape⁵² :



- de scraper les url des moteurs de recherche qu'il identifie comme des pages sur lesquels il est possible de « spammer »
- de vérifier le PageRank des pages qu'il spamme
- de donner l'accès à quantité de statistiques : backlinks répertoriés, captchas débloqués, etc.
- d'inscrire un utilisateur sur un forum, en éditant automatiquement son profil
- de créer un topic, c'est-à-dire un sujet de discussion, sur un forum et de participer à la discussion tel un utilisateur réel
- de déposer des commentaires sur les plateformes de blog qu'il a répertoriées
- de mettre en forme les commentaires via un script de content spinning intégré

⁵² Voir à ce sujet l'article de Discodog, *The Xrumer effect ce n'est pas l'outil qui fait le moine*, 14-06-2010. <http://www.discodog.fr/the-xrumer-effect-ce-nest-pas-loutil-qui-fait-le-moine.html>

- **Senuke** : c'est un logiciel là aussi très complet. Il permet de créer des comptes mail sur de services de webmails qui serviront ensuite lors de l'étape de création de profils sur les sites estampillés 2.0 : sites de bookmarking comme Diigo, *diggs-like* tels Digg ou Delicious, plateformes de blogging, article directories (répertoires d'articles). Une fois encore, en utilisant le content spinning, il est possible de choisir les différentes versions d'ancres sur lesquelles seront effectués les liens.

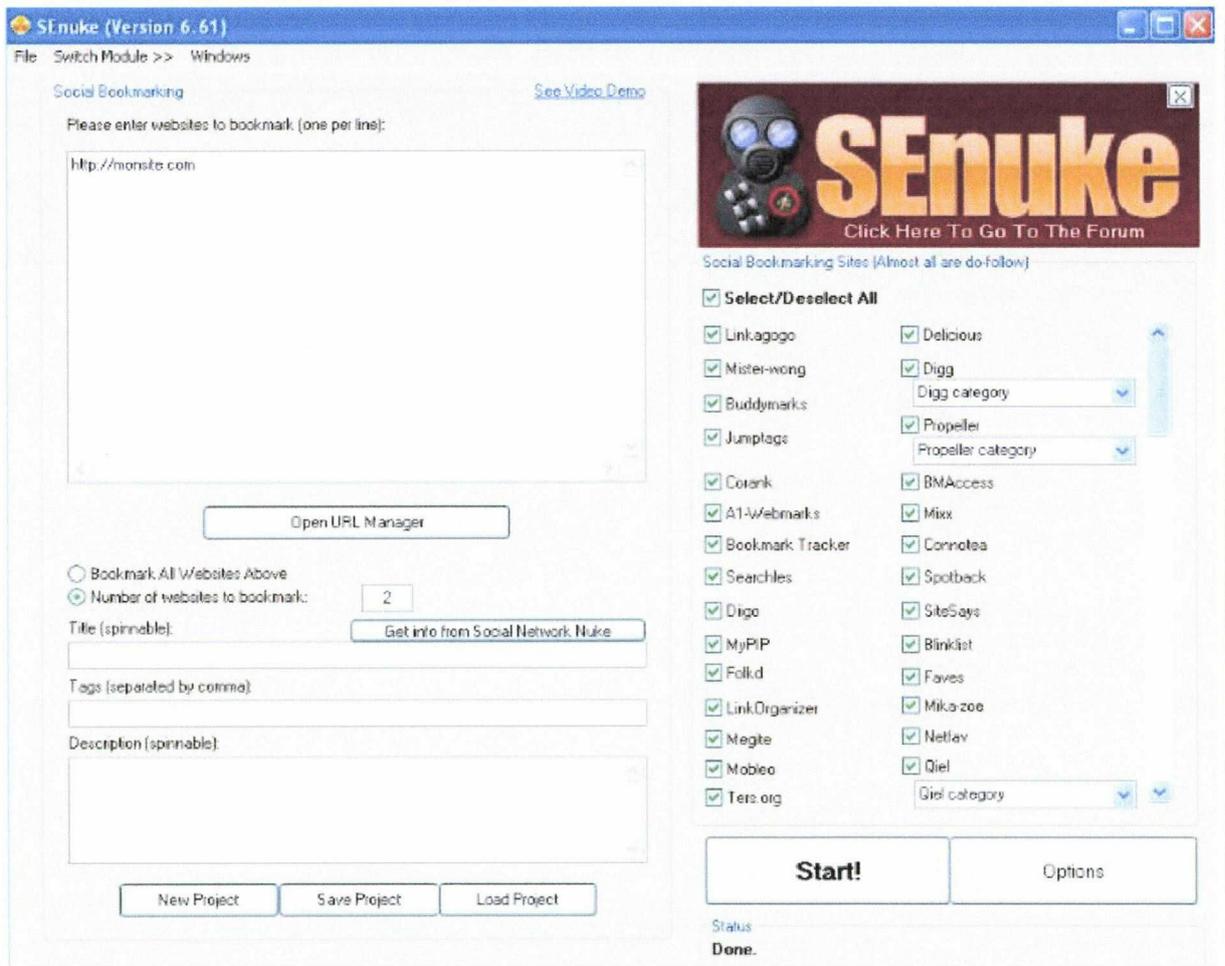


Fig.19 : capture d'écran du back office de Senuke⁵³

Tous ces logiciels utilisent ainsi le pouvoir des liens et du PageRank dans le classement des pages de résultats. Par ailleurs, il est à noter que les référenceurs black hat utilisent ces logiciels de manière combinée, de façon à consolider et à positionner un site qui sera, en quelque sorte, au sommet de la pyramide des sites sur lesquels les logiciels ont laissé une

⁵³ Crédits : SEObBlackOut, SMX Paris 2010 : Introduction aux techniques de linkbuilding borderline, 21-06-2010 <http://www.seoblackout.com/2010/06/21/smx-paris-2010/>

trace. Cette utilisation objective et évoluée des réseaux de sites est appelée le linkwheel, une vision qui tranche avec la pyramide du référencement telle que nous la présentons de manière traditionnelle. Comment cela fonctionne-t-il ?

- les sites de bookmarking et les digg-like, à la base, sont utilisés pour créer des liens vers des blogs ou forums hébergés sur des plateformes web 2.0.
- ces blogs, hébergés chacun sur un serveur différent, créent des liens uniques entre eux, de manière à augmenter leur PageRank
- chacun de ces blogs pointe vers un article soumis sur un répertoire d'articles à fort PageRank, tel que l'article directory américain
- cet article, à son tour, envoie un lien vers le site principal.

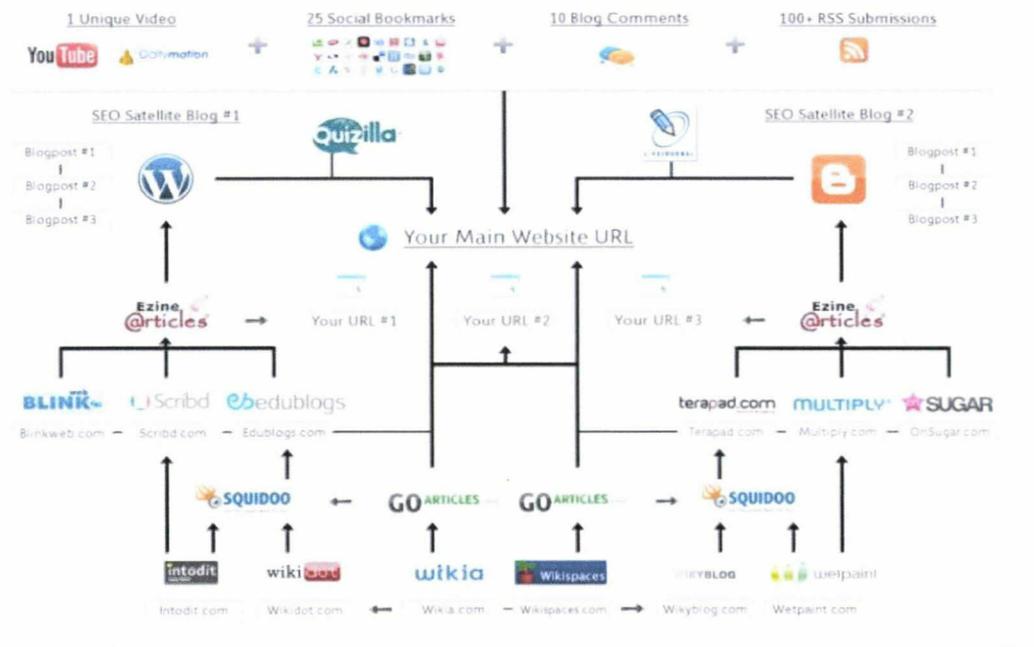


Fig.20 : schéma complexe d'une linkwheel⁵⁴

La création avancée et automatisée de backlinks sur l'ensemble des supports qui s'y prêtent sur le web est certes la plus utilisée dans le milieu du SEO Black Hat mais elle n'est pas

⁵⁴ Crédits : SEObBlackOut, SMX Paris 2010 : Introduction aux techniques de linkbuilding borderline, 21-06-2010 <http://www.seoblackout.com/2010/06/21/smx-paris-2010/>

unique. Voyons donc dans une prochaine partie un échantillon de pratiques s’assimilant encore au spamdexing.

2.3.2. Techniques et manipulations diverses

Ici nous ferons un point le plus exhaustif possible sur les différentes techniques, parfois méconnues, qui sont en vigueur dans le milieu du SEO black hat.

- **Spam de referer** : le spam de referer est une technique largement répandue dans le milieu du SEO black hat car elle est assez facile à mettre en œuvre. Le referer est l’adresse de la page sur laquelle était le visiteur qui vient d’arriver⁵⁵. Cette information est d’ailleurs transmise dans l’en-tête HTTP envoyé par un serveur. En utilisant un plug-in pour firefox tel que Ref-Control⁵⁶, il est possible de changer son referer lorsque l’on navigue sur Internet et ainsi de faire croire aux sites que l’on visite qu’ils reçoivent du trafic en provenance du site choisi comme referer. Outre le fait que cela va permettre au webmaster qui contrôle ses statistiques de découvrir le site que nous avons placé en referer et peut-être de s’y intéresser, cette technique est surtout un moyen efficace de gagner des backlinks depuis les outils de statistiques installés sur les serveurs des sites qui vont être spammés. Ainsi, des outils tels qu’AWStats ou Webalizer créent automatiquement des liens en dur vers les sites référents qu’ils détectent. Comme ces pages de statistiques sont parfois accessibles aux robots des moteurs de recherche, ces derniers y voient un ensemble de backlinks qui pointent vers le site concerné.

Links from an external page (other web sites except search engines) - Full list	103	7.1 %	109	6.5 %
- http://s[redacted]e.net/forum/forum.php	10	10		
- http://www.f[redacted]a.com/bp/c/5	10	10		
- http://s[redacted]e.net/forum/message.php	9	9		
- http://e[redacted]a.org/wiki/Awstats	8	8		
- http://e[redacted]a.org/wiki/Google	5	5		
- http://b[redacted]a.com/index.php/	4	4		
- http://l[redacted]e.com	4	4		
- http://forums.d[redacted]t.com/thread.php	3	3		
- http://www.h[redacted]x.com	3	3		
- http://www.p[redacted]w.com	3	3		
- Others	44	50		
Unknown Origin	3	0.2 %	3	0.1 %

Fig.21 : liens vers les sites référents sur AWStats.

⁵⁵ 512 Banque, *Outil de spam referer (gentil)*, 08-12-2008. <http://www.deliciouscadaver.com/outil-de-spam-referer-genti.html>

⁵⁶ Disponible ici en téléchargement : <https://addons.mozilla.org/fr/firefox/addon/953/>

- **Cloakings** : nous avons déjà présenté la technique du cloaking par IP-delivery, c'est-à-dire en fonction de l'IP de l'entité qui demande le chargement de la page. Il existe également le cloaking sur user-agent⁵⁷ qui permet de détecter quel est l'user-agent de celui qui se connecte au site et ainsi de délivrer au cas par cas, un contenu ciblé. Voici le code commenté PHP qui permet d'utiliser cette technique.

```
<?php
$trouve=strpos($_SERVER["HTTP_USER_AGENT"],"Googlebot");
if($trouve!==false){ // le visiteur est Googlebot, lui présenter la page cloakée
?>
<html>
... page cloakée pour Googlebot. </html>
<?php
}
else{ // le visiteur n'est pas googlebot, redirection vers une page standard?>
<html>
... page standard...
</html>
<?php
}
?>
```

Fig.22 : cloaking sur User-agent

Enfin, il existe également le cloaking par reverse DNS⁵⁸, qui va permettre de retrouver le DNS d'un visiteur à partir de son IP, et ce grâce à une fonction PHP : `gethostbyaddr()`. En disposant d'une liste des DNS des robots des moteurs de recherche, et en utilisant une fonction permettant de rechercher des correspondances dans les chaînes de caractères, il est ainsi possible de présenter aux robots des contenus différents de ceux présentés aux internautes.

- **Position relative en CSS et z-index** : il existe une propriété de positionnement CSS, `position:relative` permettant de placer un élément du code source avec des coordonnées négatives, vers la gauche ou la droite (ex : `position:relative;left:-1000px`) sans pour autant interférer avec le flux des autres éléments, de façon à ce que cet élément ne soit pas visible par les internautes, mais bel et bien scanné par les robots. C'est ainsi une

⁵⁷ User-agent : application cliente utilisée avec un protocole réseau particulier, par exemple un navigateur ou un robot comme Googlebot, Yahoo!Slurp,

⁵⁸ DNS : Système permettant d'établir une correspondance entre une adresse IP et un nom de domaine

autre manière de truffer une page de mots-clés sans que cela soit visible. Dans le même ordre d'idée, la propriété z-index permet de placer un élément en dessous d'un autre dans la page. Ainsi, il est possible de placer une première couche de contenu, volontairement suroptimisée, puis de la cacher en plaçant au-dessus un autre élément grâce à la propriété z-index. Lorsque l'on désactive la feuille de style CSS, il est par contre assez aisé d'identifier la supercherie.

Il existe encore une quantité très importante de techniques qualifiées de black hat comme l'achat de liens, l'injection d'URL⁵⁹, mais l'objectif n'est pas de dresser un catalogue des pratiques mais plutôt de montrer comment, avec des outils parfois très simples et un minimum de connaissances en programmation, il est possible de manipuler les résultats des moteurs de recherche. Voyons désormais comment les black hat se protègent.

2.3.3. Se protéger

Les référenceurs qui penchent du côté « black hat » ont tout intérêt à ce que leurs actions soient les plus discrètes possibles sur la toile, de manière, tout d'abord, à ce que Google ne puisse pas remonter jusqu'à eux et ensuite que les webmasters ne les dénoncent pas par le biais du « spam report » évoqué plus haut.

Pour se protéger, il existe différentes techniques et solutions à mettre en œuvre :

- utiliser la propriété « **noarchive** » : la propriété noarchive peut être ajoutée dans la balise meta destinée aux robots, dans la partie <head> du code html d'une page web. Elle indique en effet aux robots qu'ils ne doivent pas sauvegarder dans leur système de cache, la page dans laquelle figure cette balise. La version cache est, pour les robots des moteurs de recherche, la page que les spiders récupèrent et stockent sur un serveur, chaque fois qu'ils parcourent une page web. Elle est parfois utilisée par les internautes pour retrouver un contenu qui n'est plus disponible en ligne, mais qui est encore sauvegardé dans le cache des moteurs.

⁵⁹ Voir à ce sujet Pagasa, *Injection d'URL*, 19-12-2007. <http://www.pagasa.net/injection-durl/>

```
<META NAME="ROBOTS" CONTENT="NOARCHIVE">
```

Fig.23 : balise meta « noarchive »

En utilisant cette balise, le lien vers la version en cache, accessible depuis la page de résultats ne sera donc pas accessible. De ce fait il est impossible pour quiconque de visualiser ce que les robots ont indexé. Il s'agit là d'une technique efficace pour camoufler un cloaking par exemple. Il est également possible de définir cela directement dans le fichier .htaccess, par le biais de ce code :

```
<Files ~ >  
header set X-robots-tag "noarchive"  
</Files>
```

Ou encore de passer cette valeur directement dans le header HTTP, via un script PHP :

```
<?php header ('X-Robots-Tag: noarchive'); ?>
```

- utiliser plusieurs **hébergeurs** différents : une deuxième règle d'or du référencement black hat est de varier les hébergements et les IP des sites que l'on essaie de faire progresser, de façon à ce que Google ne puisse pas « tracer » les propriétaires des sites ou qu'un webmaster ne puisse pas découvrir le réseau de sites. En effet, il existe aujourd'hui des logiciels permettant de retrouver tous les noms de domaine hébergés sur un même serveur, créant donc le risque de démasquer un référenceur black hat. Dans la mesure du possible, il est également conseillé de masquer les informations Whois, ce que proposent certains services d'hébergement.
- **ne pas lier** les sites entre eux : afin que les sites MFA ou d'autres sites destinés à être monétisés ne soient pas démasqués, une des règles d'or du référencement black hat est de ne pas créer de liens bidirectionnels.
- ne pas ajouter de code **Google Analytics** : les référenceurs black hat conseillent de ne pas insérer de code de tracking Google Analytics de manière à transmettre le moins

d'informations possibles à Google sur leurs pratiques et ne pas être « profilés ». Ils conseillent également la prudence vis-à-vis de la régie AdSense.

- **bloquer l'accès** des robots d'indexation à certains fichiers : lorsque les référenceurs black hat utilisent par exemple un script JavaScript leur permettant, aux yeux de Google, de masquer un lien, ils font en sorte que l'accès à ce fichier soit bloqué via le fichier robots.txt.
- utiliser des **proxies** pour rester anonyme : les proxies sont des serveurs qui relaient des requêtes entre un poste client et un serveur. Ils permettent donc de camoufler l'identité du poste client lorsque celui-ci va utiliser des méthodes black hat. En utilisant des proxies performants, il devient donc possible de rendre anonyme le spamdexing et ainsi de ne pas subir de blocage de la part des sites spammés, alors incapables d'identifier l'IP du visiteur.

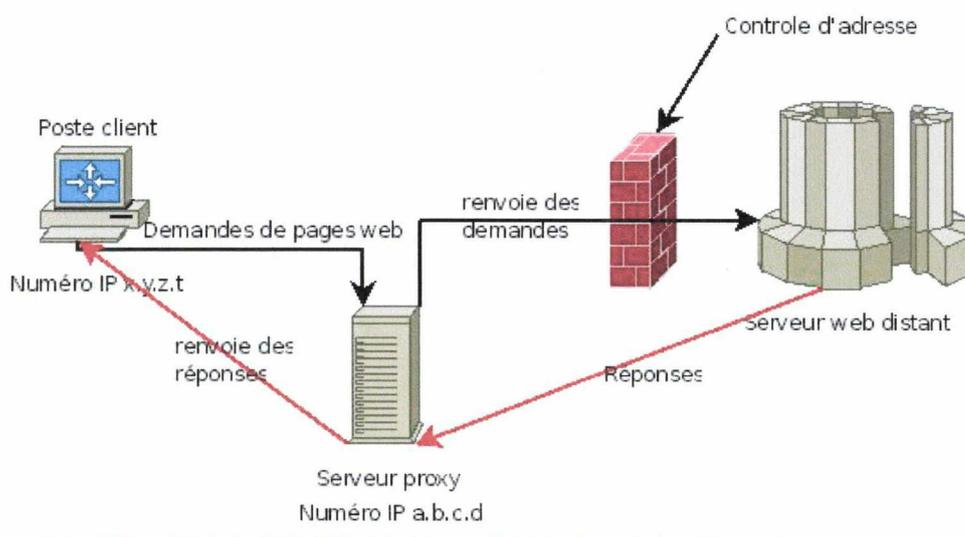


Fig.24 : schéma de fonctionnement d'un proxy

Nous pouvons donc constater que le milieu du black hat SEO ne manque pas d'ingéniosité pour trouver des techniques de spamdexing. Face à de telles pratiques, la question est à présent de savoir comment doit se positionner un chargé de référencement ?

3. Comment doit se positionner le référenceur ?

Comme nous pouvons le voir, la tentation peut être grande pour un webmaster d'utiliser à bon escient les techniques de spamdexing. Si l'utilisation du black hat pour un site personnel n'engage que la responsabilité du webmaster, elle comporte davantage de risques en agence. En effet, l'utilisation du black hat fragilise les performances du site des clients, qui peuvent du jour au lendemain se voir affubler d'une pénalité. Il s'agit donc de trouver un bon compromis entre pratique éthique et suroptimisation.

3.1. Où s'arrêtent les bonnes pratiques et où commence le spam ?

Pour référencer éthiquement un site web, il est possible de s'inspirer de ce que les référenceurs black hat utilisent, sans pour autant reproduire à l'identique leurs manœuvres. Voyons tout d'abord, ce que les « black hat » apportent au SEO de manière général.

3.1.1. Ce qu'apportent les chapeaux noirs au SEO

Tout d'abord, nous pouvons constater que les référenceurs qui basculent du côté des techniques black hat permettent à la discipline du SEO d'avancer, de progresser. Ils donnent une certaine ouverture d'esprit à leurs collègues white hat. En effet, en expérimentant certaines techniques et en fleurant avec le blacklisting, ils permettent aux « white hat » de distinguer ce qu'il est possible de faire de ce qui ne l'est pas. Par exemple, ils étudient les différents brevets de Google (et des autres moteurs de recherche) qui ont été déposés pour cerner de plus près comment fonctionne un moteur de recherche. Par ce biais, ils contribuent à mieux définir ce qui est considéré comme spam. Dans un article intitulé « *Web Spam : le guide SEO spamdexing* »⁶⁰, SEO Black Out explique que les moteurs de recherche ont accumulé une sorte de base de connaissances qui leur permet de profiler, de détecter, ou du moins de suspecter une tentative de spam. Ainsi, des critères comme l'extension de nom de domaine (un .biz serait par exemple plus suspect), la longueur du texte de la page, le nombre d'ancres de liens, l'évolution rapide ou plus lente du nombre de liens entrants sont analysés par les équipes anti-spam de façon régulière. Ce qui ressort de cet article, c'est que

⁶⁰ *Web spam : le guide SEO spamdexing*, 23-07-2010. <http://www.seoblackout.com/2010/07/23/web-spam-seo/>

les ingénieurs qui veillent au bon fonctionnement des moteurs de recherche établissent un certain nombre de grilles de statistiques : sur la base de ces chiffres, ils avancent dans la lutte contre le spam. Pour les référenceurs dont la pratique est orientée white hat, il est donc primordial de suivre l'activité des équipes anti spam, pour axer ces efforts sur ce qu'il est ou non possible de faire dans le cadre d'une optimisation du référencement naturel.

Par ailleurs, le deuxième bénéfice du black hat tient à la façon dont ces référenceurs envisagent le métier de SEO. En effet, ils montrent énormément de curiosité et d'inventivité dans leur manière de concevoir la discipline. Un exemple concret est par exemple leur vision de la création de liens. Si la vision traditionnelle du référenceur est de se consacrer sur la création de liens depuis des annuaires, communiqués de presse ou encore digg-like et de se concentrer sur l'originalité du site et de son contenu pour que les webmasters créent naturellement des liens vers le site (principe du linkbaiting), les référenceurs black hat, eux, font preuve de plus d'imagination, de créativité, en pratiquant une « chasse au liens » permanente. Ainsi, ils permettent d'identifier quelles sont les plateformes sur lesquelles il est possible de déposer un lien, comme par exemple :

- les forums : une signature dans un forum avec un lien discret vers son site peut être tout à fait assimilé à du white hat
- les blogs : l'ajout de commentaires suite à un article peut tout à fait se rapprocher du white hat si cela est effectué de manière pertinente, en apportant une réelle plus-value à l'article. Ainsi, Sylvain Peyronnet dans un article intitulé « *Comment spammer un blog dofollow* »⁶¹ donne aux webmasters des conseils sur la façon dont les commentaires peuvent participer à la discussion dans les blogs, sans pour autant que ces derniers soient rejetés par l'administrateur pour spam.

En outre, les référenceurs black hat obligent les moteurs de recherche à constamment améliorer leur technologie, de façon ce qu'ils retournent les résultats les plus pertinents, correspondant aux intentions de recherche des internautes. Laurent Bourrelly parle ainsi, à

⁶¹ PEYRONNET, Sylvain, *Comment spammer un blog dofollow ?*, 22-11-2009. <http://blog.axe-net.fr/comment-spammer-un-blog-dofollow/>

ce propos, d' « émulation saine »⁶². Le duel constant entre les référenceurs et moteurs de recherche serait au final un duel galvanisant, duquel seul l'internaute sortirait gagnant, car bénéficiant d'un service informationnel optimal.

Enfin, nous pourrions presque clamer que le référencement black hat est utile dans la mesure où il est à l'origine du développement de nombreux outils qui facilitent le travail quotidien des référenceurs. En effet, les adeptes de l'automatisation créent parfois des outils qui peuvent être utilisés dans le cadre d'une pratique « white hat ». A titre d'exemple, intéressons-nous à un scrapper qui permet d'aspirer l'ensemble des requêtes de Google Suggest, relatives à un mot-clé. Google Suggest est un service proposant à l'internaute les requêtes les plus populaires en fonction du mot-clé qu'il est en train de saisir. Il est très utilisé dans le référencement, notamment pour connaître les différentes associations de mots-clés qui sont frappés par les internautes. En « scrappant » Google Suggest, nous pouvons donc récupérer l'ensemble des variantes de Google Suggest pour un même mot-clé et selon plusieurs niveaux de profondeur.

Script Google Suggest

référencement

Langue : FR ▾
Profondeur : 1 ▾
Alphabet :

Keyword : et

```
0 ; référencement
0 ; référencement google
0 ; référencement gratuit
0 ; référencement naturel
0 ; référencement google gratuit
0 ; référencement site
0 ; référencement yahoo
0 ; référencement newsgroup
0 ; référencement web
0 ; référencement site web
0 ; référencement internet
0 ; référencement internet gratuit
0 ; référencement internet google
0 ; référencement internet wiki
0 ; référencement internet comment ca marche
0 ; référencement internet définition
0 ; référencement internet forum
0 ; référencement internet garanti
0 ; référencement internet pour les nuls
0 ; référencement site internet
0 ; référencement site internet gratuit
0 ; référencement site internet google
0 ; référencement site internet sur google
0 ; référencement blog
0 ; référencement blogger
0 ; référencement blogspot
0 ; référencement blog google
0 ; référencement blog gratuit
0 ; référencement blog wordpress
```

Il suffit d'entrer dans le champ de recherche le mot-clé désiré (ci-dessus) pour voir apparaître l'ensemble des requêtes associées à ce mot-clé qui sont fréquemment recherchées par les internautes sur le moteur de recherche Google (ci-contre).

⁶² BOURRELLY, Laurent, *Ne tirez pas sur le référencement black hat*, 18-09-2009.
<http://www.laurentbourrelly.com/blog/237.php>

3.1.2. Bénéfices et limites des actions black hat

Lorsqu'un webmaster prend la décision d'utiliser consciemment des techniques qui sont contraires aux règles énoncées dans les guidelines d'un moteur comme Google, et s'il ne prend pas un minimum de précautions, il s'expose à un certain nombre de risques. Nous pouvons nous ici nous demander quelles sont les bénéfices et les limites des différentes actions black hat.

Tout d'abord, il est à noter que l'utilisation de techniques black hat dans des secteurs concurrentiels, où les gains financiers mis en jeu sont énormes, porte parfois ses fruits. Ainsi, dans le domaine du poker ou des jeux en ligne, il arrive parfois que de nouveaux sites fassent leur apparition surprise, très rapidement, dans la première page des SERP, car ils ont eu recours à des techniques de spamdexing et sont passés entre les mailles du filet des moteurs de recherche. De même, quelques exemples de ce que nous pouvons qualifier de « gros sites » ont acquis une position favorable et durable dans les résultats de recherche des moteurs de recherche grâce à une utilisation totale ou partielle du black hat. Nous pouvons ici appuyer notre raisonnement sur deux exemples concrets :

- dans un article intitulé « *Qype, une des plus belles réussites du référencement...BlackHat ?* »⁶³, l'auteur du blog 404 création explique que le site Qype, un guide des bonnes adresses orienté 2.0, s'est appuyé sur un certain nombre de techniques black hat pour assurer son bon positionnement sur les moteurs de recherche, sur des requêtes souvent génériques et concurrentielles. La technique reposant essentiellement ici sur la création d'un réseau de sites satellites, ou plutôt de blogs sur lesquels nous retrouvons un contenu tantôt pertinent, tantôt indigeste avec un simple copier/coller de sites existants et qui intègre de nombreux liens vers certaines fiches publiées par le site Qype.fr, avec des ancres parfaitement optimisées. L'auteur de cet article dénonce cette pratique effectuée en toute impunité, sans craindre le retour de bâton de Google, alors que la majeure partie des blogs sont notamment installés sur BlogSpot, la plateforme officielle appartenant à ce même Google.

⁶³ *Qype, une des plus belles réussites du référencement...BlackHat ?*, <http://www.404-creation.com/referencement/qype-referencement-blackhat.php>.

- dans un article intitulé « *SEO : les gros sites peuvent-ils tout se permettre ?* »⁶⁴, Virginie Clève dresse un bilan assez éloquent sur les pratiques « spammantes » d'un certain nombre de mastodontes du web, qui restent cependant impunis. Elle s'étonne notamment de constater que certains de ces sites abusent de techniques black hat évidentes et facilement démasquables. Ainsi, les différentes techniques utilisées sont assez basiques et vont du simple texte caché à l'offuscation de liens en javascript, en passant par les liens invisibles permettant de dissimuler les actions de linking interne trop envahissantes pour l'internaute, comme ci-dessous :



Fig.25 : article d'un grand quotidien national (CSS activés)



Fig.26 : Le même article, en désactivant une classe CSS sur les liens

⁶⁴CLEVE, Virginie, *SEO : les gros sites peuvent-ils tout se permettre ?*, 28-06-2010. <http://www.cafe-referencement.com/lectures/seo-les-gros-sites-peuvent-ils-tout-se-permettre-269>

Après avoir dénoncé ces faits, Virgine Clève en vient à s'interroger sur l'efficacité des cellules anti-webspam et du spam report. Du moins, il semble qu'elle mette le doigt sur un phénomène empirique tendant à démontrer qu'il existe « deux poids deux mesures » dans la lutte menée par Google et consorts dans la lutte contre le spam : si les « petits sites » sont vite écartés des premiers rangs des SERP lorsqu'ils sont confondus par les équipes anti spam, le traitement réservé aux sites disposant d'une plus grande notoriété et utilisant des techniques black hat est différent. Les sanctions sont rares ou ne sont destinées qu'à montrer l'exemple de la toute puissance des moteurs, comme ce fut le cas avec Pixmania dont nous avons parlé précédemment. Dans un article publié sur SEOMoz⁶⁵, un webmaster anglais s'inquiète lui aussi du manque de réactivité de Google dans la lutte contre le spam et constate avec étonnement que des sites ayant touché de près ou de loin au black hat se retrouvent en bonne position sur des requêtes génériques telles que « seo software », « nanny services » ou encore « french doors ».

Face à tant d'exemples d'utilisation on ne peut plus bénéfique du black hat, nous pouvons néanmoins rétorquer que la menace est bien réelle, et que s'adonner au SEO black hat sans prendre les précautions et la discrétion qui s'imposent peut conduire à des sanctions lourdes de conséquence : comptes AdSense bannis pour ceux qui créent des MFA, hébergement et adresses IP surveillées, etc. En effet, en fréquentant les forums black hat, nous pouvons nous apercevoir que, bien souvent, certains novices en matière de black hat, se retrouvent du jour au lendemain confrontés à une suppression totale de leur site de l'index de Google. La limite des actions black hat est bel et bien là : pour pouvoir les utiliser, il faut savoir comment les utiliser correctement et savoir garder l'anonymat. Il est donc nécessaire d'effectuer un travail de veille et de « cache-cache »⁶⁶ permanent avec les moteurs, ce qui s'avère extrêmement chronophage. Nous pouvons alors en conclure qu'en agence, ce type de procédé est à proscrire, tant le risque de mise au ban des moteurs de recherche est important.

⁶⁵ RANDFISH, *I'm getting more worried about the effectiveness of webspam*, 17-08-2010.

<http://www.seomoz.org/blog/im-getting-more-worried-about-the-effectiveness-of-webspam>

⁶⁶ PEYRONNET, Sylvain, *SEO : black hat ou white hat ?*, 05-07-2009. <http://blog.axe-net.fr/seo-black-hat-ou-white-hat/>

Etant donné que le référencement black hat peut amener à de jolis succès, notamment à court terme, mais qu'il comporte tout de même un certain nombre de risques et notamment le spectre du spam report, une tierce catégorie de référenceurs est apparue, qui jongle entre le référencement éthique et le spamdexing : les grey hat.

3.1.3. Le grey hat, un entre-deux ?

Le grey hat SEO désigne une vision ou plutôt une pratique du référencement qui balance entre les techniques approuvées par les moteurs de recherche et les techniques réprouvées. En effet, les techniques qu'utilisent les grey hat se situent à la charnière entre les deux influences et peuvent basculer du côté black hat ou du côté white hat suivant la façon dont elles sont implémentées sur un site web. Voici ici une infographie publiée sur le site de SEOMoz résumant bien la frontière entre les deux mouvances.



Fig.27 : infographie du SEO⁶⁷

⁶⁷ RANDFISH, 4 essential SEO infographics, 10-08-2009. <http://www.seomoz.org/blog/4-essential-seo-infographics>

Ainsi, parmi les techniques de référencement les plus controversées, qui peuvent être étiquetées « grey hat », nous pouvons distinguer :

- **l'achat de liens** : s'il est pratiqué dans l'optique d'améliorer de manière artificielle le classement d'un site dans les résultats des moteurs de recherche, l'achat de liens est considéré comme « illégitime » par Google. Par contre, s'il est utilisé dans un objectif publicitaire, il est toléré. C'est par exemple le cas des liens d'affiliation, qui sont, quoiqu'il en soit, toujours assortis de l'attribut « nofollow », indiquant aux moteurs de recherche qu'ils ne doivent pas être pris en compte dans le transfert de popularité. De la même manière, des liens peuvent parfois être achetés dans le cadre de sponsoring ou de partenariat avec d'autres sites. Dans ce cas, l'ambiguïté est beaucoup plus grande pour les moteurs de recherche, qui auront du mal à identifier s'il n'y a pas velléité de manipuler l'algorithme de classement.
- **l'échange de liens** : bien qu'utilisé par de nombreuses agences de référencement, l'échange de liens d'un site vers un autre n'est pas considéré par Google comme une technique loyale, dans la mesure où elle n'est pas réellement le fruit d'un échange de bons procédés. L'objectif premier d'un échange de liens est de gonfler le PageRank d'un site de manière artificielle. Cependant, il est difficile pour Google d'identifier ce genre de procédés, hors-mis si un spam report lui est transmis. Il s'agit donc là aussi d'une pratique grey hat.
- **le spam de blog** : comme son nom l'indique, cette pratique semble a priori tout droit se ranger du côté du black hat, cependant, il n'est pas si facile de trancher étant donné que l'ajout de commentaires sur un blog est parfois pertinent. C'est d'ailleurs même l'essence du web : l'échange. Il est vrai que si cette pratique est automatisée, alors on a affaire au black hat le plus évident mais si elle est réalisée de manière manuelle, elle peut être considérée comme une ressource utile pour le référencement white hat.
- **le PageRank Sculpting** : le PageRank Sculpting désigne l'ensemble des techniques visant à optimiser le transfert de PageRank vers les pages profondes d'un site. En effet, effectuer trop de liens sortants depuis la page d'accueil d'un site dilue la valeur du

PageRank transmise aux autres pages. Certains webmasters ont donc imaginé des techniques permettant de cacher certains liens aux yeux de Google et des autres moteurs de recherche : redirections 302 (temporaires) de la page sur laquelle pointe le lien, lien encapsulé dans un fichier javascript, lien en flash ou en JQuery, etc. La position de Google sur ce phénomène est assez floue, même si l'exemple du nofollow (voir plus haut) avait pour un temps sonné le glas de la sculpture de PageRank. Aujourd'hui, il est possible de réaliser du PageRank Sculpting de manière white hat, par exemple en regroupant des contenus sur une seule page, en supprimant certaines pages inutiles ou encore en limitant le nombre de liens externes.

Dans l'ensemble des blogs et autres forums que nous avons pu consulter, la définition du grey hat est une version positive et quelque peu édulcorée du black hat. En effet, il s'agit des webmasters qui maîtrisent les techniques classiques de référencement sur le bout des doigts et qui essaient, tout en restant dans un cadre éthique, d'exploiter les failles des moteurs de recherche. Un article très intéressant paru sur le blog de SEO Player, « *Les techniques SEO black hat au devant de la scène* »⁶⁸ paru le 1^{er} décembre 2008 montre à quel point il est extrêmement délicat de trancher entre ce qui est du côté du black hat et ce qui est du côté du white hat. Sylvain Peyronnet se demande également s'il existe « un juste milieu »⁶⁹. Selon lui, le terme de grey hat renvoie à tous les leviers du SEO qui ne sont pas textuellement réprouvés par Google, mais pour lesquels le moteur de recherche n'adopte pas une position tranchée. Selon lui, ces techniques ne peuvent déboucher que sur des pénalités minimales et sont donc parfois utiles dans des secteurs concurrentiels.

Mon avis sur la question est qu'il est quasiment impossible de ne pratiquer que du white hat et que tout référenceur est amené à un moment ou un autre à utiliser des petites astuces qui vont lui permettre de forcer quelque peu l'optimisation du site sans pour autant basculer dans la suroptimisation. Ces techniques reposent avant tout sur une bonne connaissance du langage de balisage HTML, ainsi que du langage de présentation CSS.

⁶⁸ SEOPAYER, *Les techniques SEO black hat au devant de la scène*, 01-10-2008.

<http://www.seoplayer.com/optimisations-seo/les-techniques-seo-black-hat-au-devant-de-la-scene.html>

⁶⁹ PEYRONNET, Sylvain, *Black hat or white hat le SEO*, 05-07-2009. <http://blog.axe-net.fr/seo-black-hat-ou-white-hat/>

3.2. Google est-il responsable du spamdexing ?

Sans prendre de parti réducteur, il est aujourd'hui logique de se concentrer sur Google lorsque nous parlons d'optimisation pour les moteurs de recherche, tant son emprise est consternante sur le monde de la recherche d'information. Comme nous avons pu le voir, les actions black hat visent à manipuler l'index du moteur de recherche américain, mais nous pouvons nous demander dans quelle mesure Google est responsable de ce spamdexing.

3.2.1. Les failles de l'algorithme du géant de Mountain View

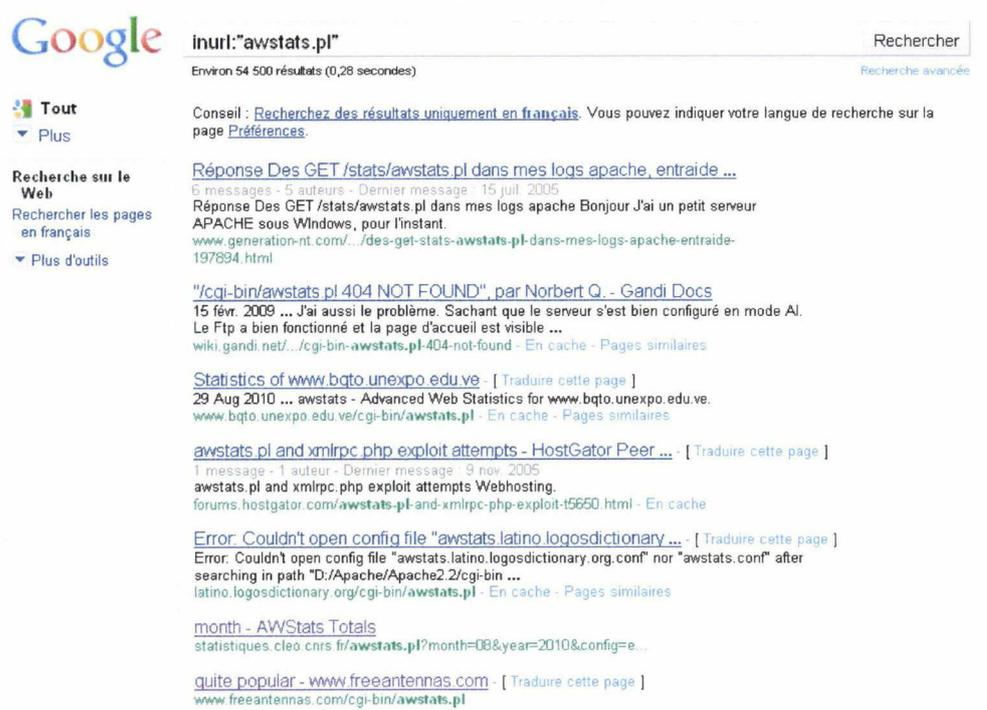
Tout d'abord, nous pouvons constater que les référenceurs black hat « ne font » qu'exploiter avec malice un certain nombre de failles décelées dans les algorithmes des moteurs de recherche. Ainsi, dans le milieu du black hat, c'est le propre outil de Google qui se retourne contre lui. Nous pouvons même dire que Google est leur ami. Pourquoi ? Tout simplement, parce que les black hat utilisent à leur avantage les opérateurs de recherche spécifiques de Google afin d'effectuer des recherches avancées et de retrouver les sites sur lesquels ils vont pouvoir s'adonner aux pratiques de spamdexing les plus redoutables.

Voici quelques listes de commandes qui vont permettre d'identifier les sites sur lesquels il est possible de poser des liens :

- **la commande « site : »** : cette commande bien connue des référenceurs, permet d'afficher l'ensemble des pages d'un site qui ont été indexées par Google. Par exemple, `site:www.univ-lille3.fr` affichera l'ensemble des pages du site de l'université de Lille 3 présentes dans l'index de Google. Dans le milieu black hat cette commande est utilisée à d'autres fins, notamment pour repérer les pages indexées par Google dont l'extension de nom de domaine est reconnue pour disposer d'un bon trustrank et qui sont susceptibles d'accueillir un commentaire ou un lien vers le site dont les webmasters black hat cherchent à favoriser le positionnement. Les principaux domaines considérés comme « trustés » par Google sont les noms de domaines en `.edu` et en `.gov`. Ces sites disposent par ailleurs, généralement, d'un très bon PageRank. A partir de là, en utilisant par exemple la commande `site :*edu` « *add your link* » sur Google, il est possible de retourner l'ensemble des pages indexées par Google dont l'extension de nom de domaine est en

.edu et qui dans la page, contiennent l'expression « add your link », c'est-à-dire ajouter votre lien⁷⁰.

- la commande « `inurl:` » : la commande `inurl` permet de rechercher des mots-clés présents dans l'URL d'une page web. Utilisée de manière black hat, cette commande peut être notamment utile pour identifier les sites qui disposent d'un outil de statistiques tel que AWStats, et sur lesquels il sera donc aisé de pratiquer le spam de referer dont nous avons parlé précédemment. Sur l'exemple ci-dessous, les deux derniers résultats sont des sites qui ont installé l'outil AWStats sur leur serveur. En visitant ces sites avec le referer du site que nous cherchons à positionner, nous obtenons donc un backlink de manière assez simple.



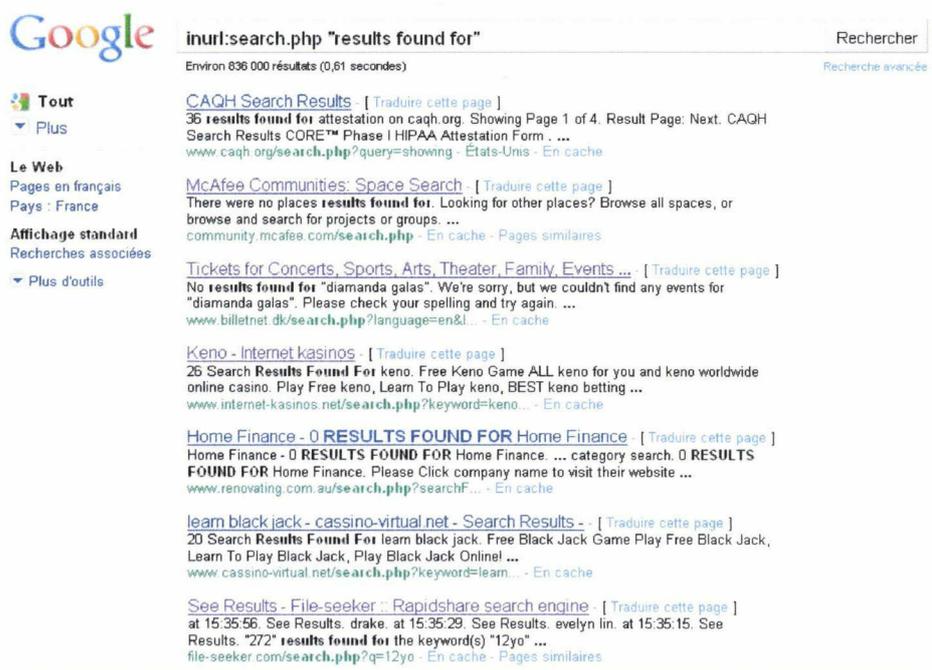
The screenshot shows a Google search for the query `inurl:"awstats.pl"`. The search results page displays approximately 54,500 results. The top results include:

- A forum post titled "Réponse Des GET /stats/awstats.pl dans mes logs apache, entraide ..." with 6 messages and 5 authors, dated 15 July 2005. The post discusses a problem with Apache logs and provides a link to a forum thread.
- A forum post titled ""/cgi-bin/awstats.pl 404 NOT FOUND", par Norbert Q. - Gandi Docs" dated 15 February 2009, discussing a 404 error and providing a link to a forum thread.
- A forum post titled "Statistics of www.bqto.unexpo.edu.ve - [Traduire cette page]" dated 29 August 2010, providing statistics for the website.
- A forum post titled "awstats.pl and xmlrpc.php exploit attempts - HostGator Peer ..." dated 9 November 2005, discussing exploit attempts.
- A forum post titled "Error: Couldn't open config file 'awstats.latino.logosdictionary.org.conf'" dated 15 February 2009, discussing a configuration error.
- A forum post titled "month - AWStats Totals" dated 15 February 2009, providing statistics for the website.
- A forum post titled "quite populaire - www.freeantennas.com - [Traduire cette page]" dated 15 February 2009, providing statistics for the website.

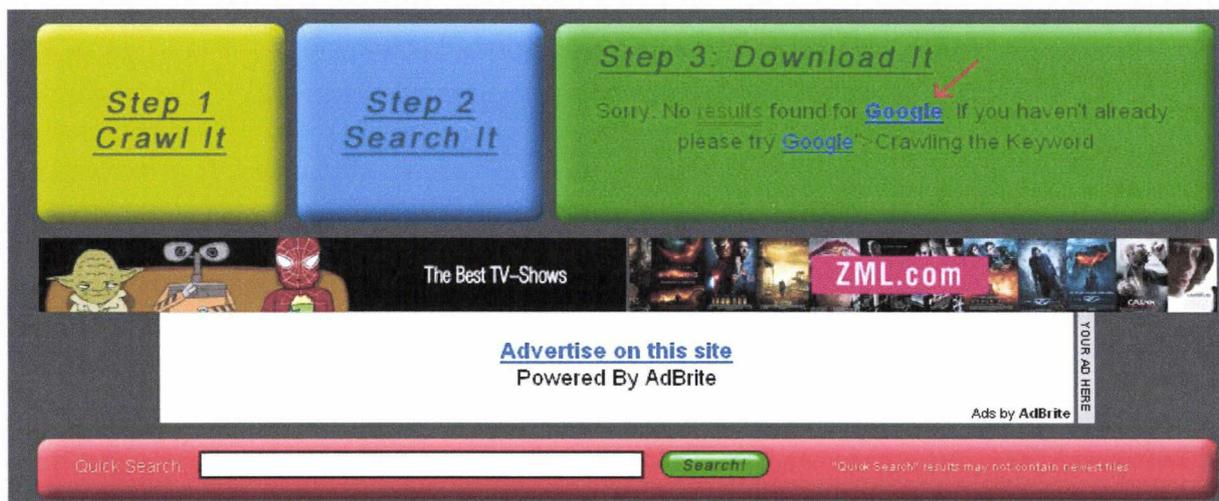
Autre exemple avec la commande `inurl:"edu/forum/profile.php"` : en saisissant cette requête dans Google, le moteur de recherche affiche l'ensemble des liens vers les profils des membres de forums de sites en .edu. Il suffit alors de s'inscrire sur un de ces forums puis d'ajouter un lien en signature de profil vers le site de notre choix.

⁷⁰ SEOBLOCKOUT, *SMX Paris 2010 : Introduction aux techniques de linkbuilding borderline*, 21-06-2010. <http://www.seoblockout.com/2010/06/21/smx-paris-2010/>

- **commandes combinées** : parfois les black hat se servent de commandes plus complexes pour repérer des failles XSS ou des « empreintes » de CMS susceptibles d’être réappropriées pour le spam :
 - **failles XSS** (Cross Site Scripting) : il s’agit de failles de sécurité des sites web qui passent par l’injection de données dans un site via les paramètres d’URL. Si ces données sont interprétées par les navigateurs, alors c’est qu’il existe une faille sur le site. En matière de black hat, ce sont les moteurs de recherche internes des sites web qui sont visés. En saisissant par exemple : `inurl:search.php "results found for"` dans la barre de recherche Google, nous avons accès à une liste de liens vers les pages de résultats des moteurs de recherche internes de certains sites web. Pour être bref, l’astuce consiste à entrer dans le champ de recherche un lien html du type `Mot-clé` et à vérifier si le module de recherche interprète le lien, sans filtrer les variables. Si les caractères HTML sont interprétés, cela signifie que le site ne s’est pas protégé via la fonction `htmlentities`⁷¹. Il suffit alors de faire quelques liens vers cette page pour que Google en ait connaissance et indexe le backlink créé.



⁷¹ `Htmlentities` : fonction permettant de convertir une chaîne de caractères en entités HTML, empêchant ainsi un code HTML d’être activé



- **empreintes** : la majeure partie des CMS open source laissent des empreintes ou « footprints » caractéristiques. Sachant qu'ils sont utilisés par des millions de webmasters dans le monde, l'identification de ces empreintes peut être potentiellement utilisée dans le cadre du spam. Par exemple, la requête "Powered by BlogEngine.NET" "add comment" permet de relever l'ensemble des sites qui utilisent la plateforme BlogEngine.NET et de trouver directement les pages sur lesquelles poser nos liens. L'empreinte « Powered by BlogEngine.NET » est en effet le texte présent dans le footer du CMS par défaut (voir ci-contre).

Powered by BlogEngine.NET 1.6.1.0 © Copyright 2007 - 2010
[Subscribe](#) | [License](#) | [Contact](#) | [Log in](#)

Comme nous pouvons le voir, les black hat utilisent les propres outils de Google pour arriver à leurs fins. Google n'étant au final qu'une machine à indexer du texte, c'est sa fonction première qui est détournée dans le sens d'une pratique black hat du SEO. Abordons à présent une autre question, liée à l'influence de Google dans le web.

3.2.2. Optimisation d'un site : pour Google ou les internautes ?

Aujourd'hui, au vu des contraintes imposées par Google et par les géants de la recherche pour qu'un site internet acquière une bonne place parmi les résultats de recherche, nous pouvons nous demander dans quelle mesure les sites web sont créés pour les visiteurs, et dans quelle mesure ils sont créés uniquement pour « plaire » à Google. Est-ce que Google n'est pas en train de créer un système qui encourage le spamdexing par la toute puissance de ses critères de classement ?

Dans un article intitulé « *J'écris pour Google* »⁷², Sylvain Peyronnet avance l'hypothèse ou plutôt le constat que sur de nombreux sites web, nous ne créons pas du texte pour qu'il soit lu, mais pour qu'il soit « trouvé ». Il se trouve que les webmasters qui s'intéressent de près au référencement ne sont plus « libres de [leur] prose » et qu'ils réfléchissent d'abord en termes de mots-clés et d'indexation avant de penser au confort de lecture de leurs visiteurs. Ils se sont résolus à fournir à Google un contenu « formaté » grâce auquel ils auront toutes les chances de se retrouver en bonne place dans les SERP. Ainsi, pour Jean-Marc Hardy⁷³, les textes insérés dans certains sites web ne sont pas destinés en priorité à être lus. Leur présence se justifie par les impératifs du référencement. Parfois, ces textes se présentent sous la forme de longs pavés, écrits avec une petite taille de police et un faible interligne. De quoi décourager la lecture et diriger le regard des lecteurs vers ce qui doit être mis en avant : un encart de contact, un formulaire de réservation, une publicité, etc.

Ainsi, l'inventivité et la création artistique sont écrasées par l'omniprésence des critères du SEO : pas trop d'animations Flash dans les pages, ni de menus en full flash, pas trop d'images, du texte en dur avant tout, etc. La pertinence humaine d'un site, son aspect visuel ou son originalité, sont autant de points qui sont ignorés par la technologie aveugle de Google. Sébastien Billard se pose également la question⁷⁴ de savoir si Google n'est pas en train de « pourrir le web » mais il est moins catégorique dans sa réponse, en évoquant notamment le fait que Google ne fait que reprendre à son compte les standards définis par

⁷² PEYRONNET, Sylvain, *J'écris pour Google*, 25/07/2010. <http://blog.axe-net.fr/j-ecris-pour-google/>

⁷³ HARDY, Jean-Marc, *Ces textes destinés à ne pas être lus*, 20-05-2010.

<http://blog.60questions.net/index.php/2010/05/20/373-ces-textes-qui-sont-faits-pour-ne-surtout-pas-etre-lus>

⁷⁴ BILLARD, Sébastien, *Google dégueulasse-t-il le web ?*, 21-07-2010.

<http://s.billard.free.fr/referencement/?2010/07/21/616-google-degueulasse-t-il-le-web>

le W3C et les « bonnes pratiques en matière d'accessibilité » : un site où les attributs alt sont remplis et où les technologies utilisées sont basiques, va de paire avec une utilisation optimale du web par les personnes handicapées ou par celles qui disposent d'un équipement sommaire pour surfer sur le web.

Enfin, nous pouvons prendre le problème du spamdexing dans l'autre sens, et se demander si Google, lorsqu'il a introduit le linking au cœur de son algorithme de classement des pages web, n'a pas tout simplement ouvert la voie aux pratiques black hat, et ainsi entraîné une pollution inévitable de son index. C'est notamment la position soutenue par l'auteur du blog Renarddudezert⁷⁵ qui voit dans l'introduction du concept de PageRank il y a quelques années, l'élément déclencheur du spamdexing à grande échelle. Selon lui, un rouage aussi important de l'algorithme de Google n'aurait pas dû être communiqué au grand public et Google aurait dû continuer à appliquer sa politique du secret et, au final, ne chercher à satisfaire que les utilisateurs de son moteur de recherche sans donner autant d'indications aux webmasters. La naissance des linksfarm, du content spinning ou des logiciels d'automatisation est selon l'auteur du blog une conséquence inévitable des annonces successives de Google sur l'importance des liens entrants dans le milieu du référencement. Dans ce contexte, il est donc regrettable de constater que les webmasters n'agissent plus avec spontanéité mais qu'ils doivent inévitablement penser à Google avant de lancer un site.

Comme nous pouvons le remarquer, et même si cela est sujet à polémique, Google a d'une certaine manière, par son évolution, contribué à l'augmentation du spam. Pour disposer du meilleur classement possible et passer devant les concurrents, outre le spamdexing, il existe aujourd'hui des méthodes totalement contraires à l'éthique que nous regroupons sous le nom de « negative SEO ».

3.2.3. Dérives du référencement et negative SEO

Le negative SEO est un ensemble de pratiques qui visent à faire descendre voire à supprimer un site concurrent des pages de résultats des moteurs de recherche, en tentant de le rendre

⁷⁵ RDD, *Comment Google en voulant assainir le web a développé l'effet inverse*, 08-07-2010.
<http://www.renarddudezert.com/2010/07/08/google-et-le-declin-des-bonnes-pratique.html>

moins crédible aux yeux desdits moteurs de recherche et qu'il subisse des pénalités. Ces méthodes sont donc bien entendues tout à fait déloyales et peuvent faire l'objet de poursuites judiciaires dans certains cas. Voici un tour d'horizon des principales techniques utilisées :

- **302 hijacking** : il s'agit de créer de toute pièce un site web qui reprend le contenu d'un autre site concurrent, mais qui est redirigé via une redirection 302 (temporaire) vers le site du concurrent. Si sur une seule page du site créé, on n'implémente pas de redirection 302, alors il est possible que Google vienne indexer cette page et qu'il supprime des SERP la page du site concurrent pour cause de duplicate content. Google a en effet longtemps eu du mal à gérer les redirections 302, notamment avec les annuaires. Certains annuaires, en effet, utilisent des redirections 302 à des fins de statistiques. Parfois ces annuaires sont positionnés devant le site lui-même, car Google considérerait que la page vers laquelle pointe le lien n'est que temporaire⁷⁶ et qu'elle ne doit pas être affichée dans les SERP.
- **Mauvais voisinage** : la technique consiste à faire quelques liens vers le site du concurrent sur des mots-clés ayant pour thématique la pornographie, le poker ou tout autre expression pour laquelle les filtres de Google sont les plus actifs. Au final, le site concurrent risque donc d'être filtré par la technologie SafeSearch de Google et de ne plus apparaître en bonne place dans les résultats de recherche. Il est également possible de réaliser des liens depuis des sites blacklistés.
- **Suppression de backlinks** : cette méthode est beaucoup plus simple mais non moins efficace. Elle consiste à se faire passer pour l'administrateur du site web du concurrent et à demander par mail la suppression des backlinks créés depuis des sites web partenaires. Le site concurrent perdra ainsi une partie de son « linkjuice » et se verra décrédibilisé aux yeux des moteurs de recherche (du moins en ce qui concerne le critère de popularité). Il est aussi possible de modifier directement les liens des sites concurrents soumis dans les

⁷⁶ BILLARD, Sébastien, *Update Allegra : Google a des ratés*, 09-02-2005.

<http://s.billard.free.fr/referencement/index.php?2005/02/09/36-update-allegra-google-a-des-rates>

annuaires en envoyant un e-mail à l'administrateur de l'annuaire et en lui demandant de remplacer le lien par un lien vers notre site.

- **Duplicate content** : cette technique repose sur la duplication de contenu et consiste à créer des splogs où sera effectué un simple copier/coller du contenu des pages des concurrents, de façon à ce que Google détecte le duplicate content et applique un filtre sur l'URL originale du concurrent. Une autre manière de s'y prendre peut être de passer par des proxies pour visiter le site du concurrent et de s'arranger pour que ces URL « proxifiées » soient indexées par Google. Ainsi, un même contenu sera accessible via différentes URLs aux yeux de Google et sanctionné pour cause de duplicate content.
- **Echange de lien unilatéral** : il est possible lors d'un échange de liens de faire croire au webmaster avec qui nous procédons à l'échange que le lien a bien été inséré sur le site, qu'il est « en dur » mais d'indiquer aux robots de ne pas suivre de lien. Pour cela, il faut utiliser la directive X-Robots-Tag dans l'en-tête HTTP⁷⁷ et spécifier la valeur nofollow. Le lien n'aura alors aucune valeur en termes de référencement. Il faut donc penser à vérifier les en-têtes HTTP lors d'un échange de lien.

```
HTTP/1.1 200 OK
Date: Mon, 14 Sep 2009 15:25:07 GMT
Server: Apache
X-Powered-By: PHP/5.2.6-1+lenny3
X-Robots-Tag : nofollow
Test : HEAD
Vary: Accept-Encoding
Content-Type: text/html; charset=ISO
```

Fig.28 : directive http X-Robots-Tag

Ces différents moyens de pénaliser un site concurrent sont clairement à rapprocher du SEO Black Hat. Ils sont réalisés au mépris de l'éthique et engagent la responsabilité de ceux qui la

⁷⁷TASSEL, Olivier, *Manipuler Googlebot avec la directive HTTP X-Robots-Tag*, 11-02-2010. <http://www.olivier-tassel.fr/x-robots-tag>

pratiquent, notamment au niveau juridique, avec, en cas de duplicate content volontaire, la menace du droit de la propriété intellectuelle.

Conclusion

Alors qu'au début du web et des moteurs de recherche, l'optimisation d'un site pouvait très bien être assurée par un webmaster, il se trouve qu'aujourd'hui la discipline a évolué et que la charge de travail a considérablement augmenté. Google, pour ne citer que lui, réajuste son algorithme à intervalles réguliers, introduit de nouveaux critères d'optimisation et met à jour ses guidelines, ce qui entraîne des changements perpétuels dans la manière d'aborder le référencement. La discipline est donc devenue une activité professionnelle à part entière. Cependant, au fil des années, nous pouvons constater que la stratégie de base du SEO n'a pas sensiblement bougé : pour qu'un site dispose d'un bon ranking, il lui faut un contenu optimisé et des liens entrants.

Or, l'acquisition de liens effectuée de manière naturelle, autrement dit de façon white hat, est un processus extrêmement chronophage. Pour obtenir des résultats satisfaisants, notamment dans les secteurs très concurrentiels, plusieurs mois voire des années sont parfois nécessaires. Pour contourner cette difficulté, un ensemble de techniques dites de black hat SEO ont émergé. Ces techniques permettent de positionner de manière rapide un site web parmi les premiers résultats de recherche, en manipulant essentiellement son linkbuilding à grands coups d'automatisation. Néanmoins, leur effet est généralement de courte durée car les moteurs de recherche disposent d'équipes anti spam qui veillent à maintenir un index propre. Le SEO a ainsi évolué en fonction des déclarations de Google : aujourd'hui le bourrage de mots-clés dans les meta keywords, le texte caché ou encore le cloaking sont proscrits alors qu'à l'arrivée sur le marché des moteurs de recherche, elles suffisaient à assurer un positionnement performant.

Nous avons mis le doigt dans ce mémoire sur la fragilité du référencement orienté black hat : en effet, une technique black hat n'est jamais viable indéfiniment et les adeptes du black hat sont sans cesse à la recherche de nouvelles failles dans l'algorithme des moteurs de recherche. Avec l'arrivée imminente du HTML 5, la dernière révision du langage HTML dont la spécification devrait être terminée fin 2010 et l'introduction de nouvelles balises sémantiques, nous pouvons présager des bouleversements dans le référencement. Du côté du black hat, les webmasters seront certainement amenés à s'adapter. Nous pouvons ainsi

nous demander si Google, qui mise beaucoup sur le HTML 5 (en témoignent les différents sites de promotion du nouveau langage qu'il promeut), peut faire de cette technologie son nouvel allié dans la chasse au spam.

Bibliographie

▪ Ouvrages

- **ANDRIEU, Olivier**, *Réussir son référencement web*, 2^{ème} édition, Eyrolles : Paris, 2009, 442 page. ISBN : 978 2 212 12646 4
- **CHU, Nicolas**, *Réussir un projet de site web*, 4^{ème} édition, Eyrolles : Paris, 2006. 244 pages. ISBN : 2 212 11974 7
- **ESKENAZI, Jean-Pierre**, *Référencement, comment référencer son site web*, Webedition : Versailles, 1999. ISBN : 2 9512348 1 3
- **GREGOIRE Gilles**, *Le référencement sur Google, Le Guide Complet*, Micro Application, Paris, 2008, 304 pages. ISBN : 978 2 300 01320 1
- **MICHELI Régis** et **ALBERICI Pascal**, *Les clés du référencement sur le web, 5 étapes pour développer votre visibilité*, BOD, Paris, 2009, 236 pages. ISBN : 978 2 810 61096 9
- **WARBESSON, Karine**, *Créez votre site web, le guide complet*, 2^{ème} édition, Micro Application : Paris, 2007. ISBN : 978 2 300 01123 8

▪ Articles sur Internet

- **404 Creation**, *Qype, une des plus belles réussites du référencement...BlackHat ?*, [En ligne]. Disponible à l'adresse suivante :
<http://www.404-creation.com/referencement/qype-referencement-blackhat.php>
- **512 Banque**, *Outil de spam referer (gentil)*, 08-12-2008. [En ligne]. Disponible à l'adresse suivante :
<http://www.deliciouscadaver.com/outil-de-spam-referer-genti.html>
- **512Banque**, « *Le web scraping ou comment piller les sites à la volée* », 12-06-2009. [En ligne]. Disponible à l'adresse suivante :
<http://www.deliciouscadaver.com/le-web-scraping-ou-comment-piller-les-sites-a-la-volee.html>
- **BARDON, Aurélien**, *Test de la balise meta description*, 27-08-2009: [En ligne]. Disponible à l'adresse suivante :

<http://www.laboratoire-referencement.fr/balise-meta-description.php>

- ✕ **BILLARD, Sébastien**, *Google dégueulasse-t-il le web ?*, 21-07-2010. [En ligne].

Disponible à l'adresse suivante :

<http://s.billard.free.fr/referencement/?2010/07/21/616-google-degueulasse-t-il-le-web>

- ✕ **BILLARD, Sébastien**, *Google dégueulasse-t-il le web?*, 27-07-2010. [En ligne].

Disponible à l'adresse suivante :

<http://s.billard.free.fr/referencement/?2010/07/21/616-google-degueulasse-t-il-le-web>

- **BILLARD, Sébastien**, *Update Allegra : Google a des ratés*, 09-02-2005. [En ligne].

Disponible à l'adresse suivante :

<http://s.billard.free.fr/referencement/index.php?2005/02/09/36-update-allegra-google-a-des-rates>

- ✕ **BOURRELLY, Laurent**, *Ne tirez pas sur le référencement black hat*, 18-09-2009. [En ligne]. Disponible à l'adresse suivante :

<http://www.laurentburrelly.com/blog/237.php>

- ✕ **BOURRELLY, Laurent**, *Ne tirez pas sur le référencement black hat*, 18-09-2007. [En ligne]. Disponible à l'adresse suivante :

<http://www.laurentburrelly.com/blog/237.php>

- **CLEVE, Virginie**, *SEO : les gros sites peuvent-ils tout se permettre ?*, 28-06-2010.

[En ligne]. Disponible à l'adresse suivante :

<http://www.cafe-referencement.com/lectures/seo-les-gros-sites-peuvent-ils-tout-se-permettre-269>

- **CROUZILLACQ, Philippe**, *Le groupe 3 Suisses assigne l'Afnic dans une affaire de typosquatting*, 20-07-2007. [En ligne]. Disponible à l'adresse suivante :

<http://www.01net.com/editorial/355270/le-groupe-3-suisses-assigne-lafnic-dans-une-affaire-de-typosquatting/?forum=355270&post=129171>

- **CUTTS, Matt**, *Calling for link spam report*, 03-03-2010. [En ligne]. Disponible à l'adresse suivante :

<http://www.mattcutts.com/blog/calling-for-link-spam-reports/>

- **DIMEGLIO, Arnaud**, *Le droit du spamdexing*, 27-01-2004. [En ligne]. Disponible à l'adresse suivante :
<http://www.journaldunet.com/juridique/juridique040127.shtml>
- **Discodog**, *The Xrumer effect ce n'est pas l'outil qui fait le moine*, 14-06-2010. [En ligne]. Disponible à l'adresse suivante :
<http://www.discodog.fr/the-xrumer-effect-ce-nest-pas-loutil-qui-fait-le-moine.html>
- **DUFFEZ, Olivier**, *Google a répertorié 1000 milliards de pages web*, 25 juillet 2008. [En ligne]. Disponible à l'adresse suivante :
<http://www.webrankinfo.com/actualites/200807-1000-milliards-de-pages-sur-le-web.htm>
- **DUFFEZ, Olivier**, *Google officialise son infrastructure Caffeine*, 09-06-2010. [En ligne]. Disponible à l'adresse suivante :
<http://www.webrankinfo.com/dossiers/indexation/caffeine>
- **HARDY, Jean-Marc**, *Ces textes destinés à ne pas être lus*, 20-05-2010. [En ligne]. Disponible à l'adresse suivante :
<http://blog.60questions.net/index.php/2010/05/20/373-ces-textes-qui-sont-faits-pour-ne-surtout-pas-etre-lus>
- **LACREUSE, Alex**, *Découvrez le pourcentage de Français qui utilisent Google*, 20-10-2009. [En ligne]. Disponible à l'adresse suivante :
<http://www.lepost.fr/article/2009/10/20/1751299-decouvrez-le-pourcentage-d-internautes-francais-qui-utilisent-google.html>
- **Pagasa**, *Injection d'URL*, 19-12-07. [En ligne]. Disponible à l'adresse suivante :
<http://www.pagasa.net/injection-durl/>
- **PEYRONNET, Sylvain**, *Les MFA | Made For Adsense | sont-ils tous des pollueurs ?*, 04-10-2009. [En ligne]. Disponible à l'adresse suivante :
<http://blog.axe-net.fr/les-mfa-made-for-adsense-sont-ils-tous-des-pollueurs/>
- **PEYRONNET, Sylvain**, *Black hat or white hat le SEO*, 05-07-2009. [En ligne]. Disponible à l'adresse suivante :
<http://blog.axe-net.fr/seo-black-hat-ou-white-hat/>

- **PEYRONNET, Sylvain**, *Comment spammer un blog dofollow ?*, 22-11-2009. [En ligne]. Disponible à l'adresse suivante :
<http://blog.axe-net.fr/comment-spammer-un-blog-dofollow/>
- **PEYRONNET, Sylvain**, *Crédit, sexe, viagra, poker, soyez patients !*, 2010. [En ligne]. Disponible à l'adresse suivante :
<http://blog.axe-net.fr/credit-viagra-sexe-poker-soyez-patient/>
- **PEYRONNET, Sylvain**, *J'écris pour Google*, 25-07-2010. [En ligne]. Disponible à l'adresse suivante :
<http://blog.axe-net.fr/j-ecris-pour-google/>
- **PEYRONNET, Sylvain**, *SEO : black hat ou white hat ?*, 05-07-2009. [En ligne]. Disponible à l'adresse suivante :
<http://blog.axe-net.fr/seo-black-hat-ou-white-hat/>.
- **RANFISH**, *4 essential SEO infographics*, 10-08-2009. [En ligne]. Disponible à l'adresse suivante :
<http://www.seomoz.org/blog/4-essential-seo-infographics>
- **RANFISH**, *I'm getting more worried about the effectiveness of webspam*, 17-08-2010. [En ligne]. Disponible à l'adresse suivante :
<http://www.seomoz.org/blog/im-getting-more-worried-about-the-effectiveness-of-webspam>
- **RANFISH**, *Online Poker – Too competitive for white hat SEO ?*, 18-02-2007. [En ligne]. Disponible à l'adresse suivante :
<http://www.seomoz.org/blog/online-poker-too-competitive-for-white-hat-seo>
- **RDD**, *Comment Google en voulant assainir le web a développé l'effet inverse*, 08-07-2010. [En ligne]. Disponible à l'adresse suivante :
<http://www.renardudezert.com/2010/07/08/google-et-le-declin-des-bonnes-pratique.html>
- **SEO Black Out**, *SMX Paris 2010 : Introduction aux techniques de linkbuilding borderline*, 21-06-2010. [En ligne]. Disponible à l'adresse suivante :
<http://www.seoblackout.com/2010/06/21/smx-paris-2010/>
- **SEO Black Out**, *Web spam : le guide SEO spamdexing*, 23-07-2010. [En ligne]. Disponible à l'adresse suivante :

<http://www.seoblackout.com/2010/07/23/web-spam-seo/>

- **SEOPLAYER**, *Les techniques SEO black hat au devant de la scène*, 01-10-2008. [En ligne]. Disponible à l'adresse suivante :
<http://www.seoplayer.com/optimisations-seo/les-techniques-seo-black-hat-au-devant-de-la-scene.html>
- **TASSEL**, Olivier, *Manipuler Googlebot avec la directive HTTP X-Robots-Tag*, 11-02-2010. [En ligne]. Disponible à l'adresse suivante :
<http://www.olivier-tassel.fr/x-robots-tag>

ANNEXES

Table des figures

Figures 1,2,3	<i>p 9</i>
Figure 4	<i>p 14</i>
Figure 5	<i>p 15</i>
Figure 6	<i>p 16</i>
Figures 7	<i>p 23</i>
Figure 8	<i>p 26</i>
Figure 9	<i>p 28</i>
Figure 10	<i>p 34</i>
Figure 11	<i>p 35</i>
Figure 12	<i>p 45</i>
Figure 13	<i>p 46</i>
Figure 14	<i>p 47</i>
Figure 15	<i>p 50</i>
Figure 16	<i>p 51</i>
Figure 17	<i>p 52</i>
Figure 18	<i>p 58</i>
Figure 19	<i>p 61</i>
Figure 20	<i>p 62</i>
Figure 21	<i>p 63</i>
Figure 22	<i>p 64</i>
Figure 23	<i>p 66</i>
Figure 24	<i>p 67</i>
Figure 25,26	<i>p 72</i>
Figure 27	<i>p 74</i>
Figure 28	<i>p 84</i>

Table des annexes

Calcul du PageRank

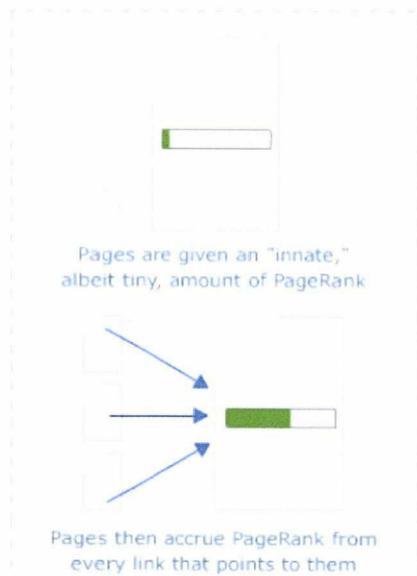
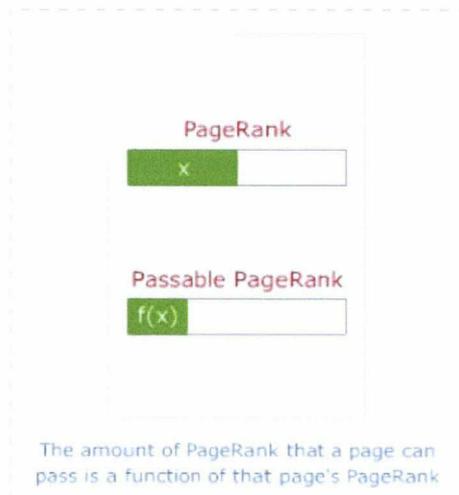
p 65

Distribution du PageRank

p 68

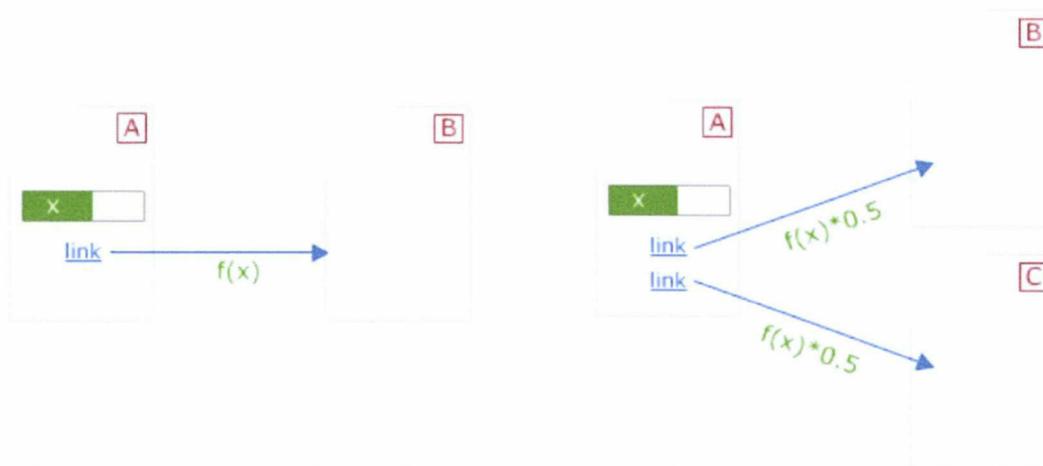


The Flow of PageRank





PageRank is Split Evenly Between the Links on a Page



Résumé

Aujourd'hui et malgré le succès grandissant des réseaux sociaux, les moteurs de recherche conservent une hégémonie incontestable sur la recherche en ligne. Face à une demande qui a explosé, le nombre de sites web a également subi une croissance exponentielle. Les résultats de recherche sont devenus de véritables terrains d'affrontement sur lesquels nous pouvons observer, à l'œuvre, les techniques de référencement les plus avancées. Ce mémoire questionne donc l'utilisation des techniques de référencement interdites par les moteurs (ou « black hat ») et essaie de démêler intérêts, avantages et limites de ces méthodes. Il s'avère tout d'abord que la frontière est bien plus mince qu'elle n'y paraît entre un référencement dit « white hat » et un référencement effectué sur la base de fondations black hat, mais également que « black hat » ne signifie pas forcément contraire à l'éthique. Cependant ces techniques de spamdexing présentent des risques et sont à éviter dans toute agence SEO, qui porte sur elle la responsabilité économique des clients qu'elle accompagne dans leur stratégie web.

Mots-clés : référencement naturel, moteurs de recherche, black hat SEO, white hat SEO, référencement abusif, webspam

Abstract

Today, despite the growing popularity of social networks, search engines keep an unchallenged hegemony on the online search. The demand has exploded and so did the number of websites. The search result pages have become, in a way, a field of battle on which we can observe how the most advanced techniques of SEO are used. Therefore, this training report is about the use of SEO techniques banned by search engines (or "black hat"). It tries to show what the interests, benefits and limitations of these methods are. Firstly, the border is thinner than we could think between SEO called "white hat" and SEO called "black hat". Then, "black hat" does not necessarily mean unethical. However, these spamdexing techniques are risked and should be avoided in every SEO agency. Indeed, business issues are threatened if black hat SEO strategies are revealed.

Keywords : search engine optimization, search engines, black hat SEO, white hat SEO, spamdexing, webspam