Robots.txt, crawl et indexation (1ère partie)

Retour au sommaire de la lettre

Domaine :	Recherche	**Référencement**
Niveau :	Pour tous	**Avancé**

Le fichier robots.txt est un grand classique du Web et du SEO. Pourtant, êtes-vous sûr de l'utiliser à bon escient et de bien comprendre son fonctionnement ? En effet, une utilisation erronée de ce fichier peut amener des soucis d'indexation, voire provoquer des pertes de référencement. Voici donc, dans cet article en deux parties, un état de l'art de la meilleure façon d'utiliser le fichier robots.txt pour mieux contrôler la vision de votre site qu'auront les moteurs de recherche. Vous risquez d'être surpris....

Le fichier robots.txt est l'un des plus anciens outils mis à la disposition des webmasters pour contrôler le comportement des robots d'exploration du web sur leurs sites. On pourrait donc imaginer que le rôle du robots.txt est connu, que la syntaxe de ses directives est maîtrisée, et que l'impact de leur utilisation est évalué correctement. Il n'en est rien...

Dans la pratique, le référenceur rencontre très souvent des robots.txt utilisés à mauvais escient, et même certains cas dans lesquels il peut jouer un rôle très néfaste pour un bon référencement. Et la plupart des erreurs commises à propos du robots.txt tirent leur origine d'une mauvaise interprétation du rôle de ce fichier...

Ne pas confondre "crawl" et "indexation"

L'une des erreurs les plus répandues parmi les webmasters (et, hélas, parmi les "pros" du référencement, les questions posées lors de l'examen CESEO à propos du robots.txt font souvent des dégâts chez les candidats), c'est de confondre "téléchargement d'un contenu" et "indexation".

Les directives d'un fichier robots.txt ont pour objectif unique d'indiquer aux moteurs (en tout cas à ceux qui respectent le protocole robots.txt) que le webmaster ne souhaite pas que certaines URL soient téléchargées. Mais qu'en est-il de leur indexation ?

Où l'on apprend que l'indexation ignore totalement le robots.txt

Lorsqu'un moteur de recherche explore le web, il découvre de nombreuses URL en analysant le contenu des pages web. Ces URL sont à leur tour téléchargées, le "crawler" y découvre de nouvelles URL, et ainsi de suite.

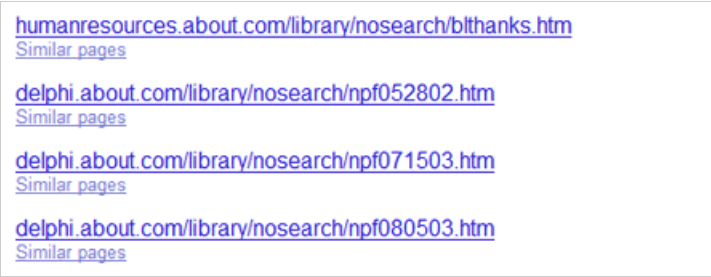
Si on arrête le processus, on constate que les URL explorées peuvent être classées en deux groupes :
- les URL découvertes et que le robot d'exploration a téléchargées : leur contenu (code source) est donc connu par le moteur de recherche ;
- les URL découvertes, mais dont le contenu n'a pas été téléchargé.

Au risque d'en surprendre certains, ces deux types d'URL figurent dans les pages de résultat d'un moteur...

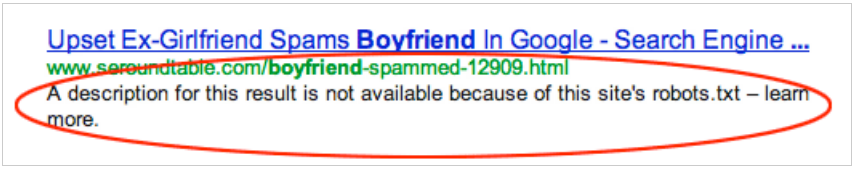
En effet, un moteur de recherche peut parfaitement "indexer" (placer dans sa base de données) une URL dont il ne connait pas le contenu. L'information contenue dans les liens pointant vers celle-ci est suffisamment riche pour "classer" la page, et pour la faire apparaître sur certaines requêtes. Les "textes d'ancre" (*anchor texts*) peuvent en particulier servir à cette fin.

Ce comportement existait dans les plus anciens moteurs, qui ne parvenaient à explorer qu'une petite partie du world wide web, et a été conservée dans les moteurs modernes comme Google.

Bien sûr, comme le contenu de la page n'est pas connu, le moteur n'est pas capable de récupérer un titre et une description pour la page. C'est pourquoi les pages non téléchargées apparaissent dans les résultats de Google sans titre ni description... Mais il peut arriver dans quelques cas particuliers que Google "recrée" un titre et une description à partir d'informations externes au contenu de la page. Par exemple, à ses débuts, le moteur fabriquait un snippet pour les pages d'eBay, dont le téléchargement était bloqué via le fichier robots.txt



Apparence classique de snippets dans les pages de résultats de Google pour des pages bloquées par un robots.txt : le snippet est réduit à une simple URL. Si Google peut « reconstituer » un titre à partir d'autres sources d'informations, le snippet peut éventuellement afficher un titre, une URL, mais la description est absente.



Récemment, un message explicite a fait son apparition dans le cas où l'URL est bloquée. Notez que Google a réussi à attribuer un « titre » à la page dont il ne peut pas lire la balise <title>...

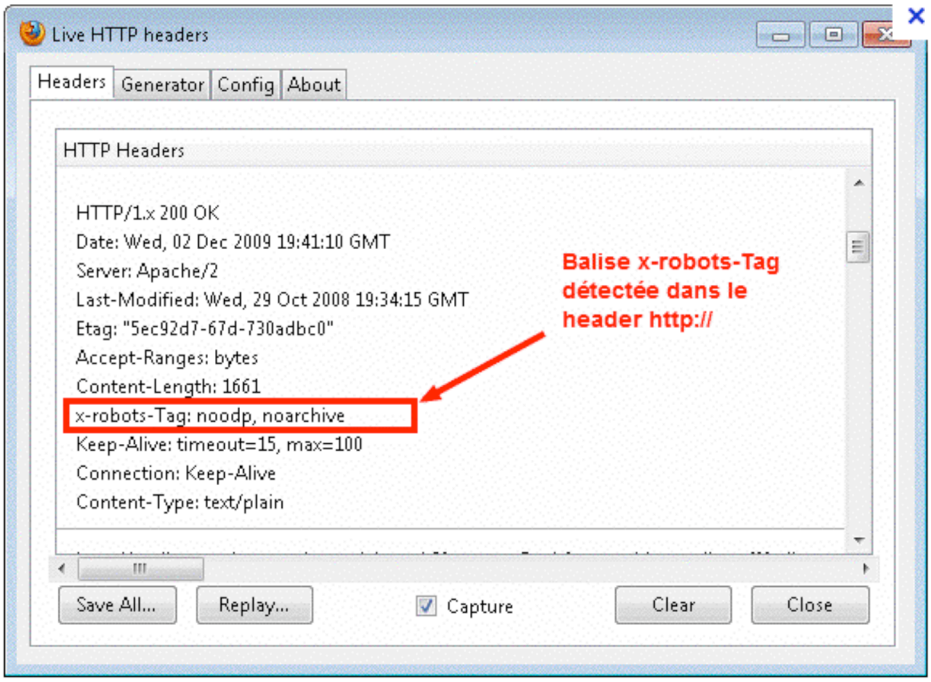
Bref, pour qu'une URL apparaisse dans les pages de résultats, il faut et il suffit que celle-ci ait été découverte, soit au cours de l'exploration, soit en analysant d'autres sources d'information (comme les URL récoltées par les barres d'outils, ou les URL récoltées dans les sitemaps).

Conclusion : interdire le téléchargement du contenu d'une page à un moteur *via* le fichier robots.txt n'a pas pour résultat d'empêcher l'indexation de cette page. Pour cela il existe les directives "noindex".

Utiliser le bon outil pour empêcher l'indexation d'une url : Meta robots noindex et x-robots-tag

La seule solution efficace pour empêcher l'indexation d'une URL est de placer une balise meta "robots" dans l'entête de la page html avec l'attribut "noindex". Habituellement, quand on fait des formations à des "pros" du référencement, on trouve toujours un esprit fort à ce stade pour dire : "ok, ok, mais comment je fais pour empêcher l'indexation d'un pdf alors, si ce n'est en plaçant les URL de ces pdf dans un robots.txt" ?

En fait, il existe une solution trop méconnue : la directive x-robots-tag, supportée par les grands moteurs (Bing, Ask, Google), et qui se place... dans l'en-tête HTTP. X-robots-tag fonctionne exactement comme la meta "robots", sauf qu'elle permet d'indiquer au moteur de ne pas indexer n'importe quel type de contenu (flash, ppt, pdf, images, word...), et pas uniquement des pages html.



La balise x-robots-Tag se place dans le header http:// (à ne pas confondre avec l'entête «<head>...</head> » d'une page html). Elle accepte les valeurs d'attribut classiques pour la balise meta robots : nofollow, noindex, noarchive etc...

Si vous êtes attentifs, vous aurez noté que le moteur ne peut pas découvrir la balise meta robots, comme la directive x-robots-tag, si aucune requête http:// n'est envoyée vers cette URL. Ce qui signifie donc que si une url est placée parmi les URL bloquées par un robots.txt, il n'y a par contre aucun moyen de faire savoir au moteur qu'on ne veut pas que cette URL soit indexée...

Cela signifie donc que toute tentative d'empêcher l'indexation d'une URL *via* un robots.txt revient en fait à se tirer une balle dans le pied !

Utiliser un robots.txt pour empêcher l'indexation d'URL est pourtant un réflexe courant chez pas mal de webmasters... Ce sont souvent les mêmes, par méconnaissance du mécanisme d'indexation, qui accusent Google de ne pas respecter le protocole robots.txt, et en argumentant en montrant que les URL incriminées sont bel et bien indexées. Le fait qu'elles soient indexées ne prouve rien, puisque que nous avons vu qu'une URL pouvait être indexée sans que le contenu ait été téléchargé... Pour prouver qu'un moteur ne respecte pas le fichier robots.txt, il faut consulter les logs et constater que Bingbot ou Googlebot a effectivement téléchargé une URL en principe bloquée.

Mais, il arrive de temps en temps que l'on identifie ce genre de bizarreries...

Les cas où Google peut réellement télécharger des URL bloquées

Car il existe réellement quelques cas qui peuvent conduire un moteur à ne pas respecter une directive dans un robots.txt. Il peut s'agir, soit d'erreurs subtiles commises par le webmaster et qui empêchent le robots.txt de remplir son office, soit d'erreurs de conception du moteur Google (si si, il y en a...).

Le robots.txt : un fichier qui n'a pas d'effets immédiats

Contrairement à une idée fortement répandue, les moteurs ne consultent pas le robots.txt à chaque requête. Ils le consultent régulièrement, à un rythme propre à chaque site et qui dépend du rythme d'exploration des URL de ce site par le robot.

Par conséquent, il est important de modifier le robots.txt AVANT de créer une famille d'URL bloquées par le robots.txt. Si le fichier est créé en même temps, voire après, les URL en question seront téléchargées, et leur contenu indexé. Notons néanmoins que ces URL ne seront pas recrawlées après la lecture de la nouvelle version du robots.txt, et que les contenus téléchargés seront éliminés de l'index à la prochaine mise à jour des informations, donc l'impact de ces ratés initiaux est souvent minime.

Un fichier qui doit être disponible en permanence

Même si le fichier n'est pas consulté à chaque requête, il est consulté souvent, et il est donc particulièrement important que ce fichier soit accessible en permanence. Or parfois, pour des raisons techniques, le fichier peut ne pas être disponible et renvoyer une erreur 500 par exemple. Ces problèmes sont depuis peu signalés par Google *via* un message envoyé aux contacts identifiés dans Google Webmaster Tools. Il convient de prendre ces messages particulièrement au sérieux, car un trop grand nombre d'erreurs constatées lors de l'interrogation du fichier robots.txt finit par se traduire par une baisse du crawl...


wmt-noreply@google.com Sep 9

to me

Au cours des 24 dernières heures, Googlebot a rencontré 57 erreurs en tentant d'accéder au fichier robots.txt. Pour veiller à ne pas explorer les pages répertoriées dans ce fichier, nous avons reporté notre exploration. Le taux d'erreurs liées au fichier robots.txt à l'échelle de votre site est de 8.7 %. Vous trouverez davantage d'informations détaillées sur ces erreurs dans les Outils pour les webmasters.

Exemple de message d'alerte à propos de l'accès au robots.txt envoyé par Google depuis un compte GWT.

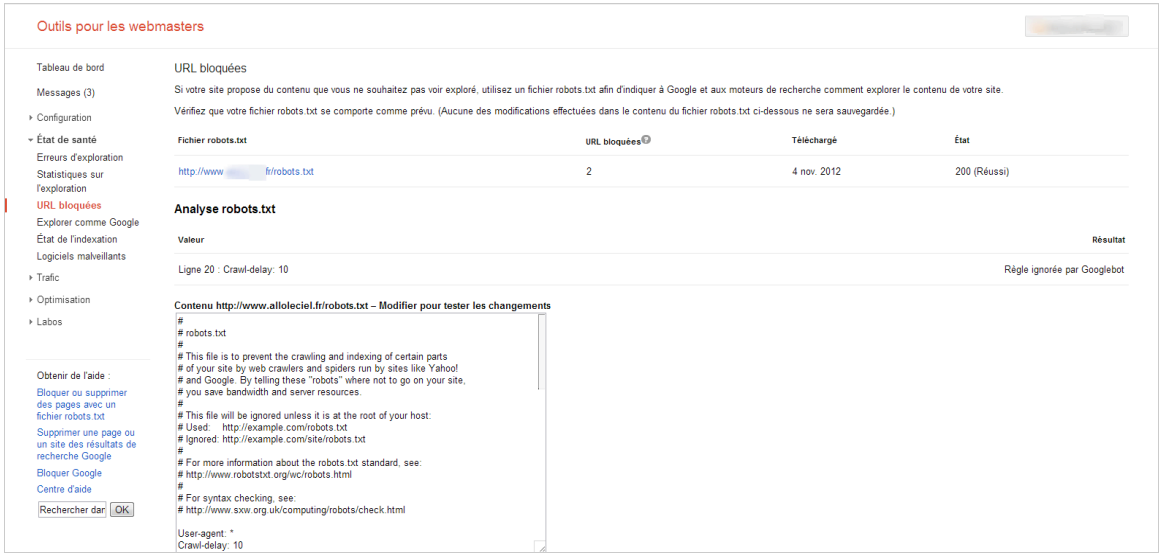
Les erreurs de syntaxe les plus communes

Si l'on constate qu'une directive de robots.txt n'est pas respectée et que des URL bloquées sont téléchargées, il faut avoir le réflexe de vérifier la syntaxe de son fichier robots.txt. L'une des erreurs les plus fréquentes est la présence de lignes blanches après la ligne indiquant le user-agent concerné par certaines directives. Si la présence de lignes blanches entre les blocs destiné à un user-agent particulier sont autorisées, ce n'est pas dans les blocs eux-mêmes. Le résultat de cette ligne blanche sera que la suite du bloc sera confondue avec un nouveau bloc dont le user agent ne sera pas reconnu. En fonction des moteurs, au mieux, seules les lignes du bloc qui suivent la ligne blanche seront ignorées, au pire cela affectera tous les blocs suivants, voire même la compréhension de tout le fichier robots.txt.

```
User-agent: *                    User-agent: *
# Directories
Disallow: /includes/             Disallow: /includes/
Disallow: /misc/                 Disallow: /misc/
Disallow: /modules/              Disallow: /modules/
Disallow: /profiles/             Disallow: /profiles/
Disallow: /scripts/              Disallow: /scripts/
Disallow: /themes/               Disallow: /themes/
```

Dans le bloc de droite, la ligne de commentaire est remplacée par une ligne blanche. La suite est considérée comme un bloc de directives indépendant. Le robots.txt n'est plus valide et peut entrainer un comportement inattendu de la part des moteurs de recherche

Ces lignes sont parfois invisibles dans certains éditeurs de textes donc il peut être intéressant de tester la syntaxe à l'aide :
- soit de l'outil *ad hoc* des Google Webmaster Tools ;
- soit de l'outil http://tool.motoricerca.info/robots-checker.phtml qui a le mérite de détecter des problèmes pour tous les robots d'exploration et pas seulement pour Googlebot (il existe d'autres outils du même genre mais c'est le plus fiable que nous connaissions).



L'outil de vérification du robots.txt fourni par Google dans ses webmasters tools permet de vérifier que la syntaxe de son fichier est bien reconnue par Googlebot. Mais attention : certaines syntaxes ne fonctionnent qu'avec Googlebot, et il faut penser à vérifier que son fichier est conforme aux standards des autres moteurs. Ici la ligne 20 est détectée comme inutile.

L'autre erreur fréquente est de créer des robots.txt dans lesquels les instructions sont contradictoires. Par exemple, une URL est en *allo/* pour tous les *user agents*, et se retrouve en *disallow* pour un *user agent* particulier.

Il faut savoir que les moteurs de recherche tiennent toujours compte du bloc de directives le plus spécifique. Si vous avez créé un bloc pour Googlebot, les directives pour le *user-agent* * seront ignorées !

Avec ce genre de syntaxe, il faut maîtriser parfaitement la hiérarchie de prise en compte par les robots des instructions imbriquées.

```
user-agent: googlebot-news
(group 1)

user-agent: *
(group 2)

user-agent: googlebot
(group 3)
```

*Pour le crawler Googlebot-news, seul le premier groupe de directives sera pris en compte. Pour le crawler Googlebot-images non mentionné dans le fichier, c'est le bloc de directives « * » qui sera utilisé. Pour le bot news, toute directive présente dans * et non reproduite dans la section googlebot-news sera ignorée.*

URL	allow:	disallow:	Verdict	Comments
http://example.com/page	/p	/	allow	
http://example.com/folder/page	/folder/	/folder	allow	
http://example.com/page.htm	/page	/*.htm	undefined	
http://example.com/	/$	/	allow	
http://example.com/page.htm	/$	/	disallow	

Hiérarchie entre directives du même groupe : Exemples d'ordre de prise en compte de directives imbriquées et éventuellement contradictoires : la règle la plus spécifique (mesurée

par la longueur du chemin qui la définit) a la priorité sur les autres. Pour les règles utilisant des caractères jokers, l'ordre de priorité n'est pas défini.

Parmi les erreurs habituelles, signalons aussi les erreurs de casse : les directives des robots.txt sont sensibles à la casse, et tout oubli d'une majuscule peut conduire à des résultats inattendus .

Quand Googlebot déraille...

Il existe un cas pour lesquels on peut constater que Google semble ignorer accidentellement les instructions du fichier robots.txt. Il semble qu'il s'agisse d'un effet de bord de ses nouveaux modes d'exploration du web dont Google n'a jamais confirmé l'existence.

Google explore toutes les sources d'URL à sa disposition : liens dans les fichiers flashs, liens dans les Powerpoint, documents Word, Pdf., etc. Mais surtout, Google analyse les scripts Javascript pour y découvrir des liens. Mieux, depuis quelque temps, il execute les Javascripts pour voir si le script ne téléporte pas l'internaute sur une autre page. Or dans un contexte précis, ce dernier comportement peut conduire Google à zapper des URL bloquées par un robots.txt.

Imaginons un annuaire qui redirige vers des pages externes, mais ne souhaite pas que du "jus de lien" soit renvoyé vers les sites référencés. Le webmaster de l'annuaire crée une redirection en Javascript du type "*refresh zéro delay*" pour une meilleure expérience utilisateur, et code un comportement différent pour les navigateurs dans lesquels le Javascript est désactivé (classiquement, c'est un comportement codé pour les moteurs de recherche). Le clic sur un *item* de l'annuaire appelle dans ce contexte une url destinée à compter les clics, puis *via* une redirection 301 une URL interne qui gère la redirection 301 vers la page externe.

La deuxième url est bloquée (théoriquement) par le robots.txt.

Dans un tel contexte, Google aura tendance à voir dans le comportement du Javascript un lien direct entre la page de l'annuaire et l'URL de la page externe. Ce chemin ne passe pas par une URL bloquée par le robots.txt, et il crawle un lien qu'il n'aurait pas du découvrir. Plus fort, il peut arriver que Google associe le contenu de la page externe avec l'URL de comptage, et dans les pages de résultats, le contenu de la page de destination, reconnaissable à sa description et son titre, est indexé sous une URL interne de l'annuaire ! (C'est ce comportement qui est utilisable en black hat, car il permet de s'approprier du contenu externe en le faisant passer pour son propre contenu. Il s'agit d'un nouvel avatar du hacking de contenu *via* des redirections. Notez bien qu'utiliser cette faille est clairement contraire aux TOS de Google).

Conclusion : Google respecte scrupuleusement les instructions du robots.txt, sauf quand on lui complique la vie en créant des chemins différents passant soit par une URL bloquée par un robots.txt, soit par une redirection différente... Notons que d'autres cas similaires ont été rapportés par le passé...

Robots.txt et Bot Herding : Caffeine en a définitivement supprimé le besoin

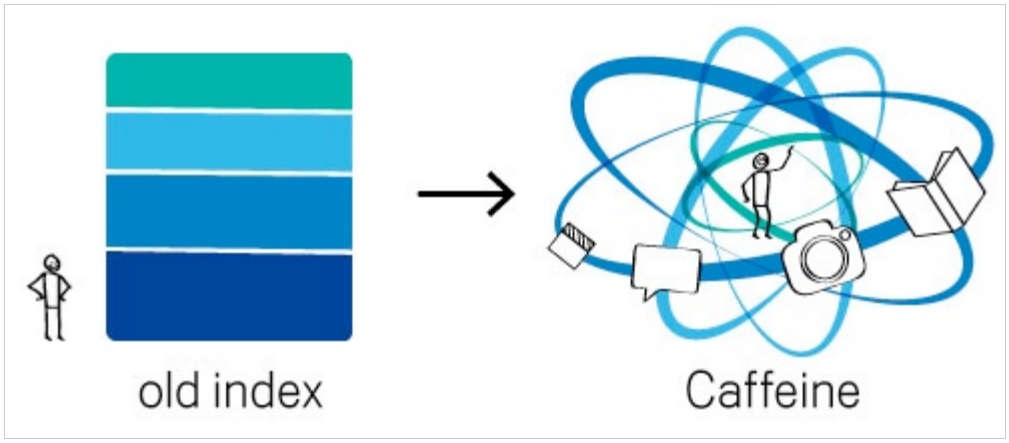
Depuis de nombreuses années, certains SEO s'ingénient à trouver des solutions pour que Google crawle la totalité des pages utiles d'un site. Dans ses premières années de fonctionnement, les capacités de crawl de Google étaient limitées et certaines pages étaient réellement "ignorées" par manque de ressources. Puis Google est passé à une logique d'exploration "ouverte" (le crawl ne s'arrête jamais) et une logique de "priorité" de crawl, qui laissait toujours des pages non crawlées.

Certains SEO ont donc mis au point des techniques dites de "bot herding", dont l'objectif était de faire aller les robots d'exploration en priorité sur les pages jugées "utiles" par les SEO. Comme pour le "PR Sculpting", ces techniques passaient (entre autres) par une logique de "blocage" de l'accès à certaines pages, *via* le cryptage de liens et le blocage de groupes d'URL *via* un robots.txt.

Ce qui est étonnant, c'est que la plupart de ces techniques, qui peuvent réellement changer la donne sur de gros sites comportant des millions d'URL s'appuyaient sur la croyance que Google attribuait une limite de pages crawlées à un site donné, et que substituer parmi ces pages crawlées des pages utiles à des pages de mauvaise qualité était une bonne idée... En réalité, cette "limite de crawl" ou "réserve de crawl" comme l'appelle certains n'existe pas : c'est un effet de bord des règles de priorité qui régissent le comportement de Googlebot. En réalité, le nombre de pages crawlées et le rythme de recrawl est déterminé par des indicateurs (comme le PageRank ou la fréquence de renouvellement du contenu), et il suffit de faire bouger ces indicateurs (par exemple en jouant sur la profondeur des pages) pour changer totalement le comportement de crawl de Google sur ces pages.

A indicateurs constants, l'une des solutions pour permettre le crawl de pages ignorées était de "sculpter" la structure du site vue par Googlebot en bloquant les URL "indésirables" avec un robots.txt. Nous verrons dans un prochain article les dégâts que ce genre d'approche peut occasionner quand on n'en maîtrise pas l'impact. Néanmoins cette approche pouvait s'avérer efficace dans certains cas, en changeant drastiquement le maillage interne.

Ces techniques de bot herding ont néanmoins perdu de leur intérêt depuis 2007 avec BigDaddy, et surtout depuis 2010 avec Caffeine. Le passage à l'infrastructure surnommée BigDaddy a été l'occasion de perfectionner le crawl, et de passer à système de crawl en "couches" recrawlées à des rythmes différents, en fonction de critères plus sophistiqués qu'auparavant. La mise en place plus récemment de l'infrastructure Caffeine a rendu le comportement de Googlebot beaucoup plus intelligent et réactif qu'auparavant, et l'infrastructure de Google lui permet maintenant de ne plus avoir de réelles limitations de crawl. Concrètement, en 2012, si une page est ignorée par les crawlers, c'est probablement parce que Google a volontairement choisi de ne pas la crawler, et le fait d'empêcher Google de crawler d'autres pages ne change plus directement son comportement (indirectement oui, car cela peut faire bouger certains indicateurs).



Une illustration fournie par Google évoquant le changement de comportement de crawl intervenu par Caffeine. Avant, les URL étaient réparties entre différentes couches, caractérisées par des comportements de crawl différent. Avec le déploiement de l'infrastructure de Caffeine, Google crawle de petits groupes d'URL, plus intelligemment et de manière plus réactive.

Bloquer des URL visibles par les internautes dans un robots.txt ne sert à rien (sauf exception)

Si nous résumons les points évoqués ci-dessus, il s'avère donc que :

- utiliser le robots.txt pour contrôler l'indexation ne fonctionne pas bien ;

- utiliser le robots.txt à des fins de bot herding ou de PR sculpting est devenu franchement inefficace quand il n'est pas contre-productif.

Parmi les utilisations courantes, on peut ajouter aussi l'élimination du DUST (*Duplicate URL / Same Text*) et des doublons d'URL. Mais on retombe là sur les mêmes remarques que celles faites à propos du contrôle de l'indexation : cela ne fonctionne pas bien, et cela se règle de manière plus sûre et plus élégante à l'aide des nouveaux outils fournis par les moteurs : la balise "canonical" d'une part, et la gestion des paramètres dans les outils pour webmasters d'autre part. Et ces deux dernières méthodes supposent que les robots d'exploration puissent crawler les pages en question... donc une fois de plus, bloquer les doublons *via* le robots.txt n'est pas une bonne idée.

D'une manière générale, en 2012, mieux vaut laisser Googlebot crawler ce qu'il veut et comme il veut, et intervenir ensuite si quelque chose ne se passe pas bien. Les deux seules catégories de cas où ce "laisser faire" est contre indiqué sont :

- les cas où il existe une différence massive entre les URL crawlées par Google et les "vraies" URL du site => interdire certaines syntaxes peut clairement améliorer la situation ;

- et bien sûr les cas où ne veut pas que les moteurs puissent accéder au contenu !

Or on trouve toujours des répertoires entiers correspondant à des pages accessibles aux internautes bloqués dans les fichiers robots.txt d'un grand nombre de sites...

Conclusion

Nous venons de le voir, le robots.txt est souvent utilisé à mauvais escient à des fins de référencement :

- c'est un outil inadapté pour contrôler son indexation ;
- d'autres méthodes sont plus appropriées pour limiter les effets de bords du DUST et du duplicate content ;
- et l'intérêt de recours au robots.txt pour faire du bot herding a beaucoup diminué depuis Caffeine.

Nous verrons de plus dans un prochain article qu'y placer des URL appartenant à l'arborescence normale de son site peut résulter en une perte massive de Pagerank interne... Bref, cet outil qui date des premiers temps du Web reste particulièrement puissant, mais s'avère mal compris et mal maîtrisé par beaucoup de webmasters. Il faut donc revenir aux fondamentaux, bien vérifier le fonctionnement de son fichier robots.txt, et l'analyser ligne par ligne.

En particulier, si vous voyez sur le robots.txt de l'un de vos sites, des lignes qui bloquent des URL faisant partie de l'arborescence normale du site, à savoir des pages que des internautes peuvent découvrir en naviguant sur votre site, il convient de se poser quelques questions. Interrogez-vous en particulier pour savoir si le motif de blocage est bien légitime, ou si vous n'êtes pas en train d'utiliser la mauvaise technique pour contrôler le comportement du moteur...

Références

Site (quasi) officiel sur le protocole robots.txt et les standards associés :
http://www.robotstxt.org/

La page de référence de Google à propos des balises meta robots et x-robots-tag :
https://developers.google.com/webmasters/control-crawl-index/docs/robots_meta_tag

Les pages de référence de Google à propos du fichier robots.txt et des extensions du standard propres à Google :

http://googlewebmastercentral.blogspot.fr/2008/06/improving-on-robots-exclusion-protocol.html

http://support.google.com/webmasters/bin/answer.py?hl=fr&answer=156449&from=40367&rd=1

https://developers.google.com/webmasters/control-crawl-index/docs/faq

Outil de validation de syntaxe de robots.txt :
http://tool.motoricerca.info/robots-checker.phtml

Philippe YONNET , ***Directeur de l'agence Search-Foresight / Groupe MyMedia. Président de l'association SEO Camp (http://www.seo-camp.org/)***