Les indicateurs sur l'état de l'indexation : une nouveauté utile des Google Webmaster Tools

Retour au sommaire de la lettre

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

En juillet dernier, les Google Webmaster Tools ont ajouté un nouvel indicateur à leurs outils : des courbes donnant le nombre de pages d'un site indexées par le moteur de recherche sur les 12 derniers mois. Cet outil, à l'usage, s'avère très intéressant si on analyse finement les chiffres fournis. Voici un décryptage des différentes informations données par cet outil et la meilleure façon de les utiliser pour votre référencement...

Le 24 juillet 2012, une nouvelle fonctionnalité est apparue dans les "Outils pour les webmasters" de Google : "l'index status" ou "état de l'indexation" en Français (http://www.abondance.com/actualites/20120726-11721-les-google-webmaster-tools-donnent-des-stats-sur-lindexation-dun-site.html). Google fournit désormais sous la forme d'un graphique l'historique du nombre de pages indexées dans le moteur. Cette information a été peu commentée et peu relevée, sans doute parce qu'elle est tombée pendant l'été, mais aussi parce que beaucoup ont considéré cette fonctionnalité des comptes GWT comme un gadget.

En réalité, les référenceurs expérimentés savent à quel point il est utile de suivre le nombre de pages indexées. En effet cette donnée peut permettre d'identifier plus facilement l'origine d'un problème de référencement, et de monitorer l'impact des changements apportés à un site.

Dans cet article, nous allons commencer par expliquer comment "lire" ces nouvelles données fournies par Google, et nous allons ensuite expliquer comment leur interprétation peut aider à établir un diagnostic précis d'un problème de référencement. Nous terminerons en expliquant comment confronter ces données à d'autres indicateurs pour en tirer des enseignements encore plus intéressants pour améliorer son référencement.

Jusqu'ici, comment obtenait-on le nombre de pages indexées d'un site ?

Rappelons d'abord ce qu'est une "page indexée" : il s'agit d'une URL susceptible d'apparaître dans les pages de résultats de Google, parce qu'elle a été ajoutée à la "base de données" du moteur (alias « l'index »).

Avant le 24 juillet dernier, les méthodes pour connaître le nombre de pages indexées était loin d'être parfaites.

Connaître le nombre de pages indexées avec site:

Pour connaître le nombre de pages indexées de son site, la méthode la plus facile d'accès est de relever le nombre de résultats renvoyés par la commande "site:" suivie du nom de domaine à étudier. Le problème est que cette méthode est très peu fiable : la commande "site:" est imprécise (les chiffres sont arrondis) et elle renvoie régulièrement n'importe quoi... Elle donne de plus des résultats différents en fonction des datacenters interrogés !

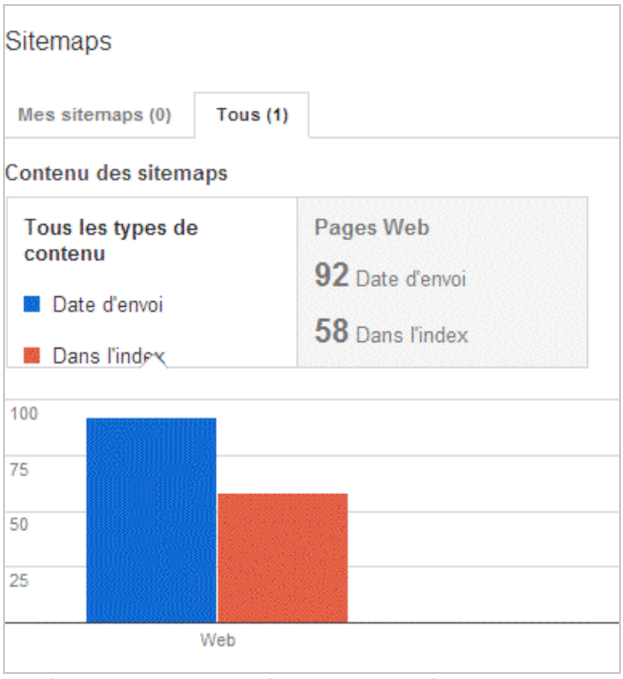
Exemple sur un site de test :
- Nombre de pages indexées fourni par GWT : 59 391
- Nombre de pages indexées fourni par "site:" sur le datacenter 1 : 58 400
- Nombre de pages indexées fourni par "site:" sur le datacenter 2 : 135 000 (???)

La méthode des sitemaps

Une autre méthode théoriquement plus précise consiste à regarder les statistiques fournies sur l'indexation des fichiers Sitemaps. Si on part du principe que le Sitemap contient 100% des

URL du site, Google donnant le nombre d'URL du Sitemap indexées à l'unité près, on peut en déduire un chiffre précis pour le nombre de pages indexées.

Mais l'hypothèse de départ est fausse : rien ne garantit jamais que la liste des URL contenues dans le sitemap recouvre parfaitement la liste des URL que Google considère comme crawlable et indexable. Il en résulte qu'en déduire le nombre de pages indexées à partir de l'indexation des Sitemaps peut conduire à ne pas voir un réel problème d'indexation, ou, si le contenu du Sitemap est inadapté, à voir des problèmes d'indexation là où il n'y a en réalité qu'un problème de sitemaps à corriger.

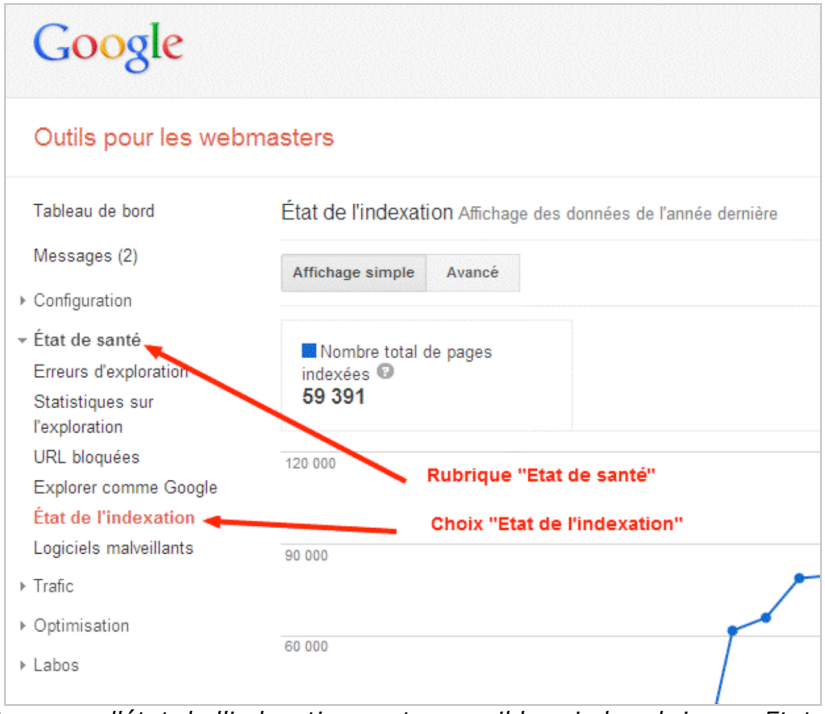


Le taux d'indexation de ce sitemap est de 63%. Mais le site comporte en réalité près de 100 000 URL, pas 92 ! La méthode des Sitemaps, pour fonctionner, demande des Sitemaps exhaustifs, et sans erreur, ce qui n'est pas toujours le cas.

Quelles informations trouve-t-on dans "l'état de l'indexation" ?

Pour découvrir ces nouvelles données, il faut aller dans la rubrique "Etat de santé" de votre compte Google Webmaster Tools, et choisir l'entrée "Etat de l'indexation".

Si vous ne disposez pas encore d'un compte GWT, rappelons que c'est un outil indispensable pour tout webmaster et pour tout référenceur : au fil du temps, Google a inclus dans ce site tant d'informations utiles dans ce compte que s'en priver pour travailler son référencement revient à rouler la nuit sans allumer ses phares !

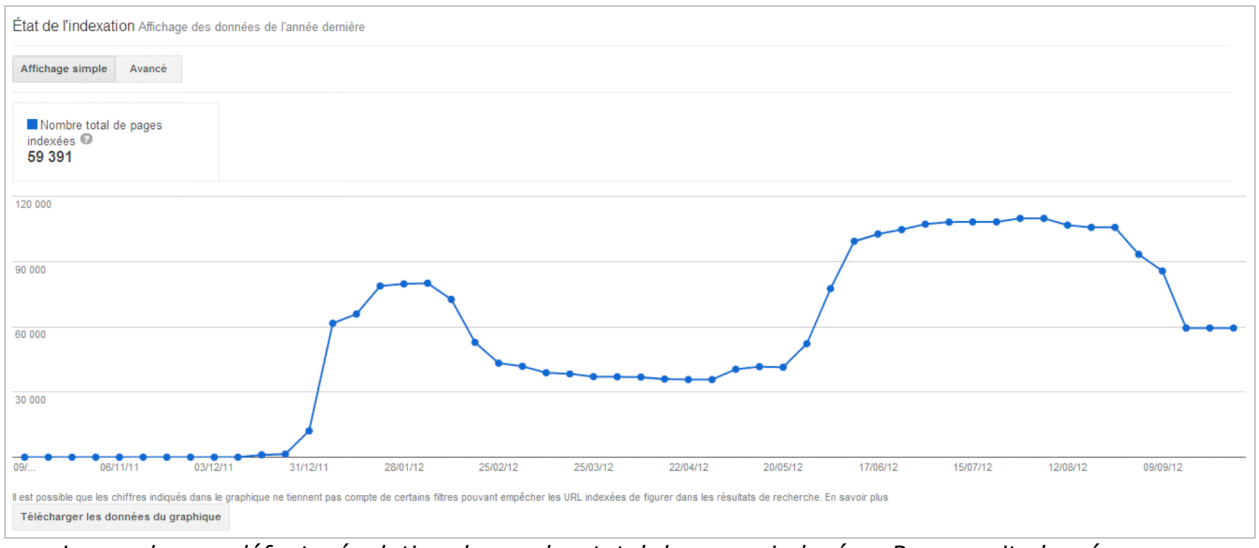


Les indicateurs sur l'état de l'indexation sont accessibles via la rubrique « Etat de Santé » de son compte « Outils pour les webmasters »

Par défaut, le graphique présente un an d'historique du nombre de pages indexées. Chaque point du graphique est espacé d'une semaine : il n'est donc pas possible d'identifier la date exacte à laquelle un changement est intervenu. Mais cette échelle de temps est suffisamment réduite pour identifier une relation de cause à effet entre un changement sur le site et un changement dans l'indexation.

Si vous souhaitez conserver les données, ou les réexploiter, il est possible de télécharger les données du graphique au format CSV (ou de les charger dans Google Docs).

Les données sont fournies au niveau "host". Rappelons que l'on peut créer un compte GWT pour un domaine, un sous-domaine, et même un sous répertoire (cette dernière possibilité est souvent méconnue). Le niveau "host" signifie que l'on peut avoir des données pour le domaine et chaque sous domaine, mais pas pour un sous répertoire.



La courbe par défaut : évolution du nombre total de pages indexées. Pour ce site lancé récemment, on voit que cet indicateur suit une courbe en montagnes russes, alors que le potentiel de contenu du site est en réalité stable autour de 100 000 pages indexables !

Les données fournies par le mode avancé

Si on clique sur le bouton gris "*Avancé*" en haut et à gauche du graphe, d'autres indicateurs plus précis deviennent accessibles.

En mode avancé, Google fournit d'autres données particulièrement utiles :
- les pages bloquées par le fichier robots.txt ;
- le nombre total de pages indexées auparavant ;
- les pages supprimées ;
- les pages non sélectionnées.

Les pages bloquées par le fichier robots.txt.

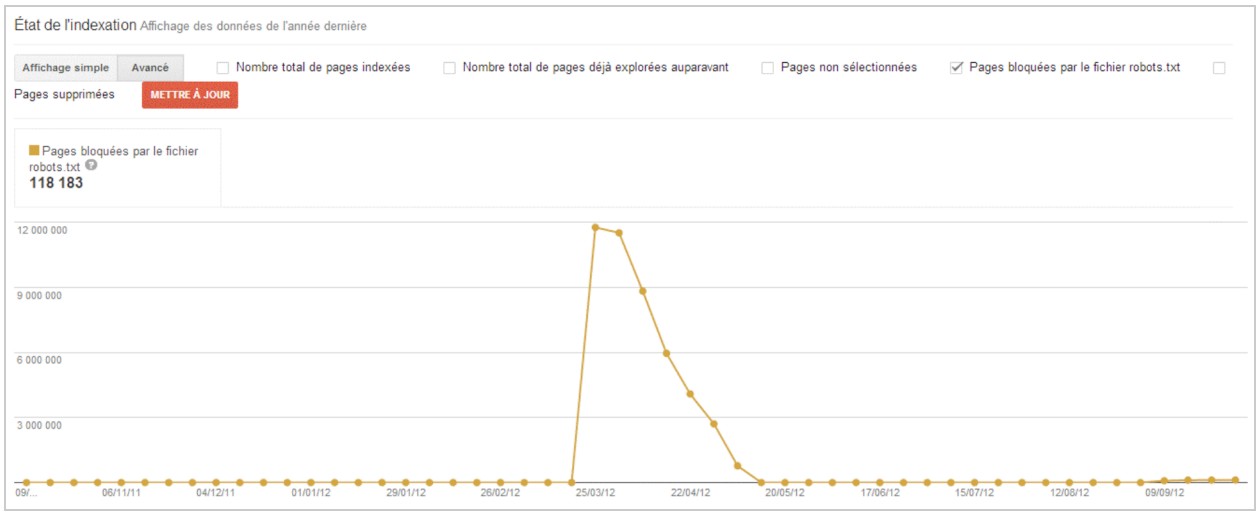
Le graphique donne l'évolution du nombre d'URL que Google a pu découvrir pendant son exploration du web (c'est-à-dire au cours du crawl de Googlebot), et que Google n'a pas pu télécharger à cause d'une ligne du fichier robots.txt

Rappelons au passage deux choses :

1. **Une url bloquée par une ligne du robots.txt peut être indexée** ! Elle apparaîtra sans titre, et sans description dans les pages de résultat, mais le robots.txt, contrairement à une légende tenace, ne permet pas de bloquer l'indexation d'une page, mais uniquement le téléchargement de son contenu par Google.

2. Par conséquent, **bloquer un grand nombre d'URL via un robots.txt sur un site conduit souvent à des pertes de pagerank interne importantes**, selon un mécanisme que Matt Cutts lui-même a exposé dans une interview réalisée par Eric Enge en 2007 (*voir la référence dans la bibliographie*).

Surveiller l'évolution du nombre de pages bloquées par un robots.txt peut permettre de détecter un problème grave pour le référencement d'un site (comme dans l'exemple ci-dessous) :



Un problème technique a engendré pour ce site un grand nombre d'URL découvertes par Google, et bloquées par un robots.txt.

Le nombre total de pages explorées auparavant

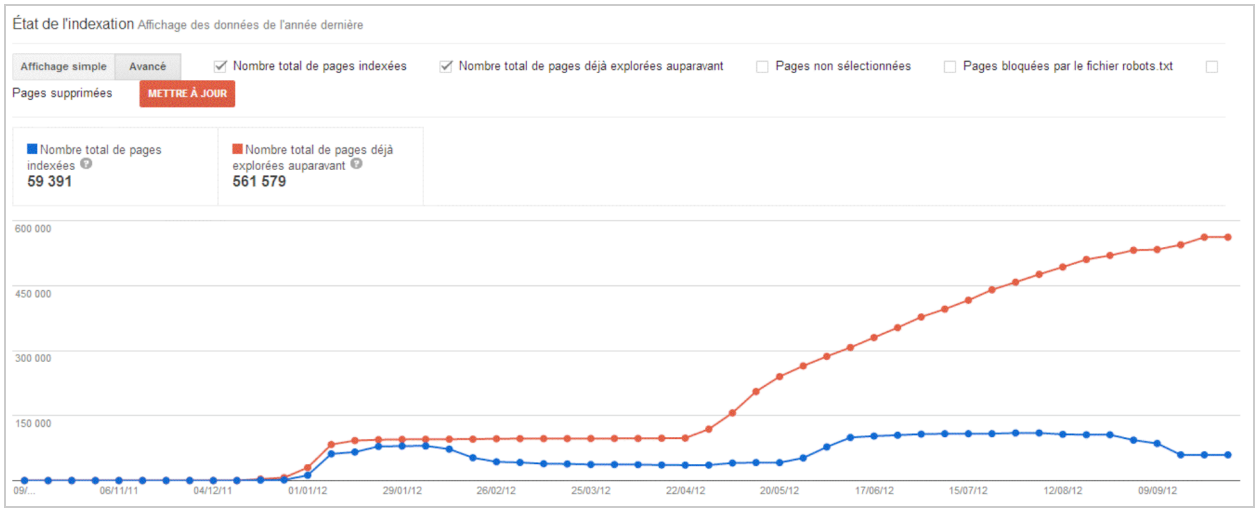
Il s'agit du nombre cumulé des URL que Google a crawlées auparavant. Ce nombre a par nature vocation à croître au fil du temps, et à demeurer bien supérieur au nombre de pages indexées.

En effet, beaucoup de facteurs peuvent conduire une page crawlée à ne pas être indexée, par exemple :
- si la page est un doublon, elle sera écartée, et seule les pages originales seront indexées ;
- si la page est en noindex, elle sera crawlée, mais non indexée ;

- si la page contient un link rel=canonical vers une autre url, elle sera également écartée ;
- si l'url est redirigée (via une redirection 301) vers une autre page, elle ne sera pas indexée.

Dans la pratique, l'information fournie par Google n'est utile que pour des sites dont la liste des URL est stable au cours du temps. Dans ce cas, voir l'évolution comparée du nombre d'URL indexées par rapport au nombre total d'URL explorées permet d'avoir une idée du ratio "*nombre de pages indexées / nombre de pages connues*" et de mesurer ses progrès en indexation (et donc en référencement).

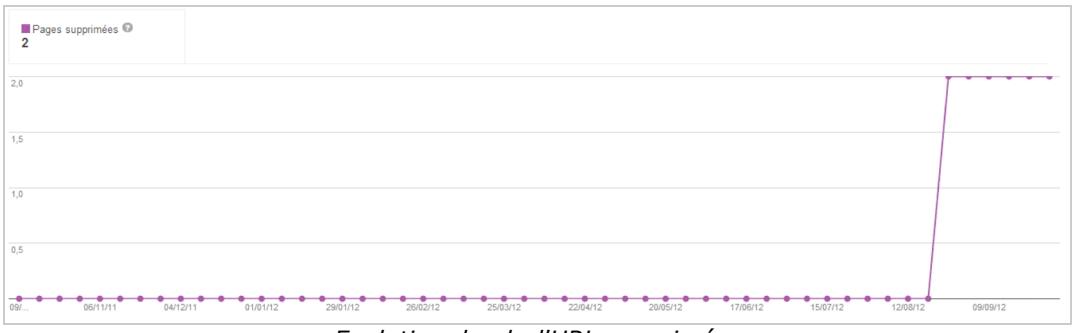
Pour les sites dont la liste des URL change de manière importante et/ou régulière, ce nombre finit par être en déphasage total avec la réalité, car il ne donne aucune idée sur le nombre d'URL que Google continue de crawler parmi celles découvertes dans le passé.



Evolution comparée du nombre de pages indexées et du nombre de pages « explorées auparavant ». Le nombre de pages du site « indexables » est proche des 100 000 en réalité, le nombre des URL découvertes est déjà en déphasage 9 mois après le lancement du site.

Le nombre total de pages supprimées

Il s'agit du nombre d'URL supprimées volontairement suite à une demande de suppression. Ce graphe est apparu récemment et n'est pas documenté dans l'aide, mais il semble que ce graphe indique à la fois les suppressions volontaires ou émanant de tiers (à confirmer).



Evolution du nb d'URL supprimées

Le nombre de pages non sélectionnées

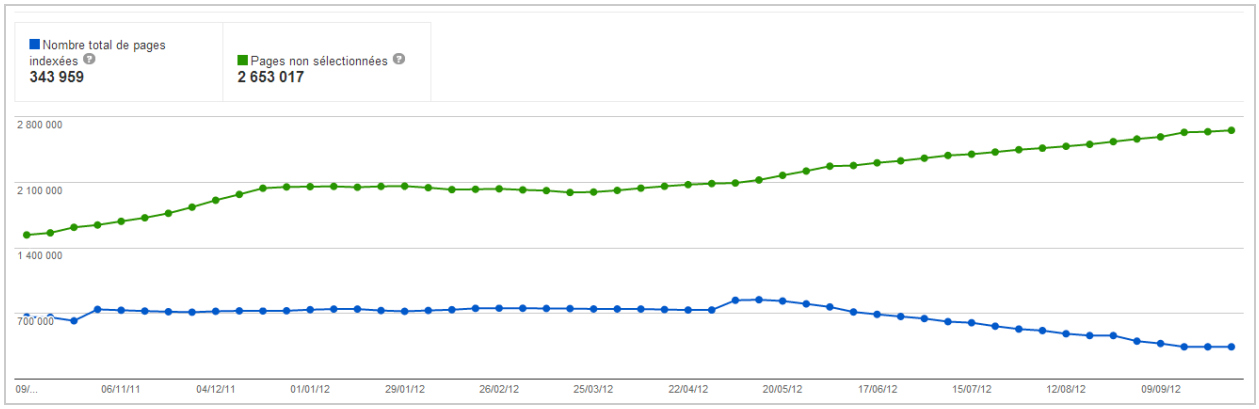
L'évolution du nombre de pages non sélectionnées est sans doute le graphe le plus utile fourni par Google.

En effet, si un grand nombre de pages crawlées ne sont pas "sélectionnées", c'est à dire indexées, on peut perdre un potentiel important pour le référencement.

Le problème c'est que Google ne fournit aucune ventilation sur les "causes" de cette non indexation. On ne sait donc pas si les URL non sélectionnées sont des pages :
- de mauvaise qualité ;

- des doublons ;
- des redirections ;
- des pages en noindex ;
- etc.

Par contre, l'observation régulière de cette courbe peut permettre de détecter un changement dans le nombre de pages non sélectionnées, et de lancer des investigations pour en comprendre l'origine.



Sur ce site, le nombre de pages « non sélectionnées » s'est mis à croître tandis que le nombre d'URL indexées s'est mis à baisser. Si ce transfert n'est pas volontaire (redirections, canonicalization par exemple), il y a lieu de s'inquiéter et rechercher la cause de cette baisse de l'indexation.

Comment utiliser ces données pour améliorer son référencement ?

Prises isolément, ces courbes d'évolution de l'indexation peuvent déjà permettre de diagnostiquer de nombreux problèmes.

1. **Une hausse du nombre de pages bloquées par un robots.txt** conduira à se demander si le phénomène est volontaire ou dû à un problème de gestion des URL.

2. **Une baisse du nombre de pages indexées**, dans un contexte où le nombre d'URL du site est censé rester stable ou même augmenter, conduira à rechercher un problème technique pouvant empêcher ou gêner le crawl ou l'indexation. Un coup d'oeil sur les autres indicateurs de la rubrique "santé du site" du compte GWT peut fournir un début d'explication.

3. **Une hausse de la proportion ou du volume d'URL non sélectionnées**, conduira à rechercher les causes de cette "non indexation".

Mais ces données sont particulièrement intéressantes à rapprocher du nombre d'URL "crawlables" et "indexables" sur un site.

En effet, imaginons que votre site comporte 100 000 pages, toutes crawlables, toutes différentes, et aucune en noindex.

Si votre nombre de pages indexées est passé de 5 000 à 10 000, c'est le signe que votre référencement s'améliore. Mais avec 10 000 pages indexées seulement, 90% du contenu de votre site n'a aucune chance d'être trouvé *via* Google : malgré ces bons chiffres, un problème important subsiste, qui gêne une indexation normale. Si vous arrivez à corriger ce problème, une explosion du trafic moteur est possible !

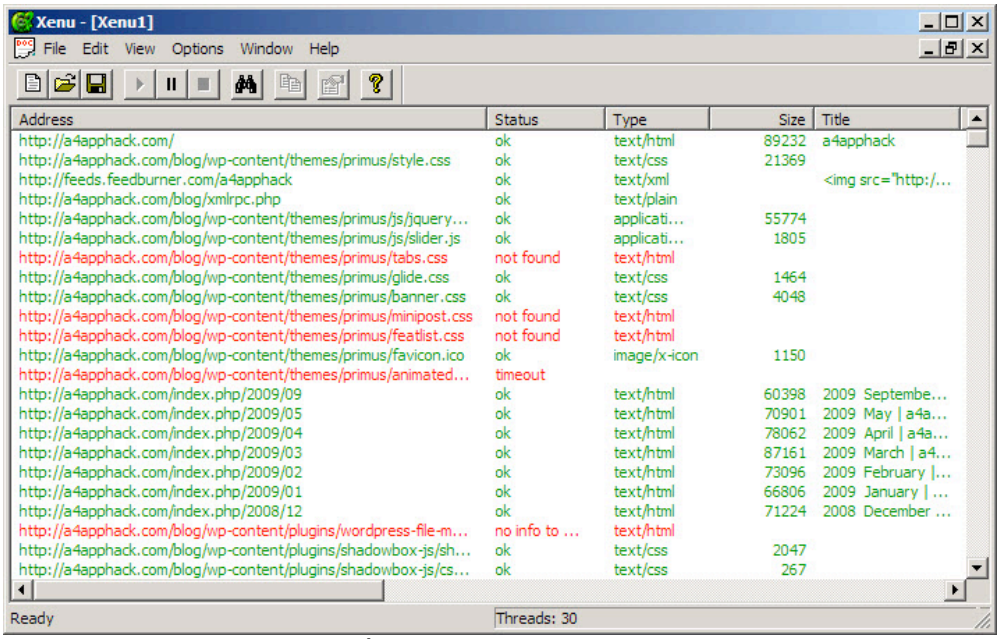
Utiliser un crawler pour déterminer la liste des URL de son site

Pour identifier le nombre de pages "crawlables", le plus simple (et le plus sûr) est d'utiliser un crawler dont le comportement sera proche de Googlebot, le robot d'exploration de Google. Le plus simple d'entre eux est le logiciel gratuit Xenu (*voir le lien en référence ci-dessous*).
Un tel logiciel, en suivant les liens sur les pages explorables, va dresser la liste :
- des pages "crawlables" ;

- des pages qui renvoient des erreurs (500, 404), qui sont par conséquent non indexables ;
- des pages redirigées (erreur 301 ou 302).

Une fois ce travail effectué, on peut travailler sur cet "inventaire instantané" et essayer de déterminer la volumétrie des pages :
- en erreur ;
- redirigées (classées comme "non sélectionnées") ;
- dupliquées ;
- normalement indexables ;
- des pages dont l'URL est présente sur le site mais dont le contenu est un doublon d'une autre page indexable.

Notons au passage que la plupart des crawlers du marché ne savent pas gérer certaines subtilités comme détecter les balises noindex, x-robots-tag, ou les balises "link rel=canonical".



Une copie d'écran de Xenu Link Sleuth en action

Si on a utilisé ces balises ou directives sur le site, il faut soit choisir un crawler plus sophistiqué (mais aucun ne gère parfaitement ces balises et directives, même ceux spécialement conçus pour le SEO), soit ajouter l'information à la liste des "URL à la main", en partant du principe que vous savez sur quelles URL vous avez volontairement ajoutés ces balises (ce qui évidemment préférable, le contraire demanderait que vous changiez d'approche ou de CMS).

Une fois ce travail réalisé, vous allez pouvoir comparer :
- le nombre d'URL indexables au nombre d'url indexées. Toute différence de volumétrie est une perte de potentiel.
- le total du nombre d'URL redirigées + le nombre d'URL canonicalisées par rapport au nombre d'URL non sélectionnées donnée par le compte GWT. La différence de volumétrie correspond aux URL écartées parce que considérées comme des doublons ou des pages de mauvaise qualité.

De nouveaux indicateurs très utiles, mais encore incomplets

Les indicateurs sur l'état de l'indexation représentent donc un effort louable de plus de la part de l'équipe de Google qui gère les outils pour les webmasters pour fournir une meilleure information sur les problèmes qui peuvent gêner ou bloquer le référencement.

Mais les données fournies sont en l'état encore difficiles à utiliser et à interpréter sans faire appel à des données extérieures. Prenons l'exemple des pages "non sélectionnées". On peut comprendre que Google ne veuille pas fournir trop d'informations sur son fonctionnement, et

donc sur la volumétrie de pages considérées comme des doublons. Mais on peut se demander comment un webmaster inexpérimenté va faire la différence, sans ventilation entre les différentes causes de non sélection, entre un fonctionnement normal et désiré de son site et un problème technique !

Même chose du côté du crawl : le volume d'URL "*ever crawled*" ne sert pas à grand-chose, le volume d'URL uniques crawlées depuis un mois serait beaucoup plus utile.

Par contre, un référenceur avancé se voit doter d'un outil de diagnostic très simple et très efficace. Surveiller ces courbes régulièrement permet incontestablement d'être alerté de l'apparition de problèmes techniques qui peuvent impacter gravement le référencement. Et une analyse approfondie régulière (tous les 3 mois) en s'appuyant sur les données issues d'un crawler peut permettre d'identifier éventuellement des chantiers efficaces pour améliorer son référencement.

Bibliographie et liens

L'annonce sur le blog de Google destiné aux webmasters de l'ajout de cette fonctionnalité :
http://googlewebmastercentral.blogspot.fr/2012/07/behold-google-index-secrets-revealed.html

Une interview de Matt Cutts par Eric Enge datant de 2007, dans laquelle Matt Cutts parle (entre autres choses très intéressantes) du problème posé par les URL bloquées par un robots.txt à mauvais escient :
http://www.stonetemple.com/articles/interview-matt-cutts.shtml

Une interview de Matt Cutts par Eric Enge datant de 2010, et contenant un grand nombre d'informations précises sur le crawl et l'indexation :
http://www.stonetemple.com/articles/interview-matt-cutts-012510.shtml

Télécharger Xenu Link Sleuth
http://home.snafu.de/tilman/xenulink.html

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