Comment mieux référencer son site avec Google Sitemaps (1ère partie)

Retour au sommaire de la lettre

Google Sitemaps est un programme initié par Google permettant de mieux référencer un site web en indiquant aux robots du moteur de recherche un "plan du site" au format XML, sous la forme d'un ou de plusieurs fichiers. Dans cette série d'articles, nous indiquons comment utiliser cette fonctionnalité pour obtenir une meilleure visibilité. Ce mois-ci : explication du concept des "Google Sitemaps".

L'ofre Google Sitemaps (https://www.google.com/webmasters/sitemaps/login) est récente puisqu'elle a été lancée par le moteur de recherche en juin dernier (http://actu.abondance.com/2005-23/google-sitemaps.php). Il s'agit d'une solution permettant de fournir aux crawlers de Google (Googlebot) un plan du site au format XML. Les robots peuvent alors identifier et aller chercher toutes les pages qui y sont décrites, selon les indications fournies dans le fichier.

Dans cette série d'articles sur ce sujet, nous allons évoquer plusieurs points :
- Explication du concept et du fonctionnement de Google Sitemaps (ce mois-ci).
- Outils automatisés disponibles pour créer un fichier Sitemaps (le mois prochain).
- Résultats de tests que nous sommes en train de mener, expliquant comme Google prend en compte les fichiers XML (le mois suivant).

Dans un premier temps, il est nécessaire de bien comprendre comment fonctionne le système, assez complet et parfois méconnu dans ses fonctionnalités avancées, proposé par Google...

Le concept de Google Sitemaps

Le concept de l'outil est extrêmement simple : vous créez un fichier XML qui contient la liste des pages de votre site, plus certaines informations sur chacune d'entre elles (fréquence de mise à jour, priorité de crawl, etc.). Vous téléchargez ce fichier sur votre serveur. Vous signalez à Google sa présence. Les robots de ce dernier viennent alors le lire et tiennent compte des données qui y sont proposées pour mieux indexer, plus en profondeur et de façon plus exhaustive, votre site. Simple non ? Encore faut-il que votre fichier soit bien créé, bien soumis et bien placé sur votre site... C'est ce que nous allons explorer tout au long de cette série d'articles...

Notez bien, cependant, que :

- L'utilisation d'un "SiteMap" n'est en rien une garantie que Google indexera TOUTES les pages qui y sont décrites. Le moteur de recherche reste maître de la façon dont il indexe les sites. Mais l'utilisation d'un tel fichier facilite, logiquement, ce processus...

- De même, Google Sitemaps n'est en rien une garantie que votre site sera mieux positionné. Cet outil n'est qu'un outil d'indexation, pas un outil de positionnement ("ranking")...

- Enfin, l'utilisation de Google Sitemaps ne remplace pas le "crawling" classique de votre site par ses robots, suivant les liens des pages web de façon traditionnelle. Les deux méthodes restent tout à fait complémentaires...

Formats du fichier à fournir à l'applicatif

Google Sitemaps reconnaît un certain nombre de formats :

- OAI-PMH (*Open Archives Initiative Protocol for Metadata Harvesting*), format inutilisable cependant pour les sites optimisés pour les mobiles. Peu utilisé, ce format (http://www.openarchives.org/OAI/openarchivesprotocol.html) est proposé par Google uniquement pour les sites utilisant déjà ce standard. Nous n'en parlerons pas ici.

- Fichiers de syndications RSS 2.0 et Atom 0.3, ce qui peut-être très intéressant si votre site utilise déjà ce type de format. Il est tout à fait possible de signaler à Google, via l'interface Google Sitemaps, vos fichiers de syndication.

- Fichiers texte (exemple : www.votresite.com/sitemaps.txt) contenant une adresse de page (url) par ligne. Le fichier ne peut contenir plus de 50 000 lignes maximum mais il est possible de créer plusieurs fichiers...

> **Important :** Nous recommandons l'utilisation du fichier texte si vous désirez uniquement fournir à Google une liste d'urls sans indiquer pour ces pages d'informations connexes (date de dernière modification, priorité d'indexation, fréquence de mise à jour). S'il s'agit juste de fournir au moteur la liste des urls de vos pages de façon "brute", l'utilisation d'un fichier texte (.txt) est certainement la voie la plus simple.
> D'autant plus que ce type de fichier est également pris en compte par Yahoo! si l'on en croit ce qui est indiqué sur sa page de soumission (http://submit.search.yahoo.com/free/request) : *You can also provide the location of a text file containing a list of URLs, one URL per line, say urllist.txt. We also recognize compressed versions of the file, say urllist.gz.*
> Vous faites ainsi d'une pierre deux coups. En revanche, Yahoo! ne reconnaît pas (encore ?) le protocole "sitemaps" de Google (voir ci-après)...

Les trois solutions ci-dessus sont intéressantes mais elles souffrent toutes d'un handicap majeur : elles ne permettent que de donner une liste d'adresses, sans informations complémentaires à leur sujet : date de dernière modification, fréquence de mise à jour, etc. C'est pour cela qu'il sera plus intéressant (mais également plus long et fastidieux...) d'utiliser le format proposé par Google ("sitemap protocol"), dont il est important de signaler qu'il est fourni sous la coupe d'une licence "Creative Commons" (http://creativecommons.org/licenses/by-sa/2.0/), ce qui signifie que d'autres moteurs peuvent l'utiliser (ce qui serait d'ailleurs une excellente idée...)...

Format des fichiers "sitemap protocol"

Le format "sitemap protocol" décrit un fichier XML qui va fournir des indications pour chaque page de votre site.

Le fichier créé sera de cette forme :

```
<?xml version="1.0" encoding="UTF-8"?>
<urlset xmlns="http://www.google.com/schemas/sitemap/0.84">
   <url>
       <loc>url</loc>
       <lastmod>date</lastmod>
       <changefreq>fréquence de mise à jour</changefreq>
       <priority>priorité</priority>
   </url>
</urlset>
```

Contenant les indications suivantes :

- *urlset* (obligatoire) commence et termine (*/urlset*) le fichier en question.
- *url* (obligatoire) décrit chaque page et contient les champs suivants :

  * *loc* représente l'adresse de la page (http://www.votresite.com/page1.html). Ce champ commence par "http://" et se termine par un "/" éventuellement. Ce champ ne peut contenir plus de 2048 caractères.
  Il est important et assez strict dans son utilisation. Attention donc à bien suivre les indications suivantes :
    • Chaque url doit être indiquée de façon absolue (pas d'affichage relatif du type "../directory/page.html"), donc commencer toujours par la mention "http://...".
    • Chaque page indiquée dans le fichier doit être située dans le répertoire où se trouve le fichier sitemap ou dans un répertoire de niveau inférieur.

    Exemple : vous créez le fichier http://www.votresite.com/produits/sitemap.xml

Ce fichier peut décrire les pages suivantes :
http://www.votresite.com/produits/index.html
http://www.votresite.com/produits/gamme.html
http://www.votresite.com/produits/electricite/rupteur.html

Mais il ne pourra pas décrire les pages suivantes :
http://www.votresite.com/contact.html
http://votresite.com/produits/contact.html
https://www.votresite.com/produits/index.html (notez le "https" pour un accès sécurisé)
http://www.votresite.com/clients/reference.html

Ces pages seront refusées par Google lors de la lecture du fichier.

Pour cette raison, l'emplacement le plus logique pour un fichier sitemap sera le niveau le plus haut de l'arborescence (http://www.votresite.com/sitemap.xml). Ceci dit, rien ne vous empêche :
- De mettre le fichier sitemap dans un autre répertoire (en tenant compte, dans ce cas, des restrictions évoquées ci-dessus).
- De créer plusieurs fichiers Sitemaps pour un même site (voir ci-après).

* *lastmod* est la date de dernière modification du fichier. Cette date doit répondre au format ISO 8601 (http://www.w3.org/TR/NOTE-datetime), le plus souvent sous la forme YYYY-MM-DD soit 2005-09-15 pour le 15 septembre 2005.

* *changefreq* représente la fréquence de mise à jour de la page, à choisir parmi les possibilités suivantes : *always, hourly, daily, weekly, monthly, yearly, never*. Bien entendu, dans ce cas, il faudra faire des choix en optant pour la fréquence la plus vraisemblable si celle-ci n'est pas constante.

Important : nous vous conseillons de ne pas tricher sur ce champ. Rien ne servira d'indiquer "*hourly*" pour toutes les pages de votre site, si la majorité ne sont jamais mises à jour. Google a appris à connaître, par d'autres voies, la fréquence de mise à jour des documents qu'il indexe. Il semble évident qu'il n'appréciera que modérement si les données que vous lui fournissez sont à des années lumières de ses propres constatations sur votre site. Cela peut même vous desservir. Soyez donc le plus loyal possible sur cette indication, vous ne vous en portrez que mieux (et votre site également) à l'avenir.
D'autre part, cette indication n'est pas obligatoirement suivie à la lettre par les crawlers. Le fait d'avoir indiqué "*never*" ne signifie pas que les robots du moteur ne viendront qu'une seule et unique fois indexer la page et l'ignoreront par la suite. Il reviendront quand même, ne serait-ce que pour être sûr qu'elle existe encore...

* *priority* indique l'importance que vous donnez à la page à l'intérieur de votre site. Sa valeur va de 0 à 1 et peut être, bien entendu, décimale (0.5, 0.7, etc.). Attention : pas de virgule, c'est le point qui marquera ici la décimale. Si rien n'est indiqué, la priorité par défaut est fixée à 0.5.
Par exemple, la page d'accueil de votre site aura, vraisemblablement, une priorité de 1.

Important : là encore, jouez le jeu et indiquez des niveaux de priorité réels. Evitez de positionner ce champ à la valeur 1 pour toutes vos pages. Point important également : ce niveau de priorité ne joue aucunement sur le "ranking" de vos pages. Il s'agit uniquement de données fournies aux robots pour crawler de façon plus ou moins prioritaire vos documents...

Notez que le fichier doit être encodé au format UTF-8 obligatoirement. Certains caractères doivent donc être encodés différemment, notamment dans les urls où l'esperluette (&) doit apparaître ainsi : &, etc.

Notez également que les champs *lastmod, changefreq* et *priority* sont optionnels.

Enfin, dans le domaine des restrictions, sachez que votre fichier non compressé (vous pouvez également fournir des fichiers compressés en GZip) doit avoir une taille inférieure à 10 Mo et contenir des informations sur 50 000 pages au maximum, ce qui laisse un peu de marge (d'autant plus que vous pouvez travailler sur plusieurs fichiers...)...

Exemples de fichiers

Ainsi, un fichier ultra-simple, minimaliste, mais fonctionnel décrivant un site de 3 pages, sera le suivant :

```
<?xml version="1.0" encoding="UTF-8"?>
<urlset xmlns="http://www.google.com/schemas/sitemap/0.84">
   <url>
            <loc>http://www.votresite.com/</loc>
   </url>
   <url>
             <loc>http://www.votresite.com/produits.html</loc>
   </url>
   <url>
             <loc>http://www.votresite.com/apropos.html</loc>
   </url>
</urlset>
```

Ce fichier est très simple et n'aura qu'une fonction : signaler à Google la présence des trois pages. Cependant, il peut paraître plus simple, dans ce cas, comme nous l'avons indiqué auparavant dans cet article, d'utiliser le format texte (.txt) pour les signaler à Google.

Un fichier plus complet, contenant plus d'infos sera le suivant :

```
<?xml version="1.0" encoding="UTF-8"?>
<urlset xmlns="http://www.google.com/schemas/sitemap/0.84">
   <url>
           <loc>http://www.votresite.com/</loc>
           <lastmod>2005-09-01</lastmod>
           <changefreq>daily</changefreq>
           <priority>1</priority>
   </url>
   <url>
           <loc>http://www.votresite.com/produits.html</loc>
           <lastmod>2005-08-12</lastmod>
           <changefreq>weekly</changefreq>
           <priority>0.8</priority>
   </url>
   <url>
           <loc>http://www.votresite.com/apropos.html</loc>
           <lastmod>2005-06-01</lastmod>
           <changefreq>monthly</changefreq>
           <priority>0.5</priority>
   </url>
</urlset>
```

Chaque page se voit alors décrite avec ses 4 champs spécifiques : url, date de dernière modification, fréquence de mise à jour et priorité d'indexation.

Travail sur plusieurs fichers

Le protocole utilisé par Google autorise la possibilité de travailler sur plusieurs fichiers XML. Il faudra dans ce cas créer un nouveau fichier descriptif ("fichier mère"), nommé sitemap_index.xml, qui va contenir les indications sur les sous-fichiers ("fichiers filles") utilisés.
Sa structure est similaire à celle d'un fichier-fille. Voici le format d'un tel fichier :

```
<?xml version="1.0" encoding="UTF-8"?>
   <sitemapindex xmlns="http://www.google.com/schemas/sitemap/0.84">
   <sitemap>
      <loc>http://www.votresite.com/sitemap1.xml</loc>
```

```
   <lastmod>2005-01-01</lastmod>
</sitemap>
<sitemap>
   <loc>http://www.example.com/sitemap2.xml</loc>
   <lastmod>2005-01-01</lastmod>
</sitemap>
</sitemapindex>
```

L'option *lastmod* indique ici la date de dernière modification du fichier sitemap, et non pas des pages dont il détient la description.

Cas particulier des sous-domaines

Votre site utilise peut-être des sous-domaines, comme le site Abondance :
- www.abondance.com
- actu.abondance.com
- offre.abondance.com
- outils.abondance.com

Dans ce cas, chaque sous-domaine est considéré par Google comme un site à part entière. Le mieux est donc de créer un fichier sitemap pour chacun des sous-domaines, décrivant les pages que chacun contient. Exemple :

- www.abondance.com/sitemap-top.xml
- actu.abondance.com/sitemap-actu.xml
- offre.abondance.com/sitemap-offre.xml
- outils.abondance.com/sitemap-outils.xml

Chaque sous-domaine étant indépendant pour Google, un fichier de type "sitemapindex" (voir ci-dessus) n'est donc pas nécessaire dans ce cas... En revanche, vous devrez déclarer chacun de ces fichiers à Google. Nous y reviendrons...

Attention : www.votresite.com est considéré par Google comme un site différent de votresite.com. Ne l'oubliez pas...

La mise en place d'un fichier sitemap s'effectue donc en quatre étapes, de façon chronologique :

Etape 1 : Création du fichier

Vous avez plusieurs possibilités pour créer un fichier sitemap :

- Le créer "à la main", à l'aide d'un éditeur de texte. Cette solution sera peut-être la plus simple, voire la plus rapide pour un tout petit site. Elle deviendra rapidement fastidieuse, voire tout bonnement impossible à gérer pour de gros sites.

- Utiliser un script, un logiciel ou un site web en ligne, qui fera cette manipulation de façon automatique, tout en vous donnant la possibilité - ou non, nous y reviendrons - de modifier les résultats créés. Nous étudierons le mois prochain une liste de ces outils - qui s'avèrent vite indispensables pour des sites de moyenne et grande taille.

Le choix de l'outil - Google en propose également un - s'avère important car tous ne sont pas équivalents, loin de là, au niveau des fonctionnalités. Nous aurons l'occasion de les comparer le mois prochain...

Etape 2 : Validation du fichier

Pour être sûr que votre fichier est bien conforme au format XML, vous pouvez utiliser un certain nombre de programmes de validation dont vous trouverez une liste ici :

http://www.w3.org/XML/Schema#Tools
http://www.xml.com/pub/a/2000/12/13/schematools.html

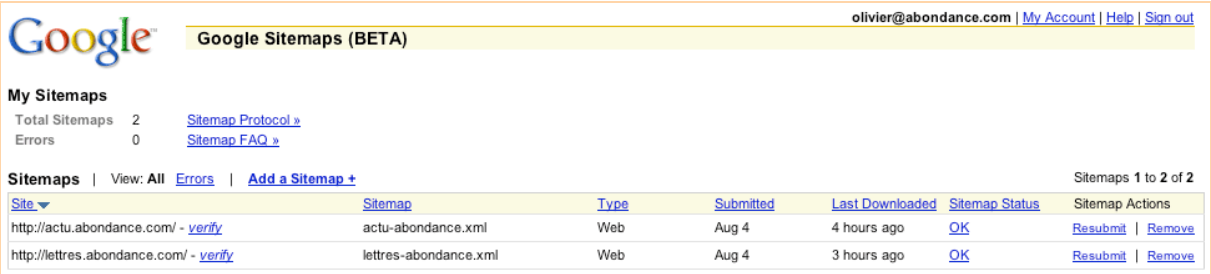
Ceci dit, au vu de la relative simplicité des fichiers Sitemaps et du fait que vous allez rapidement utiliser un applicatif qui automatise sa création, vous n'aurez rapidement plus à valider vos fichiers puisqu'on peut imaginer que les documents fournis par les logiciels sont "propres"...

D'autre part, sachez que l'interface d'administration de vos Sitemaps chez Google vous donne aussi la possibilité de tester leur validité et notamment l'exactitude des urls fournies (voir ci-dessous).

Etape 3 : Déclaration du fichier

Placer votre fichier sur votre site ne suffit pas. Il faut signaler à Google qu'il existe pour que celui-ci vienne le prendre en compte. Pour cela, vous devez disposer d'un compte Google, qu'il est possible de créer gratuitement sans aucun problème.

Une fois identifié sur la page adéquate de Google (https://www.google.com/webmasters/Sitemaps/login), vous avez accès à une interface d'administration très simple :

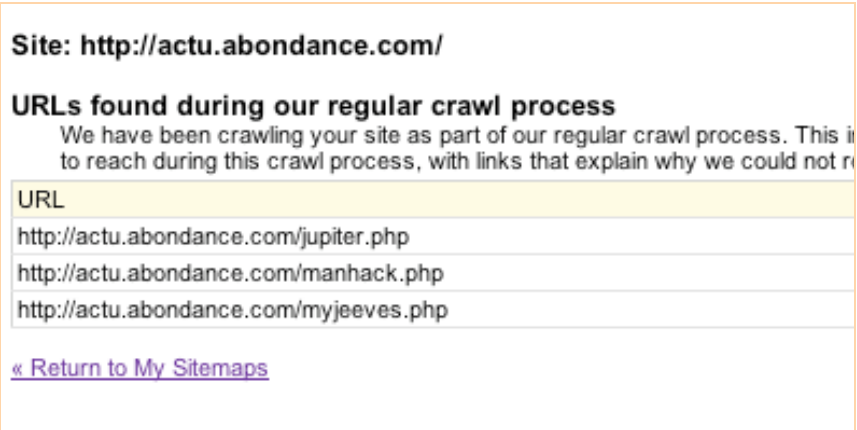


Le lien "*Add a sitemap* +" permet de signaler, sur une page de soumission spécifique, l'adresse de votre (ou de vos) fichier(s).

Devant chaque fichier enregistré, plusieurs liens sont disponibles :

- "*Verify*" vous permet, au prix d'une petite manipulation (ajout sur votre serveur d'un fichier texte dont Google vous donne le nom, la procédure étant expliquée en détail sur le site du moteur) de vérifier si toutes les urls proposées dans le fichier sont valides.

Voici un exemple de vérification qui a trouvé trois fichiers générant une erreur 404 :



- La colonne "*last downloaded*" vous indique quand votre fichier a été lu la dernière fois par Google. Information très intéressante s'il en est...

Les autres indications sont assez classiques, nous ne reviendrons pas dessus.

Etape 4 : Mise à jour du fichier

Enfin, il se peut, de façon assez logique, que le contenu de votre site change dans le temps : nouvelles pages qui apparaissent, anciennes qui disparaissent, etc. Vous devez donc mettre à jour en conséquence votre fichier sitemap. Google viendra l'indexer fréquemment, mais l'interface d'administration vous permet également de le resoumettre au besoin (lien "*Resubmit*").

Pour resoumettre votre fichier, vous pouvez également lancer sur votre navigateur une url du type :

http://www.google.com/webmasters/Sitemaps/ping?sitemap=http://www.votresite.com/sitemap.xml

En remplaçant bien évidemment l'indication http://www.votresite.com/sitemap.xml par l'adresse de votre fichier...

Nous arrêterons ici la présentation synthétique de Google Sitemaps. Logiquement, si vous avez lu attentitivement cet article, vous devez en savoir assez pour bien utiliser cette fonctionnalité du moteur. Reste maintenant à "mettre les mains dans le cambouis du moteur" et créer vos fichiers Sitemaps. C'est ce que nous verrons le mois prochain avec une comparaison des principaux outils vous permettant de le faire à votre place de façon automatique...

Documentation

Pour information, voici quelques liens importants au sujet de l'offre Google Sitemaps :

- Documentation (enfrançais) :
http://www.google.com/webmasters/sitemaps/docs/fr/about.html

- Le blog de Google dédié à l'offre Sitemaps
http://sitemaps.blogspot.com/

- Une interview de Shiva Shivakumar, responsable Sitemaps chez Google
http://blog.searchenginewatch.com/blog/050602-195224

- Le fichier Sitemap du site Google :-)
http://www.google.com/sitemap.xml