Comment utiliser intelligemment les Sitemaps XML ?

Retour au sommaire de la lettre

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

Les fichiers Sitemaps XML fêteront l'année prochaine leurs 10 ans d'existence. Parfois décriés par certains au niveau de leur utilité pour le référencement, ils s'avèrent pourtant très importants dans de nombreux cas. Encore faut-il bien les mettre en place, dans les règles de l'art, et connaître leurs nombreuses possibilités. Voici un article détaillé sur les bonnes pratiques en la matière qui devrait aider votre site à être mieux crawlé et indexé par Google, Bing et leur confrères...

L'utilité des sitemaps XML n'a jamais totalement fait l'unanimité au sein de la communauté des experts SEO. Leur impact sur le référencement naturel des sites a également souvent été remis en question. Pourtant, si on comprend leur objectif réel et si on sait comment les implémenter, un référenceur peut tirer un réel bénéfice de la soumission de ces fichiers d'URL au format XML.

L'objectif de cet article est donc de faire le point sur les fonctions assurées par les sitemaps, sur les gains réels qu'ils apportent, et sur les précautions à prendre lorsqu'on les construit. Au passage, nous rappellerons quelques astuces avancées, parfois méconnues par les webmasters.

Pourquoi le protocole Sitemap a-t-il été inventé ?

Les moteurs de recherche comme Google ont abandonné très tôt le principe de la soumission d'URL aux moteurs, pour leur préférer une exploration automatique des pages du web en suivant les liens découverts dans les pages crawlées.

Mais lorsque l'on utilise ce mode d'exploration, deux défis apparaissent : assurer une "couverture" aussi parfaite que possible du web, et garantir la "fraicheur" des données explorées.

Le problème de la couverture

Une partie des pages se situe dans une zone du web que l'on appelle souvent le "web invisible". Il s'agit de pages que les moteurs de recherche ne peuvent pas découvrir en suivant des liens placés sur des pages explorables. La structure des sites web et de nombreux problèmes techniques peuvent créer des situations (volontaires ou involontaires) où une page n'est liée à aucune page accessible aux moteurs. Le problème s'aggrave si on considère que les moteurs décident aussi d'"oublier" parfois volontairement d'explorer des pages qu'ils considèrent comme inutiles (doublons, pages extrêmement profondes etc.).

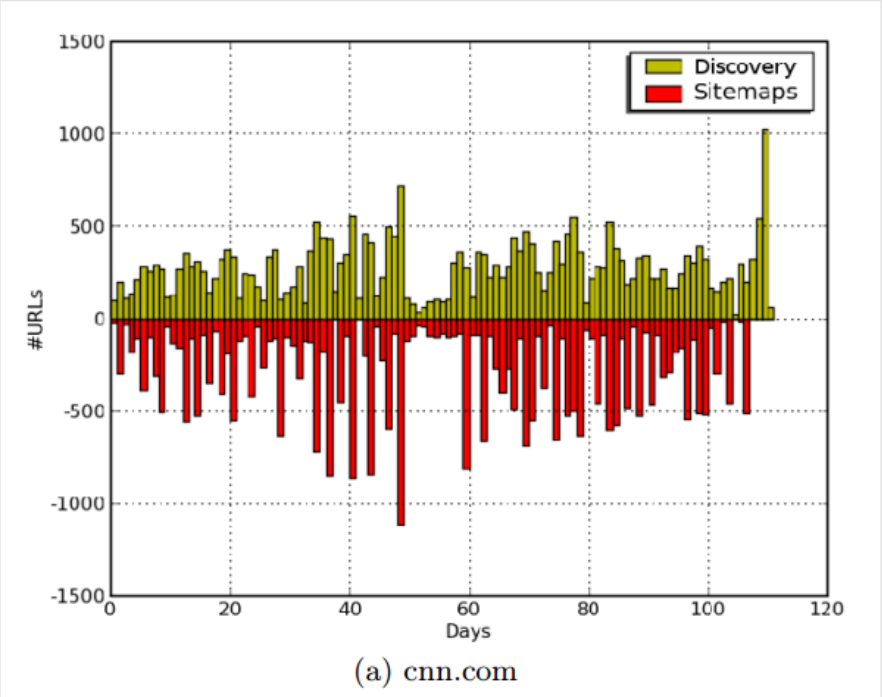
La "couverture" d'un robot d'exploration d'un moteur de recherche est caractérisée par le ratio entre les pages web explorées et les pages web intéressantes.

Améliorer la couverture est apparu comme une priorité absolue aux ingénieurs de Google, dès le début des années 2000, tant le volume des pages "ignorées" semblait grand.

La "fraîcheur" des données

L'autre défi posé par l'exploration automatique des pages web à l'aide d'un spider, est de garantir la fraîcheur des données. En effet, ce processus prend du temps, et dans la pratique, il peut s'écouler des semaines entre deux explorations d'une même page web. Si un lien vers une nouvelle page apparait sur un site, il peut donc s'écouler beaucoup

trop de temps avant que cette page ne soit détectée et explorée. Sur certains types de sites, des liens éphémères peuvent même ne jamais être détectés.



Une illustration tirée de l'article « sitemaps : beyond and above the crawl of duty ». L'analyse du nombre d'URL réellement explorées par Google sur le site CNN (en vert), comparées à celle mentionnées dans le sitemap (en rouge) montre que l'exploration passe à côté d'un grand nombre de pages. Pour un site dont le contenu évolue rapidement, les sitemaps sont indispensables pour assurer à la fois une bonne couverture et garantir la fraîcheur des résultats.

Juin 2005 : Google introduit le protocole sitemap XML

Les ingénieurs de Google se sont rendu compte que le seul moyen d'améliorer la couverture et la fraicheur des URL explorées étaient de mettre les webmasters des sites à contribution, en leur suggérant de soumettre la liste des URL de leur site au moteur.

En Juin 2005, Google a donc annoncé (http://actu.abondance.com/2005-23/google-sitemaps.php) le support d'un nouveau protocole baptisé "Sitemaps" (plans de site) conçu à cette fin, sur le modèle du protocole robots.txt. Un site officiel permet de s'informer sur le format à respecter pour construire un fichier sitemap : http://www.sitemaps.org/fr/

Qui supporte le format Sitemaps ?

Bing supporte le format depuis 2007. Baïdu et Yandex gèrent aujourd'hui aussi parfaitement le protocole sitemaps. Orange et Exalead également.

Le format sitemap XML

Décrire dans le détail le format des sitemaps XML dépasse la portée de cet article : on se reportera donc utilement à la documentation fournie sur le site du protocole sitemaps mentionné ci-dessus, et à l'aide de Google sur les sitemaps : https://support.google.com/webmasters/answer/156184?hl=fr

Rappelons simplement que les sitemaps XML doivent être encodés en UTF-8, et respecter le format habituel des fichiers XML, notamment l'obligation d'échapper les caractères réservés. Voici un exemple basique de fichier sitemap XML

```
<?xml version="1.0" encoding="utf-8"?>
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.sitemaps.org/schemas/sitemap/0.9
http://www.sitemaps.org/schemas/sitemap/0.9/sitemap.xsd">
   <url>
      <loc>http://example.com/</loc>
      <lastmod>2006-11-18</lastmod>
      <changefreq>daily</changefreq>
      <priority>0.8</priority>
   </url>
</urlset>
```

Les deux autres formats autorisés pour le sitemaps XML

Le format XML reste le format recommandé et c'est également le plus utilisé par les webmasters (à 90% d'après un chiffre communiqué par Google en 2009). Mais le protocole Sitemaps prévoit le support de deux autres formats de fichier :

- Le format texte
Il est possible de stocker les URL à soumettre dans un fichier texte brut, à raison d'une ligne par URL. Le fichier doit être encodé en UTF-8. Le format des URL doit respecter les directives défini par le protocole sitemaps, et les fichiers sont soumis aux mêmes limitations de taille que les sitemaps XML.

- Le format "flux de syndication" (flux RSS)
On peut également se servir des formats définis pour les flux de syndication, à savoir : Atom 0.3 ou 1.0, et RSS 2.0. Cette possibilité s'avère très pratique lorsque l'on utilise des CMS qui savent créer facilement des flux RSS pour le contenu d'un site mais ne disposent pas d'un outil générant des sitemaps xml automatiquement.

Le problème de la taille des sitemaps

Les fichiers sitemaps sont limités en taille :
- ils sont limités à 50 000 URL par fichier.
- et ne peuvent pas dépasser un poids de plus de 10 Mo.
Notons qu'il est possible, et même hautement conseillé, de compresser les fichiers sitemaps au format gzip. Donc dans la pratique, la limite d'URL sera atteinte avant la limite de poids du fichier.

La solution pour les "gros" sites : le fichier "sitemap index"

Pour un site de taille respectable, la limite des 50 000 URL est vite atteinte. La solution consiste donc à éclater les sitemaps en plusieurs fichier et à utiliser un fichier "sitemap index" qui indique la liste des différents fichiers xml sitemaps créés. Le format du fichier sitemap index est légèrement différent :

```
<?xml version="1.0" encoding="UTF-8"?>
<sitemapindex xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
  <sitemap>
    <loc>http://www.example.com/sitemap1.xml.gz</loc>
    <lastmod>2004-10-01T18:23:17+00:00</lastmod>
  </sitemap>
  <sitemap>
    <loc>http://www.example.com/sitemap2.xml.gz</loc>
    <lastmod>2005-01-01</lastmod>
```

```
  </sitemap>
</sitemapindex>
```

Les formats de sitemaps spécifiques

Au fil du temps, le protocole sitemap a été enrichi pour permettre l'exploration de données spécifiques. Ces formats particuliers concernent :
- les videos ;
- les images ;
- les actualités ;
- les URL mobile ;
- les pages multilingues.

Le format sitemap video :

Ce format spécifique est détaillé ici :
https://support.google.com/webmasters/answer/80472?hl=fr

Google a créé ce format car l'exploration automatique des vidéos se heurte à de nombreux problèmes techniques. Beaucoup d'implémentations de lecteurs vidéo empêchent une exploration normale des flux, il est donc fortement recommandé de construire un sitemap vidéo.

On peut également utiliser des fichiers au format mRSS.

L'ajout d'information sur les images :

Il est possible d'utiliser des balises complémentaires pour soumettre des données sur les images que l'on souhaite voir indexer pour ses pages. Ces balises sont décrites ici :
https://support.google.com/webmasters/answer/178636?hl=fr

Voici un exemple de sitemap enrichi par les balises pour les images :

```
<?xml version="1.0" encoding="UTF-8"?>
 <urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9"
  xmlns:image="http://www.google.com/schemas/sitemap-image/1.1">
 <url>
   <loc>http://example.com/sample.html</loc>
   <image:image>
     <image:loc>http://example.com/image.jpg</image:loc>
   </image:image>
   <image:image>
     <image:loc>http://example.com/photo.jpg</image:loc>
   </image:image>
 </url>
</urlset>
```

On peut déclarer au maximum 1000 images pour une page.

Les sitemaps "news" :

Un format spécifique a été développé par Google pour les pages d'actualités. Ce sitemap est conçu pour faciliter l'exploration et l'indexation des pages dans le moteur vertical Google News. Compte tenu des spécificités du crawler de Google News (et de ses limites techniques ainsi que de son caractère "capricieux"), construire un sitemap news quand on est référencé dans un site d'actualité est fortement recommandé.

Les spécificités de ce format sont détaillées ici :
https://support.google.com/news/publisher/answer/74288?hl=fr

Voici un exemple de fichier exploitant ce format :

```
<?xml version="1.0" encoding="UTF-8"?>
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9"
      xmlns:news="http://www.google.com/schemas/sitemap-news/0.9">
  <url>
    <loc>http://www.example.org/business/article55.html</loc>
    <news:news>
      <news:publication>
        <news:name>Journal L'Exemple</news:name>
        <news:language>fr</news:language>
      </news:publication>
      <news:access>Subscription</news:access>
      <news:genres>PressRelease, Blog</news:genres>
      <news:publication_date>2008-12-23</news:publication_date>
      <news:title>Les entreprises A et B envisagent une fusion</news:title>
      <news:keywords>économie, fusion, acquisition, A, B</news:keywords>
      <news:stock_tickers>NASDAQ:A, NASDAQ:B</news:stock_tickers>
    </news:news>
  </url>
</urlset>
```

Indiquer les URL pour mobile dans un sitemap :

Lorsqu'un site dispose d'URL destinées aux appareils mobiles, il est possible d'indiquer cette spécificité à l'aide de balises supplémentaires.
Ce format est détaillé ici :
https://support.google.com/webmasters/answer/34648?hl=fr

```
<?xml version="1.0" encoding="UTF-8" ?>
 <urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9"
  xmlns:mobile="http://www.google.com/schemas/sitemap-mobile/1.0">

    <url>
       <loc>http://mobile.example.com/article100.html</loc>
       <mobile:mobile/>
    </url>
</urlset>
```

Indiquer les pages multilingues dans un sitemap :

Enfin, il est possible d'ajouter des informations de type "hreflang" dans les sitemaps, pour indiquer qu'une page existe en plusieurs versions linguistiques.

Voici un exemple de ce format :

```
<?xml version="1.0" encoding="UTF-8"?>
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9"
  xmlns:xhtml="http://www.w3.org/1999/xhtml">
  <url>
    <loc>http://www.mon-domaine.com/english/</loc>
    <xhtml:link rel="alternate" hreflang="de" href="http://www.mon-
domaine.com/deutsch/" />
    <xhtml:link rel="alternate" hreflang="de-ch href="http://www.mon-
domaine.com/schweiz-deutsch/" />
    <xhtml:link rel="alternate" hreflang="en" href="http://www.mon-
domaine.com/english/" />
  </url>
  <url>
    <loc>http://www.mon-domaine.com/deutsch/</loc>
```

```
    <xhtml:link rel="alternate" hreflang="en" href="http://www.mon-
domaine.com/english/" />
    <xhtml:link rel="alternate" hreflang="de-ch" href="http://www.mon-
domaine.com/schweiz-deutsch/" />
    <xhtml:link rel="alternate" hreflang="de" href="http://www.mon-
domaine.com/deutsch/" />
  </url>
  <url>
    <loc>http://www.mon-domaine.com/schweiz-deutsch/</loc>
    <xhtml:link rel="alternate" hreflang="de" href="http://www.mon-
domaine.com/deutsch/" />
    <xhtml:link rel="alternate" hreflang="en" href="http://www.mon-
domaine.com/english/" />
    <xhtml:link rel="alternate" hreflang="de-ch" href="http://www.mon-
domaine.com/schweiz-deutsch/" />
  </url>
</urlset>
```

Ce format est décrit ici :
https://support.google.com/webmasters/answer/2620865

Remarque : la plupart de ces formats sont spécifiques à Google. Mais ils respectent tous la norme du protocole sitemap, donc ils sont tous partiellement supportés par les autres moteurs (au pire, les informations complémentaires sont ignorées).

Où placer les fichiers sitemaps ?

Le protocole sitemaps introduit une contrainte importante : seules les URL situées dans le même répertoire que le sitemap lui même seront prises en compte. Par exemple, si le fichier sitemap est placé dans ce répertoire :
www.domaine.com/rep1/
Alors ce fichier peut contenir des URL de ce type
www.domaine.com/rep1/page.html ou www.domaine.com/rep1/ssrep/page.html
mais pas
www.domaine.com/repdifferent/page.html

Le fichier sitemap index peut lui être placé n'importe où sur le site. Par contre, les fichiers sitemap individuels qu'il mentionne doivent respecter la règle ci-dessus.

Cette contrainte peut s'avérer problématique s'il est impossible de placer le fichier sitemap à la racine du site.

Google propose heureusement une solution : si le fichier est soumis via Google Webmaster Tools, les URL appartenant à un host partageant le même propriétaire validé sont autorisés. Dans ce cas, un sitemap peut contenir des URL appartenant à différents sous domaines et même à différents domaines !
http://hote1.example.com
http://hote2.example.com
http://hote3.example.com
http://hote1.example1.com
http://hote1.example.ch

Utilisez "lastmod" à bon escient

Eclater les fichiers sitemaps en plusieurs fichiers présente un avantage supplémentaire. Il peut être intéressant de placer dans des fichiers séparés :
- les URL statiques (zones du site qui ne changent pas au fil du temps) ;
- les URL récentes (pages nouvellement ajoutées) ;
- les URL éphémères (zones du site où des pages apparaissent et disparaissent fréquemment).

Cette séparation permet de jouer sur l'effet de la balise xml "lastmod", qui indique la date de dernière modification du fichier. La mise à jour correcte de cette balise permet d'assurer la prise en compte rapide des URL récentes et des URL éphémères.

Attention à la génération automatique des sitemaps

Il existe beaucoup de solutions permettant de générer automatiquement les sitemaps pour un site internet.

Ces solutions sont de deux types :
- ***les plugins ou fonctionnalités internes*** : beaucoup de CMS disposent de solutions permettant de générer un sitemap automatiquement à partir de leur base de données de contenu. Le problème est que ces solutions ignorent toutes les pages du site qui n'auraient pas été créées à l'aide du CMS : le Sitemap peut donc s'avérer incomplet dans certains cas.
- ***les crawlers générateurs de sitemap*** : dans ce cas, le sitemap est généré à partir d'un crawl du site. Le sitemap contient donc toutes les URL… explorables. Cette solution présente donc l'inconvénient d'être incapable d'inclure des URL situées dans des zones invisibles du site (cachées derrière un formulaire, ou derrière un script ajax ou actionscript). Et il faudra générer un nouveau sitemap à chaque changement sur le site (création, modification, suppression de page), pour avoir un fichier à jour, ce qui peut rapidement s'avérer fastidieux.

Il est important de ne pas accorder trop de confiance à ces générateurs automatiques de sitemaps. Un examen attentif des résultats montre :
- qu'ils sont parfois bogués, et génèrent des syntaxes d'URL erronées ;
- qu'ils sont souvent limités, et ne savent pas générer des formats spécifiques ;
- qu'ils ne tiennent pas compte des pages bloquées par le robots.txt ;
- qu'ils ne savent pas proposer une version canonique.

Dans la pratique, il faudra donc :
- soit corriger les fichiers produits par ces générateurs ;
- soit modifier les scripts pour obtenir un résultat conforme.

Eviter l'indexation des sitemaps

Les sitemaps sont des fichiers web qui peuvent aussi être explorés et indexés par des moteurs de recherche. Une requête sur "sitemap.xml" permet de vérifier que cette situation n'est pas exceptionnelle. Lorsque cela arrive, les webmasters veulent en général faire disparaître ce résultat, pour des raisons de sécurité ou pour éviter le faciliter le vol de contenu.

La solution consiste à utiliser l'attribut noindex dans une directive x-robots-tag placée dans l'en-tête http: du fichier sitemap.
Le format de cette directive est décrit ici :
https://developers.google.com/webmasters/control-crawl-index/docs/robots_meta_tag?hl=fr

Soumettre les sitemaps

Les fichiers sitemap / sitemap index ne peuvent pas être découverts automatiquement par les moteurs, car contrairement au robots.txt, le nom de ce fichier et son emplacement ne sont pas fixes.

On peut "soumettre" le fichier de deux façons différentes :
- soit en le déclarant dans le fichier robots.txt ;

- soit en le soumettant via un compte webmaster tools.

La soumission via le fichier robots.txt

La syntaxe est très simple :

Sitemap: http://www.mon-domaine.fr/sitemap1.xml
Sitemap: http://www.mon-domaine.fr/sitemap2.xml.gz

Pensez à séparer par un caractère retour chariot créant une ligne blanche les directives concernant les sitemaps. On peut déclarer autant de sitemaps qu'on le souhaite dans un fichier robots.txt

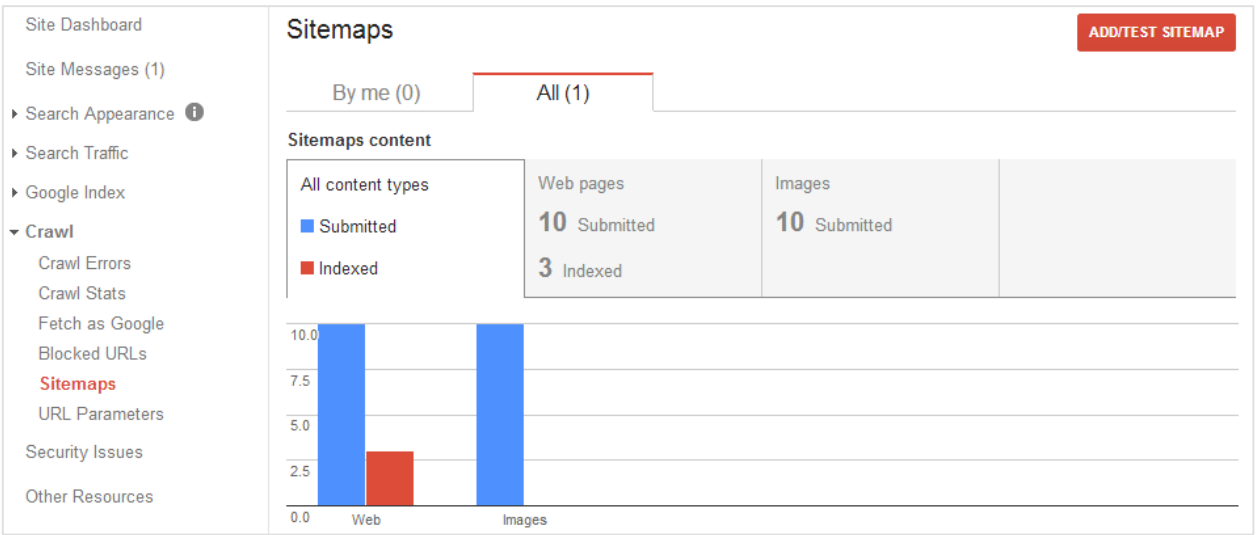
Cette solution fonctionne avec tous les moteurs de recherche supportant le protocole sitemaps.

La soumission via les webmaster tools

Certains moteurs (par exemple Google, mais également Bing) permettent la soumission des sitemaps via le compte Google Webmaster Tools. Avoir recours à cette solution n'empêche pas d'utiliser aussi la soumission via robots.txt.

La soumission via Webmaster Tools permet d'accéder à un grand nombre d'informations utiles comme :
- la confirmation de la prise en compte du fichier par le moteur ;
- la liste des erreurs rencontrée par le moteur ;
- le nombre d'URL indexées.

Il est donc possible de savoir facilement quel est le taux d'indexation d'un groupe d'URL défini par un fichier sitemap, ce qui peut fournir des renseignements utiles pour le référencement.



Vue de la rubrique sitemaps de Google Webmaster Tools : ici le sitemap soumis contient 10 URL, mais seulement 3 sont indexées.

Le Ping de sitemaps pour accélérer leur prise en compte

Il peut s'écouler un certain temps (plusieurs heures, voire plusieurs jours selon les sites) entre la soumission d'un fichier sitemap et sa prise en compte par les moteurs. C'est surtout vrai si le fichier est soumis via le fichier robots.txt uniquement.

Une façon très efficace de signaler à un moteur la présence d'un nouveau fichier sitemap est d'utiliser les URL de ping des moteurs.

Par exemple pour Google la syntaxe est :

www.google.com/webmasters/tools/ping?sitemap=http%3A%2F%2Fwww.yoursite.com%2Fsitemap.gz

Via le compte webmaster tools de Google, il suffit de cliquer sur le bouton "soumettre à nouveau" et le moteur sera averti que vous avez rafraichi le contenu du sitemap.

Les sitemaps sont-ils réellement utiles ?

L'utilité des sitemaps pour le référencement a fait l'objet dans le passé de débats riches et passionnés. Deux camps se sont longtemps opposés. Certains experts, dont Rand Fishkin de Moz, ont longtemps défendu l'idée que les sitemaps étaient potentiellement plus nuisibles qu'autre chose. A l'inverse, d'autres ont défendu l'idée que les sitemaps xml étaient indispensables pour le référencement.

Avec le temps et l'expérience, un consensus a commencé à se dégager en faveur de l'utilité des sitemaps.

L'un des arguments des "anti" sitemaps était l'existence d'un phénomène d'"aplatissement" de la structure du site. En effet, la notion d'arborescence disparait totalement dans les sitemaps, qui ne contiennent aucune information sur la manière dont les pages sont reliées entre elles. Alors que cette information est évidemment collectée lors de l'exploration automatique traditionnelle.

Mais en réalité, l'information sur les liens entre les pages, ou l'arborescence n'est pas "écrasée" en cas de recours aux sitemaps. Les moteurs ajoutent simplement à la liste des URL stockées dans l'ordonnanceur (le composant logiciel qui décide des URL à crawler) les URL manquantes découvertes dans les sitemaps.

Une autre critique adressée aux sitemaps était leur inefficacité en termes d'amélioration du référencement et du positionnement. Force est de constater que les sitemaps ont pour mission unique de résoudre des problèmes de couverture et de fraicheur lors des crawls, et que leur impact sur le référencement ne peut être qu'indirect, en favorisant la découverte et l'indexation de pages qui auraient été ignorées sans leur présence.

Avec le temps, le camp des "anti" s'est dégarni (Rand Fishkin s'est rangé dans le camp des pro-sitemaps après deux ans d'expérience), et une majorité des experts SEO, à l'instar des moteurs eux-mêmes, recommandent de construire et de soumettre des fichiers sitemaps.

La balise "priority" sert-elle à quelque chose ?

Si on ne peut pas indiquer dans un sitemap xml la position d'une url dans l'arborescence, le protocole prévoit, dans son format xml, de pouvoir indiquer via le tag "priority" une valeur entre 0,0 et 1,0 (la valeur de priorité par défaut est de 0,5 pour une url) indiquant le niveau d'importance de la page selon les webmasters. Il s'agit d'une indication fournie aux moteurs que ces derniers exploitent… ou non. Une analyse des logs montre en effet souvent que le comportement de crawl de Google n'est pas impacté par des changements de valeur de priorité. En réalité, cette balise comme la balise <changefreq> qui déclare la fréquence de changement des pages, sont ignorées si le crawler estime les données issues de l'exploration normale plus fiable. Ce comportement frustrant ne doit pas vous empêcher d'indiquer des valeurs cohérentes pour ces balises.

Des sitemaps indispensables dans les phases de lancement de sites / de création de pages

Certaines étapes de la vie d'un site peuvent être facilitées grâce à la soumission de sitemaps adaptés : il s'agit de toutes les étapes de lancement des sites ou d'ajout massif de nouvelles URL. Dans ces contextes, un fichier sitemap peut aider le moteur à découvrir et indexer rapidement les URL. C'est d'autant plus indispensable si le moteur crawle peu fréquemment votre site.

Indiquez l'url canonique dans vos sitemaps

Il faut signaler un comportement particulier de Google dans la prise en compte des URL incluses dans un sitemap : le moteur considère que la syntaxe incluse dans un sitemap est celle que le webmaster déclare préférer. Cela signifie qu'en cas de doublon d'URL, la simple mention d'une des deux syntaxes dans un sitemap pourra conduire Google à "canonicaliser" les syntaxes dupliquées, pour retenir la syntaxe du sitemap (par exemple, si le sitemap contient des URL en https, et si le site contient aussi des versions http des mêmes pages, la version indexée sera la version https).

Compte tenu de ce comportement, il est conseillé de rester cohérent, et d'adopter des conventions similaires dans les link rel canonical. Par exemple :
Si l'on retient l'URL http://www.domaine.com/page.html?param=1 comme URL préférentielle et qu'il existe une URL avec un contenu identique dont l'URL serait http://www.domaine.com/page.html?param=2 alors la première url doit être placée dans le sitemap (pas la seconde), et la deuxième URL peut contenir un *link rel=canonical* pointant vers la première URL (et pas l'inverse).

Faut-il inclure des pages en noindex ou en nofollow dans le sitemap ?

Les sitemaps servent à indiquer la liste des URL que les moteurs doivent explorer. Cela n'a rien à voir avec l'indexation. Par conséquent :
- des pages contenant une meta robots avec comme valeur d'attributs noindex, follow ont parfaitement leur place dans un fichier sitemap ;
- par contre, les pages en nofollow, logiquement, doivent être exclues ;
- de même les URL dont la syntaxe est bloquée par une directive disallow dans le robots.txt n'ont rien à faire dans le sitemaps ;

Attention à ne pas mettre en ligne des sitemaps bourrés d'erreur

Si vous voulez que les sitemaps soient correctement pris en compte par les moteurs, il est important de faire la chasse aux erreurs qu'ils peuvent contenir. En particulier
- les URL dans les sitemaps ne doivent pas renvoyer des codes 404, 500 ou tout code autre que 200 ;
- les URL dans les sitemaps ne doivent pas être redirigées ;
- les URL dans les sitemaps doivent correspondre à des pages uniques, autant que possible ;
- les URL ne doivent pas être en nofollow ou bloquées par un robots.txt.

Si le sitemap contient trop d'erreurs, le contenu du fichier ne sera pas pris en compte par le moteur de recherche. Ce taux est connu pour Bing : 10%.

Informations détaillées sur les sitemaps

Tester le sitemap

Sitemap : /fr/sitemap.xml

Ce sitemap a été envoyé le 23 août 2011 et traité le 19 avr. 2012.

Informations sur l'erreur : 2 erreurs, 0 avertissements.

Afficher : Tous | Erreurs | Avertissements — Afficher 25 lignes — 1 à 2 sur 2

#		Type	Problème	Description	Nombre de problèmes	Exemple	Ligne	Détectée
1	⚠	Erreurs	Erreur	Une erreur s'est produite lors de la tentative d'accès à votre sitemap. Veuillez vous assurer qu'il respecte nos consignes et qu'il est accessible à l'emplacement que vous avez indiqué, puis envoyez-le à nouveau.	1	**Erreur HTTP générique: 404 introuvable** Erreur HTTP : 404	-	4 juil. 2014
2	⚠	Erreurs	Le sitemap est un fichier HTML.	Votre sitemap semble correspondre à une page HTML. Veuillez utiliser un format de sitemap pris en charge.	1	Tag : html	2	19 avr. 2012

Dans la rubrique sitemaps, Google Webmaster Tools retourne les problèmes d'exploitation des sitemaps en les classant en deux types : les « erreurs » qui empêchent la prise en compte du fichier entier, et les avertissements qui concernent des URL individuelles.

Conclusion

Soumettre un Sitemap XML pour les moteurs de recherche se révèle être toujours une excellente idée. Les sitemaps news, ou video, sont même indispensables pour un bon référencement dans les moteurs verticaux de Google. Mais il ne faut pas s'attendre à un impact direct sur le positionnement des pages web de votre site : leur objectif est simplement d'aider les moteurs de recherche à découvrir l'intégralité du contenu de votre site.

Parfois décriés, les sitemaps ne présentent pas d'inconvénients majeurs pour le référencement, et ils ne présentent que des avantages si les sitemaps ne contiennent pas d'erreurs. Le soit-disant impact négatif des sitemaps pour le SEO n'est qu'une légende urbaine. En réalité, la tendance lourde est même aujourd'hui d'enrichir le contenu possible des sitemaps pour élargir leurs fonctionnalités…

Après bientôt dix ans d'existence, l'utilité des sitemaps ne cesse de progresser et ils vont probablement continuer à jouer un rôle important à l'avenir pour le référencement.

Bibliographie

Sites utiles :

Site officiel du protocole sitemaps
http://www.sitemaps.org

Page de support de Google sur les sitemaps
https://support.google.com/webmasters/answer/156184?hl=fr

Article scientifique :

Sitemaps: Above and Beyond the Crawl of Duty
http://www2009.eprints.org/100/1/p991.pdf
Uri Schonfeld, UCLA Computer Science Department. Narayanan Shivakumar, Google Inc.

Philippe YONNET , ***Directeur Général de l'agence Search-Foresight (http://www.search-foresight.com).***