## Les crawlers, des outils indispensables pour le SEO (1ère partie)

Retour au sommaire de la lettre

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

*On connait tous les robots Googlebot et Bingbot, utilisés leur moteur de recherche respectif pour explorer des milliards de sites web chaque jour. Mais il existe également des outils qui simulent ces robots et vous donnent de nombreuses indications sur l'indexabilité de votre site web. Des systèmes qui sont devenus indispensables au fil du temps dans le cadre d'un audit SEO. Cet article en deux parties vous fera découvrir ces différents outils ce mois-ci ainsi que leur utilisation "classique", avant de s'attarder à des usages plus avancés le mois prochain...*

Les crawlers font partie de ces "briques" logicielles indispensables pour créer un moteur de recherche. La raison d'être de ces programmes est de permettre la découverte des pages et des documents publiés sur le web, et de les télécharger à fins d'indexation ou d'analyse. Au fil du temps, on les a affublés d'autres jolis noms anglais tout aussi imagés : harvesters (moissonneuses), spiders (araignées), bots (robots)…

Dans cette série de deux articles, nous allons passer en revue les raisons pour lesquelles les outils de type "crawler" sont indispensables dans la boîte à outils du référenceur. Le premier article sera consacré aux utilisations classiques des outils de crawl, et le second à des utilisations plus avancées…

### Les moteurs de recherche explorent le web à l'aide de crawlers

Même si on peut trouver des exemples de moteurs de recherche qui, dans les premiers temps du web, ont référencé ou indexé des pages découvertes par "soumission" des webmasters, la règle pour les moteurs grand public consiste plutôt à utiliser un crawler pour découvrir les pages publiées sur le World Wide Web.

Le principe d'un crawler ou d'un spider qui fonctionne en mode exploration est le suivant :
1. On fournit une première URL au crawler.
2. Le crawler lance une requête http:// et télécharge le contenu de la page.
3. Le crawler analyse le contenu de la page, en extrait les informations utiles, et les stocke dans un entrepôt de données (*data repository*). Ces données sont ensuite retraitées pour créer le fameux "index" du moteur.
4. Parmi les données trouvées dans la page, figurent la mention d'autres URL dans le code HTML de la page (ou sous forme de liens hypertextes dans des PDF, des .doc Word…).
5. Ces liens sont placés dans une file d'attente.
6. Chaque lien fait à son tour l'objet d'une requête http://, le contenu de la page est téléchargé, analysé, de nouveaux liens sont découverts et ainsi (presque) à l'infini.

A chaque étape du processus, une URL est dans l'un de ces trois états :
- l'URL peut ne pas avoir été découverte encore ;
- l'URL peut avoir été découverte, mais n'a pas encore été téléchargée et son contenu analysé ;
- l'URL peut avoir été découverte, et téléchargée.

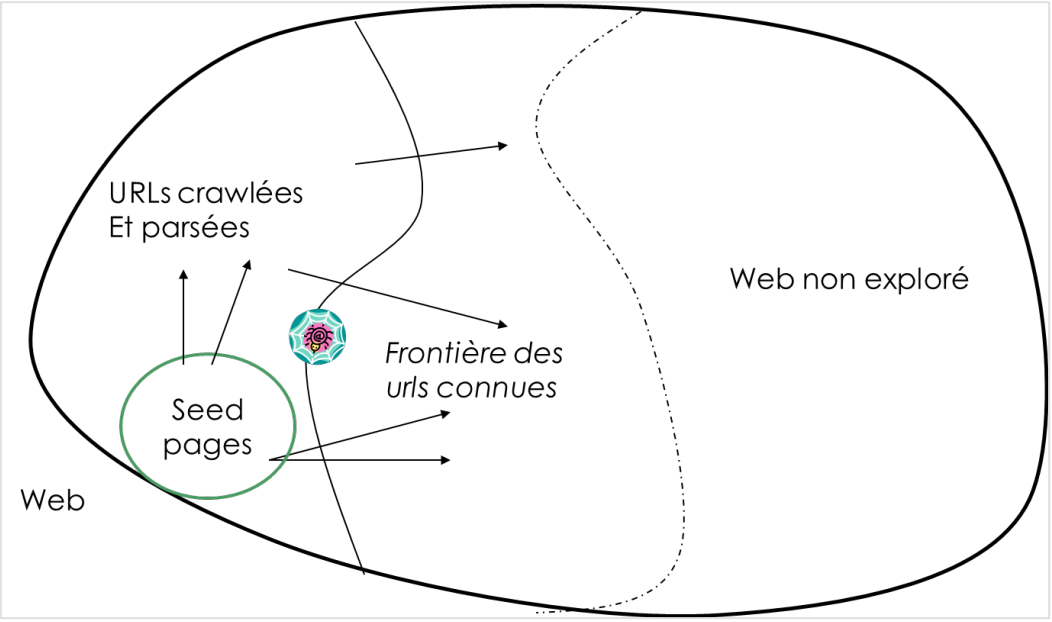



*Illustration du processus de crawl d'un moteur de recherche*

Si on choisit mal le lien de départ, le processus peut s'arrêter très vite, et des pans entiers du web peuvent ne jamais être découverts.

Pour éviter cela, un moteur qui se lance sélectionne au départ une série d'URL (baptisées "seed URL", URL semence) qui sont choisies parce qu'elles :
- contiennent beaucoup d'autres URL (par exemple des pages d'annuaire) ;
- couvrent des zones différentes du web ;
- sont proches de "noyaux du web" qui vont donner accès à un grand nombre d'URL.

Mais quoi qu'il en soit, cette méthode à ses limites, et de nombreuses URL ne peuvent pas être découvertes en utilisant cette méthode (ces URL sont placées dans une zone du web qu'on appelle communément le "web invisible").

Les problèmes qui peuvent empêcher un crawler de moteur de recherche de découvrir une URL sont assez divers :
- les URL accessibles uniquement *via* un formulaire en mode post ;
- les URL bloquées par un mot de passe, un .htaccess ;
- les URL utilisant des protocoles ou des ports exotiques ;
- les URL bloquées par un fichier robots.txt ;
- etc.

## *Pourquoi utiliser un crawler pour auditer son site ?*

Comme les moteurs de recherche utilisent leur propre crawler pour découvrir les pages d'un site, tout ce qui va gêner ou bloquer le travail du crawler, soit pour découvrir toutes les URL du site, soit pour en télécharger le contenu, peut avoir un impact sur le SEO.

En effet, une page qui n'a pas pu être découverte n'a aucune chance de figurer dans l'index de Google, donc de se positionner sur la moindre requête, et *a fortiori*, de générer le moindre trafic issu du moteur de recherche.

**Remarque** : la plupart des moteurs supportent le format XML sitemap, qui permet de fournir les URL des pages à indexer par un moteur. Force est de constater que les moteurs font toujours plus confiance aux informations sur les URL "découvertes" par le processus de crawl que *via* le sitemap XML. Ce dernier sert surtout à faire découvrir des URL que le processus de crawl échoue à découvrir, et à indiquer au moteur les préférences des webmasters, notamment en matière de syntaxe d'URL à indexer.  Mais la

possibilité d'uploader un sitemap ne permet pas de solutionner tous les problèmes de découvertes d'URL et d'indexation. La présence d'un sitemap ne garantit absolument pas que les URL contenues dans le fichier seront toutes indexées.

Si quelque chose gêne le téléchargement du contenu de la page (partiellement, ou totalement), cela aura aussi forcément un impact sur son référencement. L'idée est donc d'essayer de simuler le comportement d'un crawler de moteur de recherche sur son site, pour détecter tout ce qui peut gêner et/ou bloquer le travail du "spider".

Il existe donc toute une famille d'outils, orientés SEO ou non, qui permettent d'explorer le contenu d'un site et d'en analyser les données.

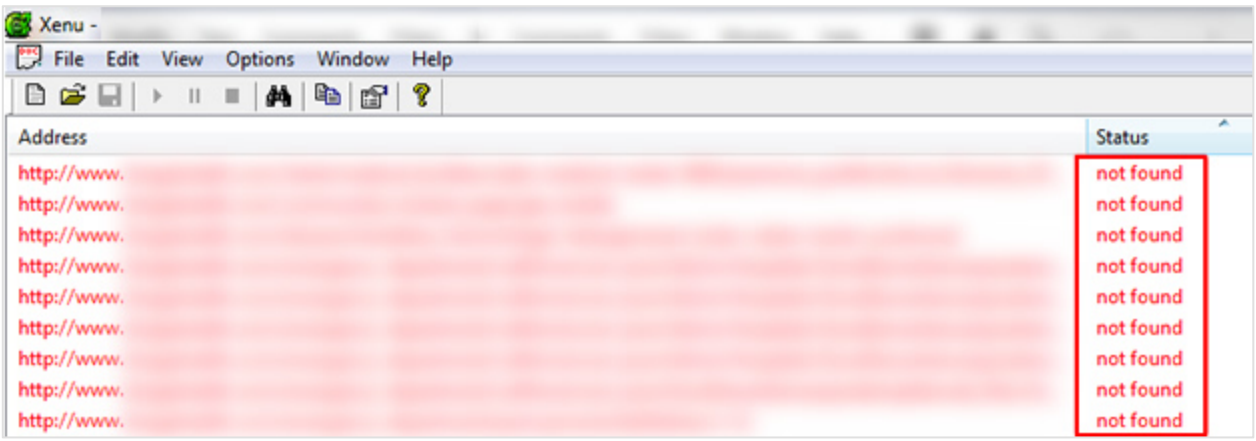
## De quoi faut-il disposer pour réellement simuler ce que Googlebot ou Bingbot peut découvrir et télécharger sur le site ?

Un crawler, pour donner des informations intéressantes pour le SEO, doit explorer les pages en suivant les mêmes règles que Google. Il est donc intéressant de vérifier que l'outil de crawl :
- respecte les directives du robots.txt (ou le fait à la demande) ;
- permet de changer le *user-agent* (pour tester l'arborescence vue depuis un mobile vs celle vue sur un desktop par exemple) ;
- respecte les directives de la balise meta robots et des directives x-robots-tag dans les balises http ;
- respecte l'attribut nofollow dans les liens ;
- est capable de détecter les balises link rel canonical, voire de les interpréter comme des redirections ;
- et idéalement, sait découvrir des liens dans des PDF, des fichier Word, des animations Flashs, du Javascript, etc.

## Les utilisations classiques des crawlers

**Découvrir les liens brisés**

Aucune URL (interne) présentées sur un site n'a de bonne raisons de pointer vers une ressource absente. La détection des urls qui renvoient un code 404 (ou 410) permet donc de résoudre un problème qui impacte les moteurs de recherche et les utilisateurs, et donc de faire la chasse aux liens brisés.



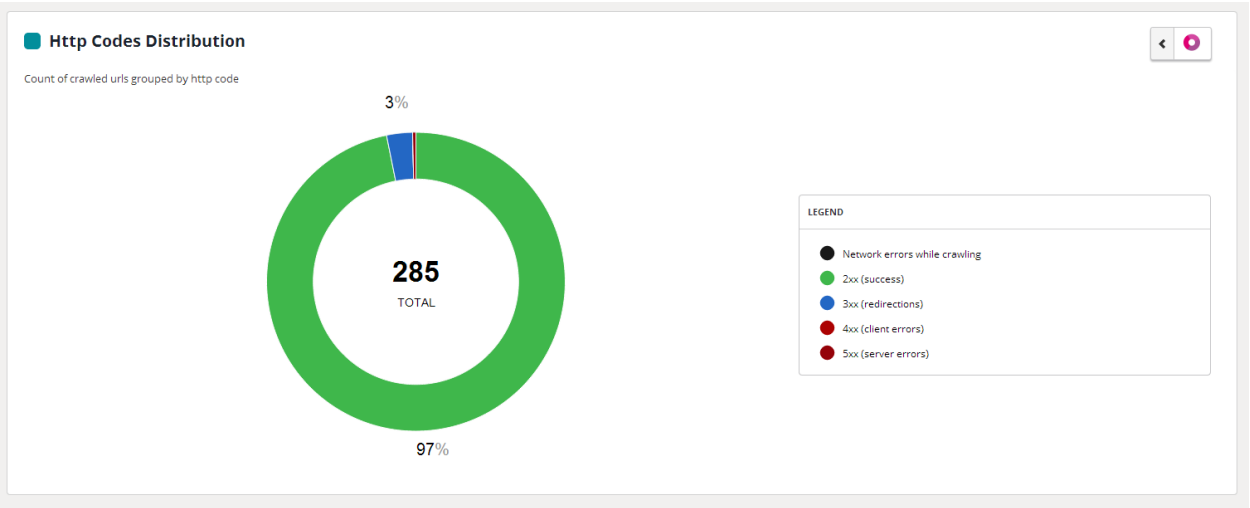

*La détection des liens brisés à l'aide de l'outil gratuit Xenu Link Sleuth*

**Découvrir les pages de redirection et tester les codes de redirection**

Détecter les redirections entre une URL présentée sur un site et une autre URL est également un test qui s'avère utile. En règle générale, le code renvoyé par le serveur doit être "301" et non "302". Repérer les redirections effectuées en 302 permet d'améliorer le comportement des crawlers sur votre plateforme web. Par ailleurs, un grand nombre de redirections peut révéler un problème d'applicatif web (un site bien fait effectue peu de redirections, sauf au cours de période de migrations techniques).

### Tester les codes d'erreurs renvoyés par le serveur web

Outre les codes d'erreur 40x et 30x, il peut être utilise de repérer la présence de "*time out*" (pages que l'on ne parvient pas à télécharger dans un délai donné) et d'erreurs 500 (erreur renvoyé par un serveur web saturé ou un script planté).




*Analyse des codes erreurs renvoyés lors d'un crawl par l'outil Botify.*

### Vérifier l'inventaire des URL explorables, et la connexité du maillage interne

Les outils de crawl du marché ont tous une fonction qui permet de récupérer dans un fichier plat, CSV ou Excel la liste des URL trouvées sur un site donné.

L'analyse de ces fichiers peut révéler très facilement des dysfonctionnements de votre applicatif web, ou des anomalies dans votre maillage interne. Parmi les anomalies classiques, on peut citer :
- la présence de syntaxes anormales, publiées sur le site ;
- l'absence de certaines pages, non explorables ;
- la présence d'autres pages, que l'on ne veut pas voir découvertes par un crawler (pages de back office par exemple).

### Détecter les balises meta manquantes ou dupliquées

La plupart des crawlers orientés SEO permettent de récupérer les infos sur les principales balises meta de la page. Il est ainsi facile de repérer les pages sur lesquelles un <title> ou un <h1> est absent, la présence de descriptions dupliquées d'une page à l'autre, ou le contenu de la balise <meta robots>.

On peut également détecter la présence de liens internes ou externes avec l'attribut "nofollow".

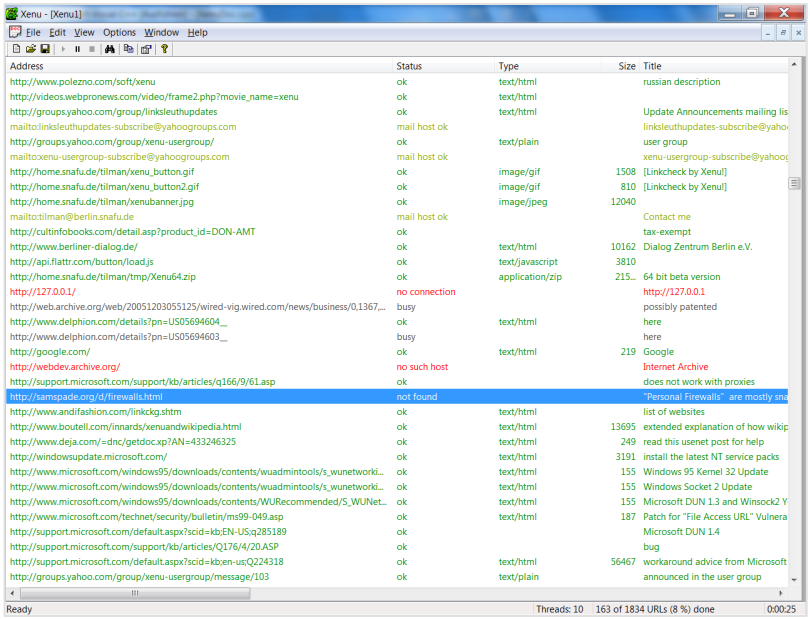
| es | Title 1 | Title 1 Length |
|---|---|---|
| 1 | Délégués \| SEO Camp | 19 |
| 1 | Commissions \| SEO Camp | 22 |
| 1 | Commissions \| SEO Camp | 22 |
| 1 | Commissions \| SEO Camp | 22 |
| 1 | Commissions \| SEO Camp | 22 |
| 1 | Annuaire des Formations Référencement 2010 \| SEO Camp | 53 |
| 1 | Annuaire des Formations Référencement 2010 \| SEO Camp | 53 |
| 1 | Annuaire des Formations Référencement 2010 \| SEO Camp | 53 |
| 1 | Annuaire des Formations Référencement 2010 \| SEO Camp | 53 |
| 1 | Annuaire des Formations Référencement 2010 \| SEO Camp | 53 |
| 1 | Annuaire des Formations Référencement 2010 \| SEO Camp | 53 |
| 1 | Annuaire des Formations Référencement 2010 \| SEO Camp | 53 |
| 1 | Annuaire des Formations Référencement 2010 \| SEO Camp | 53 |
| 1 | Inscription d'une formation SEO \| SEO Camp | 42 |
| 1 | Inscription d'une formation SEO \| SEO Camp | 42 |
| 1 | Inscription d'une formation SEO \| SEO Camp | 42 |
| 1 | Inscription d'une formation SEO \| SEO Camp | 42 |
| 1 | Inscription d'une formation SEO \| SEO Camp | 42 |
| 1 | Formation \| SEO Camp | 20 |
| 1 | Formation \| SEO Camp | 20 |
| 1 | Formation \| SEO Camp | 20 |
| 1 | Formation \| SEO Camp | 20 |
| 1 | Formation \| SEO Camp | 20 |
| 1 | Formation \| SEO Camp | 20 |

*Détection des balises dupliquées à l'aide de Screaming Frog*

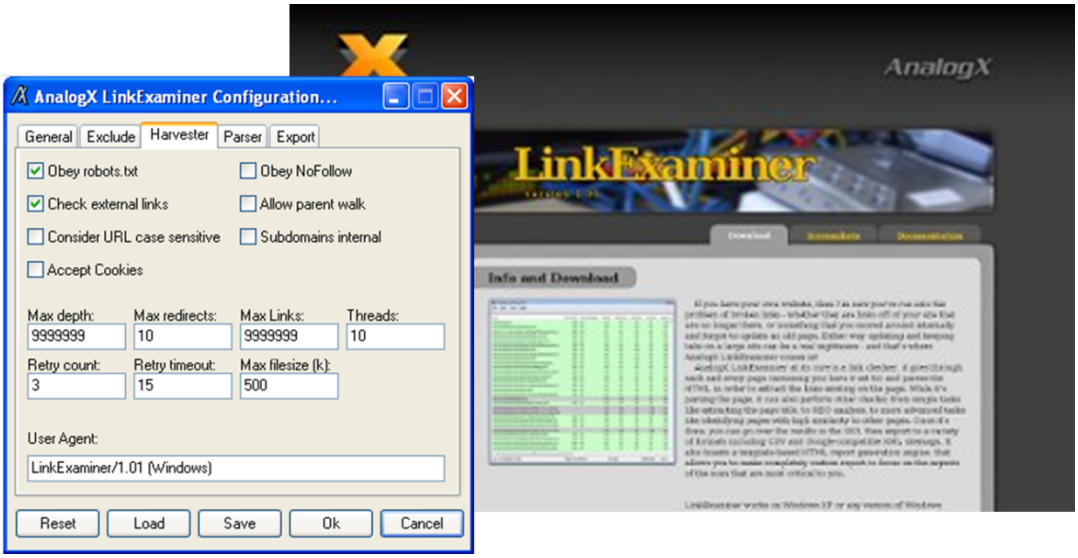
## *Quelques exemples de crawlers communément utilisés pour auditer des sites*

### Les crawlers gratuits

***Xenu Link Sleuth*** (http://home.snafu.de/tilman/xenulink.html) est un outil gratuit (et incontournable) développé il y a une quinzaine d'années. Ce petit exécutable permet de crawler facilement des sites de taille moyenne. Il sert essentiellement pour découvrir des liens brisés et les codes d'erreur envoyés par les urls, par contre il est dépourvu de fonctions SEO avancées.

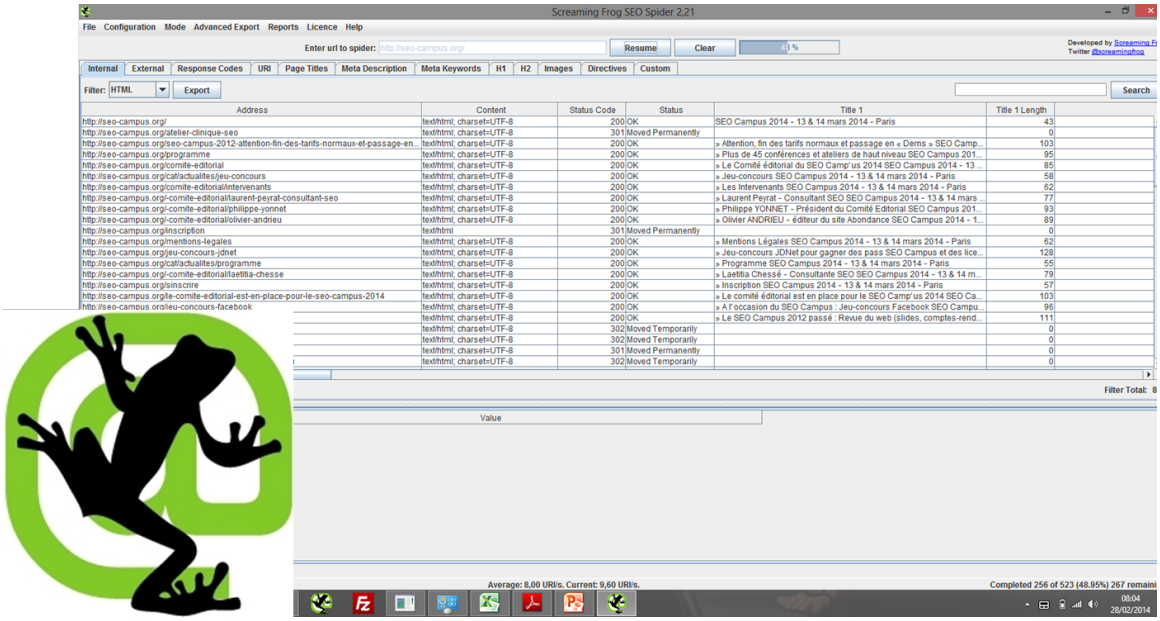



***Link Examiner d'Analogx*** (http://www.analogx.com/contents/download/Network/lnkexam/Freeware.htm) est un autre outil gratuit, mieux doté en fonctionnalités SEO, mais qui se révèle bogué à l'usage et difficile à utiliser sur des sites de plus de 50 000 URL.




***Siteliner*** (http://www.siteliner.com) est un outil en ligne, de type Saas, gratuit pour une utilisation limitée, payant sinon. Il ne permet pas de crawler de gros sites. Son principal intérêt est de disposer d'un outil de détection des pages doublons et des quasi doublons.

***Screaming Frog*** (http://www.screamingfrog.co.uk/seo-spider/) est certainement l'un des outils de prédilection des experts SEO. Pour une licence vendue à un prix accessible, on dispose d'un outil doté de nombreuses fonctionnalités avancées pour les audits SEO. Par contre, il trouve ses limites sur les gros sites, car il est tributaire de la mémoire disponible sur la machine : il peut donc se bloquer au-delà de 50 000 URL sur certaines machines, et parvenir à crawler jusqu'à 1 million d'URL avec une machine dotée d'une grande capacité en mémoire vive.




Pour crawler des sites plus gros (au-delà du million d'urls), il reste deux solutions :
- Personnaliser des outils open source.
- Ou passer par des outils du marché dédiés aux professionnels.

**Les outils open source personnalisables**

Ces outils demandent des connaissances en programmation.

***Lucène*** (https://lucene.apache.org/) : un outil de crawl et d'indexation écrit en Java.

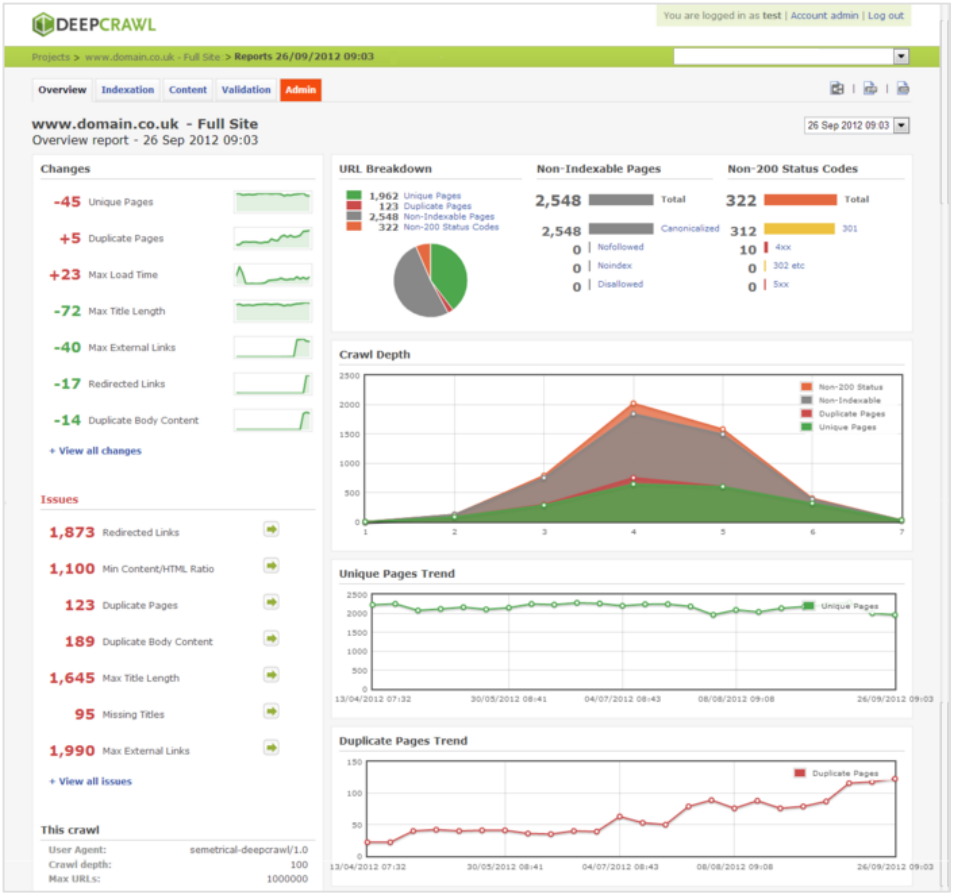
***Nutch*** (https://nutch.apache.org/), un « fork » de Lucène, toujours écrit en Java, et dotés de fonctionnalités de crawl avancées.

***Scrapy*** (http://scrapy.org/), une bibliothèque en Python permettant de créer des crawlers personnalisés.

**Les outils du marché conçus pour de gros sites**

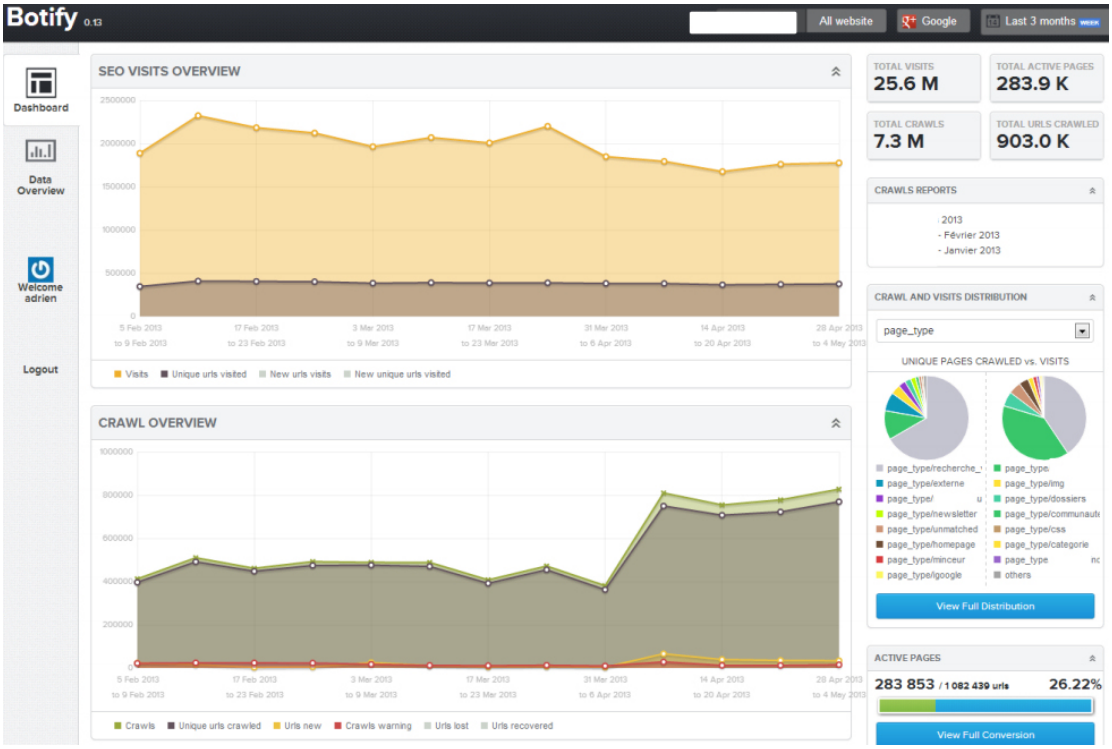
***80legs*** (http://80legs.com/) est un crawler universel en mode SaaS que l'on peut personnaliser à l'aide de « plugins » et de codes personnalisés pour le transformer en crawler avancé doté de fonctionnalités SEO .

***Deepcrawl*** (http://deepcrawl.co.uk/) est un peu plus qu'un crawler : c'est un outil d'audit et d'analyse complet pour le SEO, fonctionnant en mode SaaS, qui fournit toute une série de rapports sur la santé de votre site à partir des données recueillies par le crawler. Il a cependant un défaut : la licence de cet outil d'origine britannique est très élevé.




*Un aperçu des rapports de Deepcrawl*

***Botify*** (http://www.botify.com/) : un outil créé par des référenceurs français, doté lui aussi de fonctions de reporting SEO avancés.  La nouvelle version en mode SaaS est proposée depuis le début de l'année avec une licence beaucoup plus accessible que Deepcrawl.




## *Beaucoup de ces informations sont déjà présentes dans les webmaster tools, qu'est-ce que le crawler m'apportera de plus ?*

Les Comptes Webmaster Tools de Bing ou de Google contiennent certes beaucoup d'infos et d'indicateurs sur le crawl de Googlebot ou de Bingbot. Déjà, ces outils ne fournissent pas toutes les informations que l'on peut récupérer grâce aux crawlers les plus sophistiqués.

Mais surtout, c'est la comparaison des données renvoyées par votre crawler et celles contenues dans Webmaster Tools qui permet de mettre le doigt sur des anomalies qui nuisent à votre bon référencement.

Et c'est ce que nous étudierons dans le cadre du prochain article, qui décrira les utilisations avancées des crawlers pour le SEO. Rendez-vous dans un mois !

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