KillDuplicate : un outil pour monitorer son duplicate content externe

Par Paul Sanches

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

Le duplicate content externe (copie de vos contenus sur d'autres sites) est un fléau qui peut parfois vous poser des problèmes conséquents en termes de positionnement Google. Le moteur de recherche ne pénalise pas ce type de pratique (au sens de punition pour spam), mais il ne donne la visibilité qu'à une seule source d'informations. Et cela peut ne pas être vous, même si vous en êtes la source originale ! Pour éviter cela, l'outil KillDuplicate propose une suite complète d'outils pour détecter et éradiquer ce type de problème. Description par son concepteur...

Tout d'abord, qu'est-ce le duplicate content ou contenu dupliqué ? Voici la définition que nous en donne Google : *« Par contenu en double, on entend généralement des blocs de contenu importants, appartenant à un même domaine ou répartis sur plusieurs domaines, qui sont identiques ou sensiblement similaires. »*
Source : https://support.google.com/webmasters/answer/66359?hl=fr

Historique de l'outil

La première fois que j'ai rencontré des problèmes de contenu dupliqué, c'était en 2006 avec le site d'un client dans le domaine photovoltaïque. Après lui avoir fait son site, le travail de référencement commençait. Le site était premier sur la marque et quelques requêtes long-tail au bout de quelques jours. Un mois plus tard commence l'habituel rituel du matin : tracking des positions, trafic des sites clients et persos… constat : le client ne se positionne plus. A la saisie de la marque du site client dans Google, surprise, stupéfaction : un autre site apparaît en première position. Un clic sur ce résultat fait constater une copie conforme du site du client, sur lequel on peut surfer comme si de rien n'était… sauf qu'on n'était pas sur le domaine du client mais sur ce que l'on appelle un webproxy. Mon client venait de se faire dupliquer son site.

Un webproxy ?

Un webproxy, pour ceux qui ne connaissent pas, est un site qui est à l'origine créé pour surfer anonymement sur le web. Ppar exemple : vous êtes à la fac et l'administrateur réseau n'autorise pas le surf sur youtube, vous allez alors utiliser un webproxy pour ce faire. Vous en trouverez de nombreux sur cette adresse : http://proxy.org/.

Mais le souci avec les webproxies est que certains mettent en cache toutes les pages que vous visitez et remplacent les liens internes de ces pages par les leurs. Ces pages vont ensuite être crawlées et indexées par Google, voire même être créées grâce à Googlebot qui va parcourir tous les liens présents sur chacune des pages mises en cache.

Imaginez ensuite que ce webproxy va, au bout de quelques mois, avoir plusieurs dizaines voire centaines de milliers de pages indexées. Il va donc forcément prendre de l'autorité auprès de Google et donc potentiellement prendre la place de petits sites ou se positionner sur des requêtes « longue traine » et ainsi générer du trafic et pourquoi pas de l'argent.

Bien sûr, ce site se fera pénaliser au bout d'un certain temps. Mais il aura fait quelques dégats en attendant...

Google est impuissant face aux problèmes de duplicate content externe

Il existe de nombreux exemples où le duplicate content peut faire du mal au positionnement d'un site. Exemple sur la figure 1 avec la copie d'écran du célèbre cas où le site Google Maps avait disparu sur la requête [Google Maps] suite à la duplication du site par un WebProxy. Seule la version anglaise de Google Maps ressortait sur la requête Google Maps sur Google.fr (l'adresse maps.google.fr était dégagée par le webproxy).

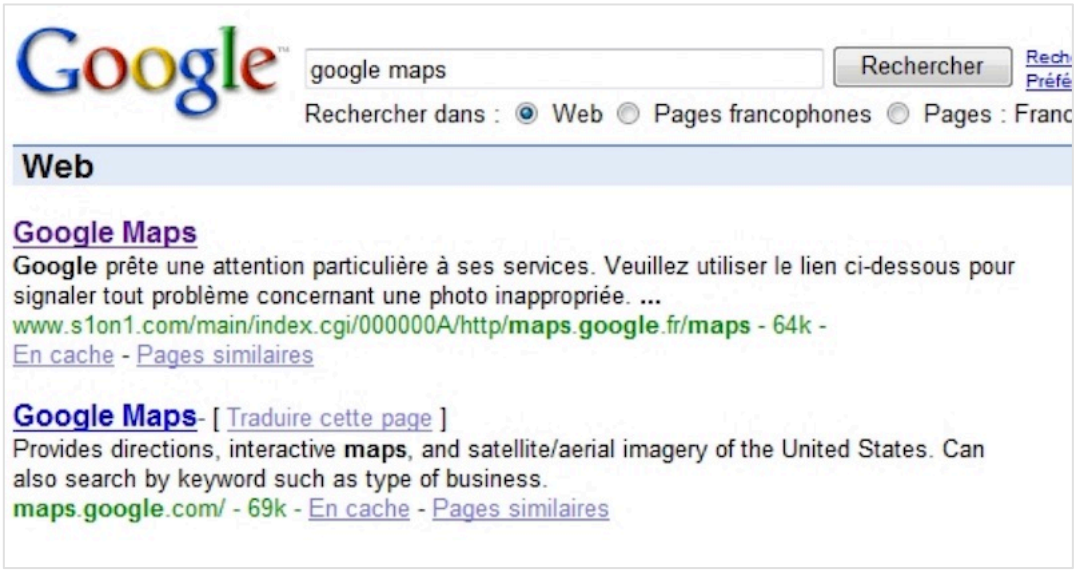

Fig.1. Google Maps remplacé dans la SERP par un webproxy.

La figure 2 montre un autre exemple de l'impuissance de Google à traiter le problème : Google proposait aux webmasters de dénoncer les sites scrapers (voleurs de contenu) qui se positionnaient mieux que les leurs. Depuis, ce formulaire a été fermé.

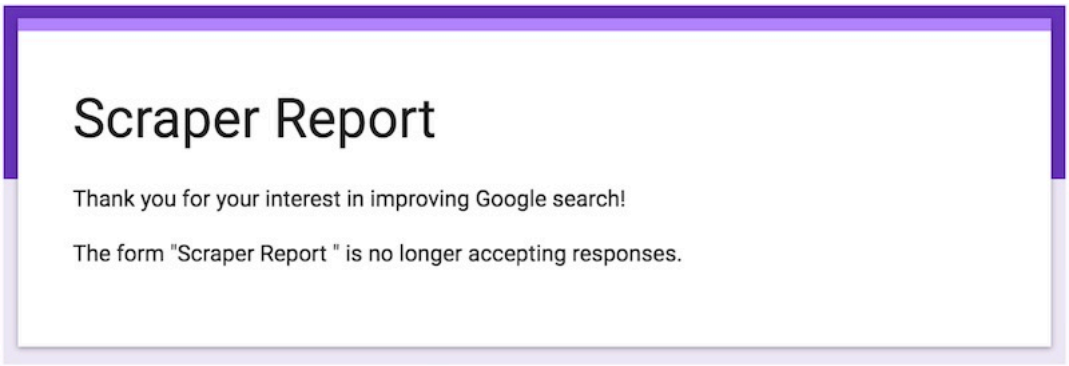

Fig.2. Google Maps remplacé dans la SERP par un webproxy.

Un autre exemple : un test de duplicate fait sur la page d'accueil du site de Matt Cutts, voir pour plus d'infos :
http://www.seoblackout.com/2014/02/01/negative-seo-duplicate-content-mattcutts/

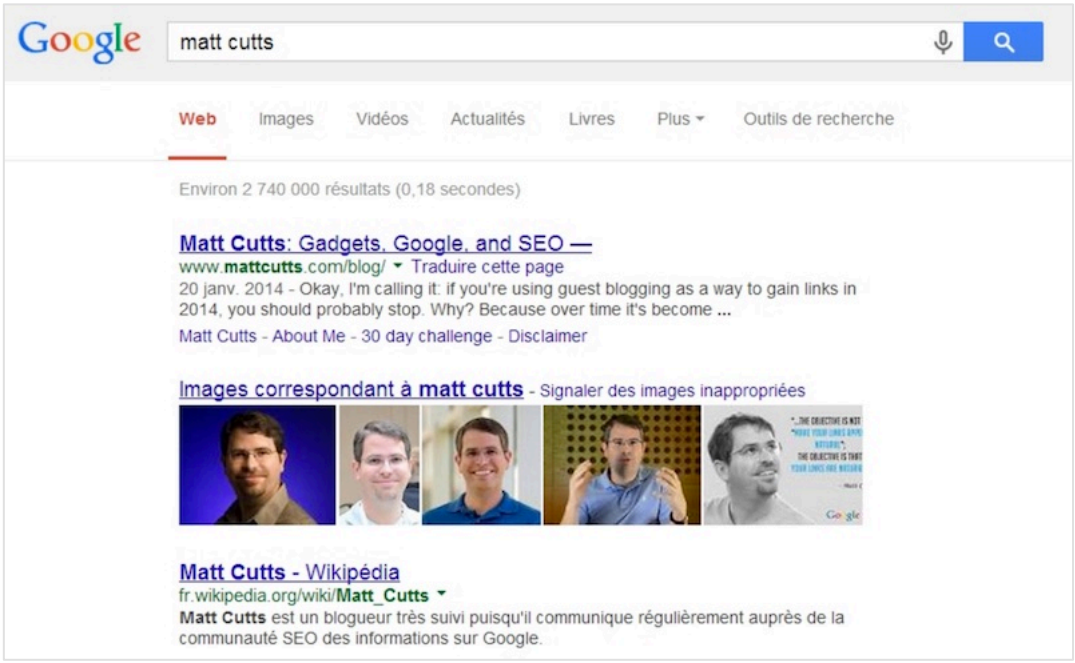

Fig.3. Comment créer quelques soucis au site perso de Matt Cutts.

Et un dernier exemple avec les polonais... Des webmasters se sont plaint un jour que leurs sites avaient disparus de Google. Leur contenu avait été placé sur des sites polonais hackés, qui avec parfois de l'autorité se positionnaieent à la place du site original qui, lui, disparaissait totalement des SERP.
Source : https://www.seroundtable.com/google-investigating-polish-hacker-who-is-stealing-webmaster-traffic-21601.html

A hacker hacks some website (let's call it "website A" - the website that is hacked) and he puts there subpages with copies of our (Polish webmasters') websites (let's call my websites as "website B"). The copies are put into an iframe. We (owners of "website B") are not hacked by him, he just copies code of our website and puts in the iframe on "website A". Now, the problem is that Google algorithm in many cases considers the malicious copy put by hacker on website A as THE ORIGINAL and the website B (our website which is the original) disappears from Google results. What's more, even after removing the hacked address of website A from the Google index our website (B) does not come back to the Google ranking.

So we are disappearing from the Google ranking without any fault of us (the hacker does not hack us but somebody else). The hacker hacks other servers because he doesn't want to be caught but in fact he could do this (make copies of our websites) on his own server and then due to algorithms failure we also would disappear from the results.

So we would like someone in Google to fix this problem because it's getting worse every day and it considers many webmasters. We are innocent victims of the algorithm problem - algorithm reads the copy as the original and then removes the original from the ranking.

Fig.4. Disparition de sites pour cause de contenu dupliqué.

Google affirme qu'il n'existe pas de pénalité pour contenu dupliqué. Et c'est vrai, il n'existe pas de pénalité ni pour le duplicate externe ni l'interne, mais le résultat est le même, sur une requête donnée, Google n'affichera pas 2 pages avec le même contenu dans la première page des SERP, il choisira la page qui a le plus d'autorité.

Et ce problème du duplicate content externe n'est toujours pas résolu chez Google. Un sondage sur Twitter indiquait ce qu'en pensaient les autres SEO :

Fig.5. Sondage Twitter sur le contenu dupliqué externe.

La figure 6 montre les commentaires laissés sur le sondage (source : https://twitter.com/Seoblackout/status/720902328040230914).

Fig.6. Commentaires sur le sondage Twitter sur le contenu dupliqué externe.

A cette même question, Zineb (Google Webmaster Trends Analyst à l'époque), répondit lors du VLC 2015 : *« Si un autre site a le même contenu que le vôtre et qu'il a de meilleurs signaux de positionnement que le votre, il risquera de se positionner mieux que vous. »*

Google ne trouve donc pas ou ne cherche pas de solutions viables pour régler cette histoire de paternité de contenu. Quand on demande à John Mueller (*Webmaster Trends Analyst* chez Google Zurich) pourquoi ne pas mettre en œuvre tel ou tel système d'authentification du contenu original, il répond en plaisantant à moitié que les spammeurs « font indexer leurs pages alors que vous n'avez pas encore appuyé sur le bouton "envoyer" pour publier votre article ».

Comment résoudre les problèmes de duplicate content externe ?

Pour gérer le sproblèmes de duplicate content externe, la première chose à faire est de monitorer ses contenus, soit à la main soit à l'aide d'outils. Imaginons que vous ayez un petit site d'une dizaine de pages, dans ce cas, vous pourrez effectuer ce travail manuellement. Vous allez copier des phrases au hasard issues de vos pages et les coller dans Google entre guillemets pour voir qui a repris votre contenu.

On trouve plusieurs types de « repreneurs de contenu » : les « gentils » et les « méchants ». Les gentils vous feront un vrai lien, suivi par les moteurs de recherche. Les méchants vous montreront une fausse page 404, une redirection 301 cloakée vers une page de conversion… Ils peuvent également mettre votre contenu sur des sites hackés avec une forte autorité.

Vous allez donc parcourir chaque résultat et visiter chaque site pour voir qui vous a pris votre contenu (est-ce qu'il a tout repris, est-ce un agrégateur ?...). Bref, vous regardez si les liens sont en dofollow (vérifier dans le cache Google) et dans ce cas, pas de souci.

En revanche dans tous les autres cas, il vous faudra évaluer la nocivité de ce duplicate et prendre les mesures nécessaires. Voici les 3 étapes classiques :

1 – Contacter le webmaster par Whois, formulaire de contact, twitter, etc. ;
2 – Contacter l'hébergeur du site ;
3 – Plainte DMCA auprès de Google, Facebook, Wordpress...

Ce travail de monitoring prend du temps, il faut surveiller le site dans le temps, gérer les contacts, archiver les demandes traitées…

C'est là qu'arrive Kill Duplicate...

Voici la raison de la création de l'outil KillDuplicate (https://www.killduplicate.com/fr), pour pouvoir monitorer de façon automatique des sites persos et les sites de clients. L'outil est pensé pour faire exactement tout le process fait manuellement pour monitorer les sites et corriger les soucis de duplicate.

Chaque semaine l'ensemble des sites gérés en interne sont scannés. Un rapport détaillé de ce qui se passe pour chaque page est fourni, avec des solutions proposées.

Sur le site, il y a écrit : « le meilleur outil d'analyse du duplicate content du marché » ?

Il n'existe pas à l'heure actuelle de solution aussi complète et dont les résultats sont aussi probants. Les concurrents utilisent les API Google, du coup les résultats donnés ne se font pas en temps réel et ne sont pas exhaustifs. Pire certains utilisent les résultats donnés par Yahoo et non Google (qui est pourtant le moteur de recherche utilisé par 95% de la population française, et 90% de la population mondiale). Avant de lancer le développement de cette solution, nous avons testé la plupart des outils disponibles sur le marché. Le résultat est simple : les résultats donnés sont incomplets, parfois faux, et aucun de ces outils ne va assez loin dans l'analyse et la gestion de projet. Voilà, ça c'est dit... :-)

Comment fonctionne techniquement l'outil ?

Selon l'abonnement, le client va renseigner les URL à scanner toutes les semaines. Pour le plus petit abonnement, on démarre avec un scan de 100 URL par semaine. Le client a la possibilité, si son budget est limité, de changer ces URL chaque semaine. Dès que le scan est terminé, on peut constater le duplicate présent ou pas sur chaque URL, le pourcentage, l'entête http renvoyé par la page, la présence ou non d'un lien nofollow, etc.

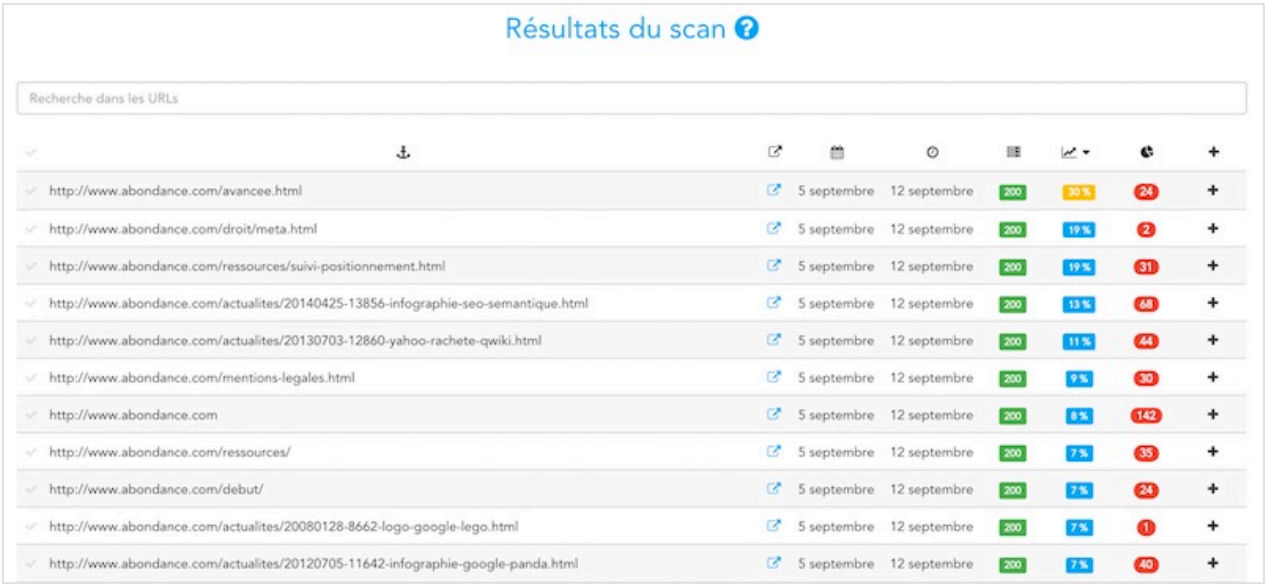

Fig.7. Page de résultats du scan du site Abondance.com

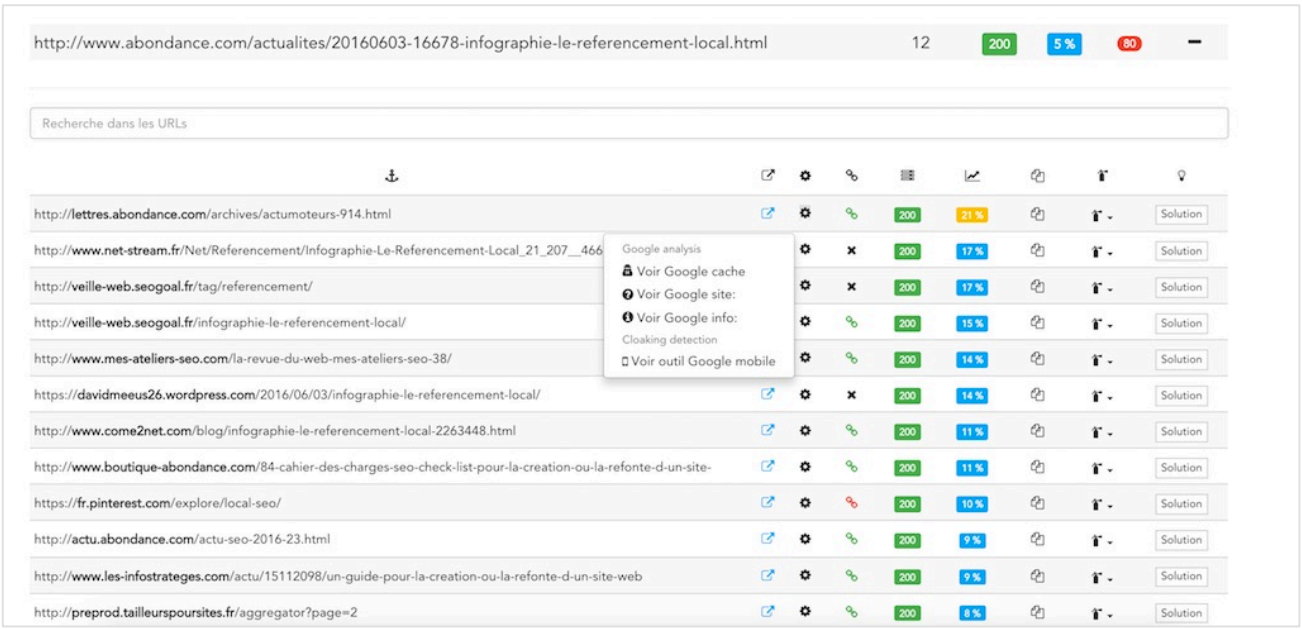

Fig.8. Page de résultats pour une URL donnée

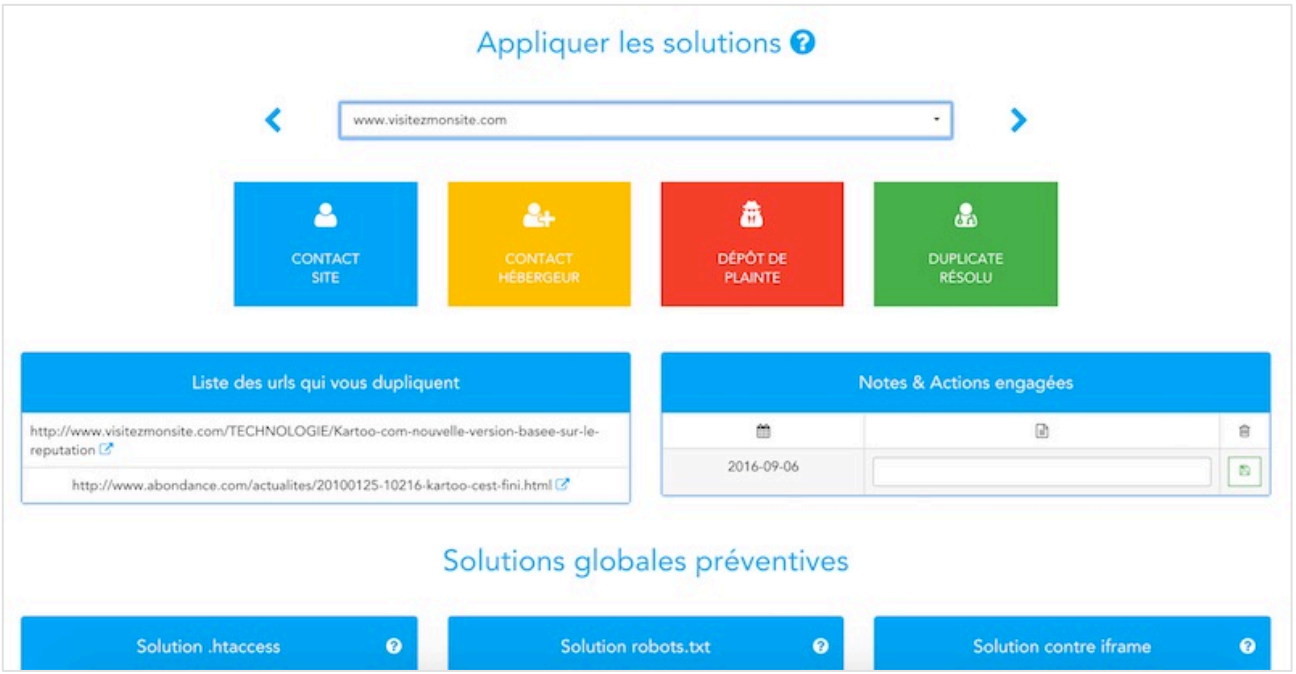

Fig.9. Page des Solutions

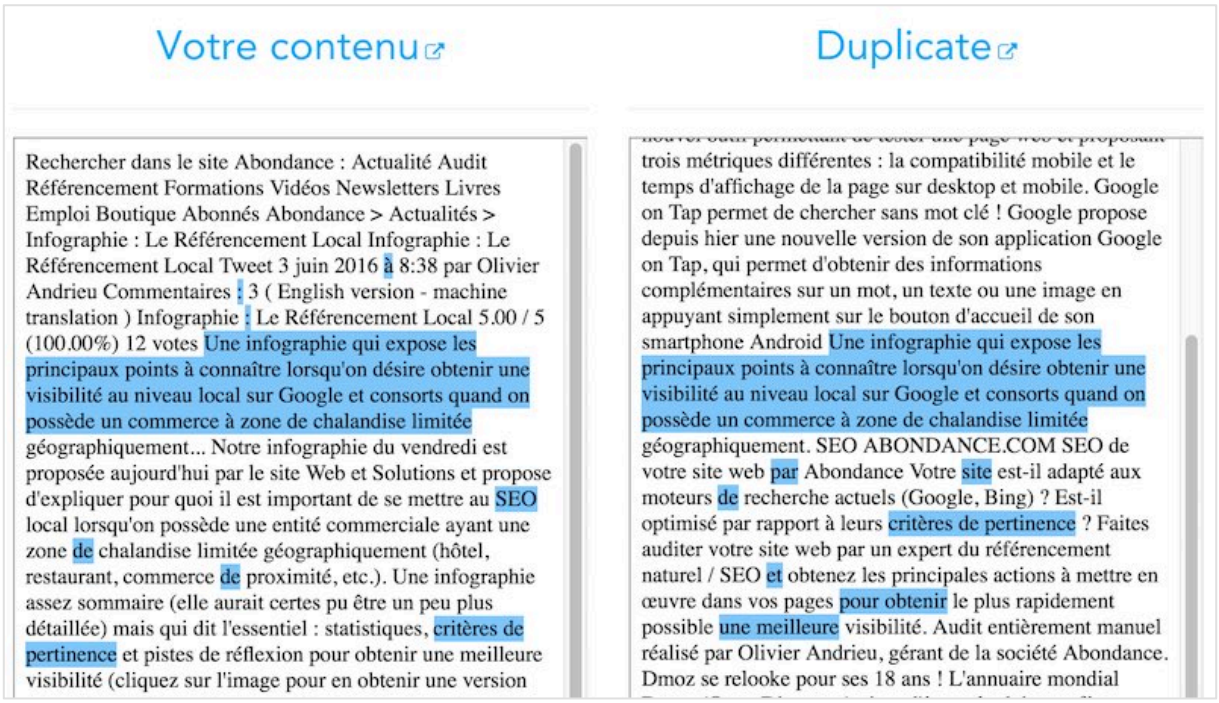

Fig.10. Visualisation du duplicate

Kill Duplicate ne montre pas uniquement le contenu volé. En effet, tout ce que l'on devait jusqu'ici faire à la main est automatisé :

- Analyse du duplicate (pourcentages, présence ou non backlink…) ;
- Scan hebdomadaire ;
- Visualisation & comparaison du duplicate ;
- Traitement par lot ;
- Proposition de solutions personnalisés et automatiques ;
- Identification des voleurs ;
- Génération de mail automatique pré-rempli (en français ou en anglais) pour demander la suppression de la page dupliquée ;
- Gestion dans le temps.

Et une Api en prime !

Avec l'API, vous pouvez utiliser Kill Duplicate dans vos applications. Vous pouvez également, avec les crédits API, tester des textes (plutôt que des URL) directement sur le site.

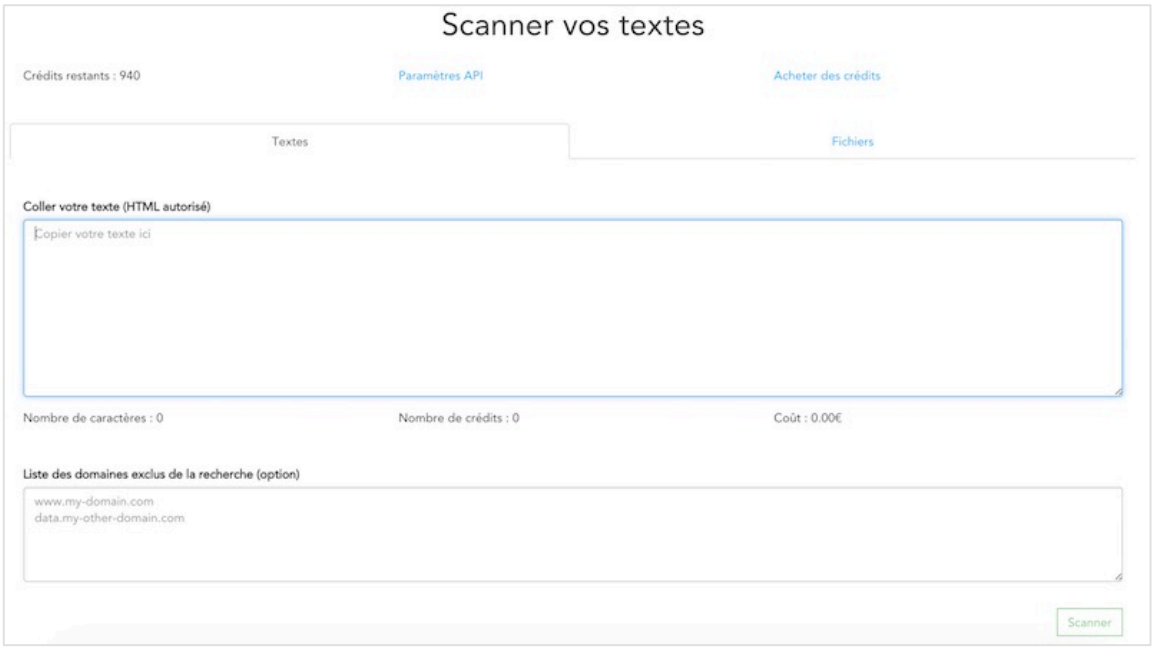

Fig.11. Scan de texte pour détection de duplicate content.

A qui s'adresse cet outil en particulier ?

KillDuplicate s'adresse à toute personne ayant un site web et souhaitant se faire connaître sur Google. Nous nous adressons aussi bien aux webmasters, aux référenceurs, aux professeurs, aux rédacteurs web, aux plateformes de rédaction et à tous ceux qui ne supportent pas qu'on leur vole leur travail. Les référenceurs peuvent également utiliser l'outil en avant-vente pour leurs audits par exemple.

Nous nous adressons également aux agences de communication qui peuvent réaliser un monitoring efficace et rapide de tous les sites de leurs clients et établir une stratégie de contenu et une veille concurrentielle SEO.

Sans oublier les e-commerçants qui souvent sont les premières cibles des voleurs de contenu.

L'outil peut également être détourné à d'autres fins. Les journalistes, par exemple, peuvent l'utiliser pour suivre la reprise de leur article et voir qui a publié et partagé l'information.

Comment l'outil peut « aider à enrayer les problèmes de duplicate content et de plagiat » ?

Simplement parce que lorsqu'on lance l'outil sur un site et qu'on contacte le webmaster, dans 95% des cas, on obtient une réponse positive à la demande et le contenu est retiré ou le backlink ajouté. Et quand on n'a pas de réponse, ce qui est parfois le cas, le passage par le bouton « contacter l'hébergeur » résoud le plus souvent le problème en moins de 24h. Dernière option possible, déposer une plaine DMCA auprès de Google (https://support.google.com/legal/troubleshooter/1114905?hl=fr) ou autre (Facebook, Wordpress, Twitter…).

Une solution automatique d'envoi de mails

Une solution automatique est possible lorsqu'il s'agit par exemple de sites qui viennent « scraper » automatiquement. On a une liste de sites qui dupliquent nos clients, on analyse ceux qui dupliquent le plus et on propose une solution automatique pour ces sites. On peut détecter leur IP et les bloquer dans le htaccess ou équivalent. En bloquant l'IP, on va empêcher le scrapeur de récupérer le contenu et quand Google va à nouveau crawler la page, le contenu dupliqué n'existera plus. Lorsqu'aucune solution automatique n'est proposée, il faut alors se tourner vers l'envoi d'un email aux webmasters, à l'hébergeur ou passer par le dépôt de plainte DMCA.

Conclusion : comment éviter le duplicate content ?

Il est possible de l'éviter mais les solutions actuelles sont beaucoup trop contraignantes : montrer un contenu différent à GoogleBot - ce que l'on appelle le cloaking - est une solution contraignante et pas en accord avec les guidelines de Google. La meilleure solution reste de surveiller régulièrement son site et d'analyser le duplicate puis d'y apporter une solution. Les webmasters réagissent rapidement, comme les hébergeurs, un coup de téléphone ou un mail et c'est réglé la plupart du temps très rapidement.

Paul Sanches, *consultant SEO, Impact SEO (http://www.impactseo.fr). Twitter : https://twitter.com/Seoblackout*