Duplicate Content et Référencement (2ème partie)

Retour au sommaire de la lettre

Tout éditeur de site web de contenu a eu, a, ou aura un jour à faire avec la notion de "duplicate content" sur les moteurs de recherche. En d'autres termes, la prise en compte par Google et consorts d'une seule version d'une page proposant un contenu qui est dupliqué à l'identique - ou presque - dans un autre document se trouvant sur le même site ou sur une autre source d'information. Quelles sont les différentes formes (nombreuses) de "duplicate content" ? Quelles solutions apporter pour éviter ce type de souci ? Cette série d'articles tente d'y voir plus clair sur cette problématique que nous avons initiée, dans la lettre précédente, par une explication générale du phénomène de "duplicate content" et la problématique du contenu "canonique" dupliqué par des sites partenaires ou "pirates"... Ce mois-ci, nous tentons de comprendre pourquoi une même page accessible depuis des urls différentes pose problème aux moteurs de recherche et à votre référencement...

Duplicate content, acte 2

Le mois dernier, nous avons exploré le concept de Duplicate Content et la problématique du contenu canonique dupliqué sur des pages d'autres sites, qu'ils soient partenaires ou non. Mais il ne s'agit pas là de la seule problématique gravitant autour du phénomène de Duplicate Content, loin de là...

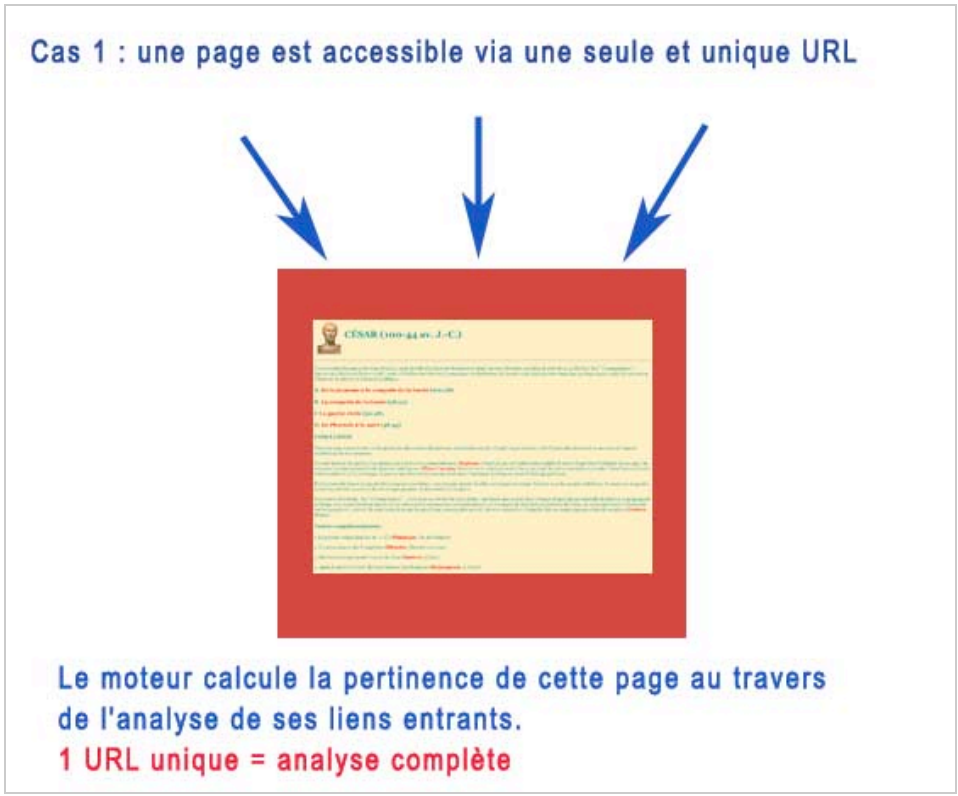
En effet, les webmasters sont souvent confrontés au fait qu'une même page web, unique – proposant strictement le même code HTML - , soit accessible par des urls différentes sur un même site, bien évidemment. Cette situation est dommageable pour un référencement et nous allons tenter d'expliquer pourquoi.

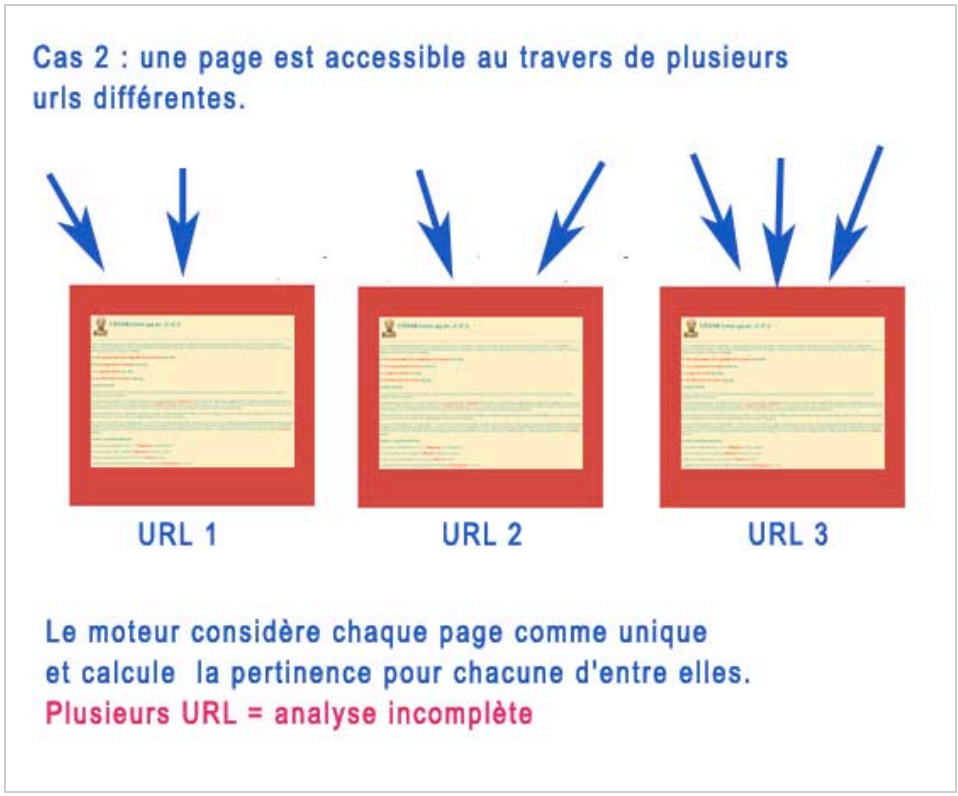
Duplicate interne, quelle incidence sur le référencement ?

On le sait, les algorithmes de pertinence des moteurs de recherche majeurs sont fortement influencés par l'analyse des liens externes et internes qui pointent vers une page et les notions de popularité (quantité et qualité des liens entrants) et réputation (textes des liens pointant vers la page). Pour prendre un premier exemple simple, chaque lien vers votre page d'accueil est une pierre de plus à ajouter à la qualité de votre référencement.

Ainsi, imaginez que votre page d'accueil (www.votresite.com) redirige l'internaute, à l'aide d'une redirection JavaScript ou 302, vers une page pour les internautes francophones disponible à l'adresse www.votresite.com/fr/accueil.php. La plupart des liens, sur le Web, pointeront pourtant vers l'adresse www.votresite.com, qui n'a pas de contenu puisqu'elle redirige automatiquement l'internaute. La "vraie" page d'accueil (www.votresite.com/fr/accueil.php), recevant peu de liens (les redirections JavaScript et 302 ne transférant pas la popularité d'une page à l'autre), elle sera très peu populaire, réduisant donc d'autant sa capacité à être bien positionnée. La page certainement la plus populaire de votre site n'a donc ici aucune utilité en termes de référencement. Rageant, non ?? Il en serait, bien sûr, autrement si une redirection 301 (qui transfère le PageRank) avait été utilisée...

De même, si votre page d'accueil, pour reprendre cet exemple, est très populaire, elle va, au travers de ses liens internes, transférer de la popularité (on parle de "*Link Juice*" ou "jus de lien") aux pages internes vers lesquelles elles pointent. Ce "Link Juice" transmis par les liens internes est important pour les moteurs de recherche. Or, si une même page est accessible par plusieurs adresses différentes, elle sera considérée comme autant de documents différents par les moteurs de recherche. Donc un document unique A sera "vu" par Google et consorts comme plusieurs pages A', A'' et A''', par exemple, chacune de ces pages détenant une fraction de la popularité globale de A. Pas top...





Il sera donc important que, sur votre site, CHAQUE PAGE SOIT ACCESSIBLE PAR UNE SEULE ET UNIQUE ADRESSE (URL) afin que l'analyse des liens internes (popularité et réputation) soit la plus fine, juste et efficace possible.

Nous allons voir, dans la suite de cet article, quels sont les cas les plus fréquents de Duplicate Content de ce type et les solutions à y apporter.

Nom de domaine dupliqué

Le premier cas est assez fréquent : votre site est accessible par plusieurs noms de domaines : www.abondance.com, www.abondance.net et www.abondance.fr par exemple. Dans ce cas, la solution est simple : une redirection au niveau de votre DNS ou, mieux, *via* un code 301 est à privilégier. Vous prenez, dans ce cas, en compte pour votre communication un seul nom de domaine (pour nous : abondance.com) et vous redirigez tous les autres vers lui. Les redirections DNS et 301 sont bien interprétées aujourd'hui par les moteurs de recherche qui comprendront aisément que toutes ces adresses pointent vers un seul et unqiue site web. Pas de soucis majeurs.

Nom de site dupliqué

Le deuxième cas est également assez fréquent : votre page d'accueil est accessible sous des adresses de type www.votresite.com ET votresite.com (sans le préfixe "www"). C'est une bonne chose pour l'internaute car cela lui rend la saisie plus simple et plus rapide, mais cela peut aussi créer un phénomène de Duplicate Content pour les moteurs en rendant un seul site accessible sous deux adresses différentes...
Un fichier .htaccess bien conçu avec une règle de réécriture privilégiant l'une ou l'autre adresse (la plupart du temps celle avec "www") résoudra le problème. Exemple :

```
RewriteEngine On
RewriteCond %{HTTP_HOST} !^www\.votresite\.com [NC]
RewriteRule (.*) http://www.votresite.com/$1 [QSA,R=301,L]
```

(*source :* http://www.webrankinfo.com/actualites/200510-contenus-dupliques.htm)

Point important : Google vous propose également, dans ses Webmaster Tools (http://www.google.fr/webmasters/), dans la zone "Outils", le choix "Définir un domaine favori", qui permet d'indiquer au moteur quelle adresse "canonique" vous désirez que Google prenne en compte pour son indexation :

Ceci dit, cet outil n'étant disponible que pour Google et pas chez ses concurrrents, cela ne vous dispensera pas de créer un fichier .htaccess idoine...

Adresse de la page d'accueil dupliquée

De la même façon, votre page d'accueil – toujours elle – est surement accessible à la fois au travers de l'adresse http://www.votresite.com/ mais également d'une adresse de type :
- http://www.votresite.com/index.html
- http://www.votresite.com/index.htm
- http://www.votresite.com/index.php
- http://www.votresite.com/accueil.php
- Etc.

Ces deux adresses (www.votresite.com et www.votresite.com/***) risquent fort d'être considérées comme deux pages différentes également par les moteurs de recherche. Le problème est donc identique au cas précédent et une redirection 301 sera la bienvenue pour n'afficher qu'une seule adresse "canonique" (là aussi, la plupart du temps www.votresite.com).

De la même façon, évitez les liens vers des adresses comme http://www.votresite.com ET http://www.votresite.com/. Le dernier slash ("/") peut, là aussi, poser problème et on n'y pense pas forcément lorsqu'on créé les liens internes du site...

Choisissez donc l'adresse "canonique" que vous voulez pour votre page d'accueil, mais unifiez-la partout sur votre site dans les liens qui pointent vers elle. Google l'explique ici : http://googlewebmastercentral.blogspot.com/2006/12/deftly-dealing-with-duplicate-content.html

Vérifiez également bien qu'au sein même de votre site, les liens, dans les pages internes, qui pointent vers votre page d'accueil pointent bien vers www.votresite.com et non pas vers un autre intitulé d'URL. On a souvent pas mal de surprises suite à cette vérification...

De la même façon, cette problématique peut se reproduire sur des pages internes et notamment des rubriques somaires (www.votresite.com/produits/ et www.votresite.com/produits/index.html) pour lesquelles le remède sera identique. Là encore, du travail de vérification en perspective...

Cas des sites dynamiques

Les sites dynamiques sont très intéressants de par les possibilités d'automatisation qu'ils proposent, mais ils sont aussi souvent source de soucis et de conflits dans les URLS. On dénombre alors plusieurs problèmes qui peuvent subvenir en termes de Dupicate Content :

Cas 1 : paramètres inversés dans l'URL

Par exemple, une même page pourra être accédée selon deux adresses :

http://www.votresite.com/catalogue?ref=123456&pays=fr**&langue=fr**

mais également :

http://www.votresite.com/catalogue?ref=123456**&langue=fr**&pays=fr

Ce sont deux pages exactement identiques, accessibles avec les mêmes paramètre mais pas dans le même ordre dans l'URL. Résultat : deux adresses distinctes et un cas classique de Duplicate Content. Là encore, il vous faudra vérifier, au sein de votre site, l'ordre dans lequel vous passez les paramètres dans vos urls et bien garder, à chaque fois, une stratégie cohérente sur l'ensemble du site à ce niveau.

Cas 2 : pagination des listes

Ce cas est fréquent dans des pages qui listent des produits ou dans des fils de discussion de forums par exemple. Une première page, listant un certain nombre d'"items" (discussions, produits, etc., le meilleur exemple restant encore une page de résultats de moteur de recherche...), sera accessible à l'adresse :

http://www.votresite.com/liste-produits?prod=telephones

Puis, si une deuxième page de produits est disponible, celle-ci sera accessible (*via* le bouton "suivant" par exemple) à l'adresse :

http://www.votresite.com/liste-produits?prod=telephones&page=2

Exemple type d'une page listant des produits

Le problème viendra si l'on revient (au travers du lien "précédent" par exemple) sur la page 1 et que l'on y accède via cette url :

http://www.votresite.com/liste-produits?prod=telephones&page=1

Cette aresse est clairement différente de celle affichée en premier pour la même page... Attention donc à faire en sorte que tout accès à la première page se fasse sous la forme d'une url unique. L'une ou l'autre (avec ou sans le paramètre indiquant le numéro de page) mais surtout unique !

Cas 3 : réécriture d'url

Si vous avez mis en place une réécriture d'url sur votre site, vous devez avoir maintenant des adresses de type :

http://www.votresite.com/catalogue-telephone-nokia-810-kwt-fr-fr

au lieu de :

http://www.votresite.com/catalogue?ref=123456&pays=fr&langue=fr

Très bien et votre référencement ne s'en portera que mieux. Mais n'oubliez pas, pour autant, de mettre en place, dans votre stratégie d'*url rewriting*, une redirection 301 depuis l'ancien intitulé (avec les "?" et les "&") vers le nouveau pour éviter tout souci. Ce serait trop bête d'oeuvrer à créer des urls "propres" pour générer en même temps un phénomène de Duplicate Content...

Cas 4 : plusieurs énoncés d'url pour une même page

Ce cas se rapproche de celui qui concerne la duplication de la page d'accueil, déjà vu auparavant. Souvent, sur une page interne, lorsqu'on clique sur le logo générique ou sur un lien de type "Accueil", on est redirigé vers une adresse de type (ici un exemple sous Lotus Notes) :

http://www.votresite.com/internet/webfr.nsf/0/08112EF3CAE171EBC12573930048A2C9?OpenDocument

au lieu de :

http://www.votresite.com/

Cela peut être dû à plusieurs raisons : vous désirez garder une indication sur la navigation de l'internaute et la page depuis laquelle il revient à l'accueil, votre système de navigation est tout simplement (*sic*) configuré ainsi, des identifiants de session sont automatiquement ajoutés dans l'url, etc.

Là encore, les moteurs de recherche vont identifier votre page d'accueil au travers de plusieurs adresses distinctes, ce qui n'est pas une bonne chose. Le remède est toujours le même :
- Soit simplifier les urls pour toujours indiquer dans les liens leur intitulé "canonique".
- Soit mettre en place des redirections 301 depuis les intitulés "développés" (http://www.votresite.com/internet/webfr.nsf/0/08112EF3CAE171EBC12573930048A2C9?OpenDocument) vers les intitulés "canoniques" (http://www.votresite.com/). Encore une fois, ce conseil est valable pour n'importe quelle page du site et pas uniquement la page d'accueil...

Pour le cas des identifiants de sessions, toujours problématiques en termes de référencement, il faudra peut-être creuser une solution plutôt basée sur des cookies, qui laisses les urls "vierges" de toute indication de navigation et évite le phénomène de Duplicate Content...

IMPORTANT : nous avons beaucoup parlé dans cet article de redirections 301. Elles sont, c'est vrai, très souvent indispensables dans une stratégie d'optimisation de site. Sachez cependant que les moteurs de recherche n'apprécient que très modérement les redirections de ce type en cascade, les unes derrière les autres. Attention donc à ne pas les multiplier, notamment sur la page d'accueil...

Conclusion

Nous n'avons pas abordé dans cet article le cas des contenus "proches" sur un site (par exemple une version Web et une version imprimable d'un même article ou des pages proposant les mêmes TITLE et balises meta "description", etc.). Cela fera l'objet du prochain article de cette série, en octobre 2008, dans cette même lettre.

Notre article de ce mois ne concerne donc que **les pages strictement identiques accessibles au travers d'urls différentes**. Le problème principal, dans ce cas, tourne toujours autour du transfert du *Link Juice* d'une page vers une autre et de l'analyse et la compréhension globale des l'interconnexion des pages de votre site par les moteurs. Une attention toute particulière devra donc être portée à ces problèmes si vous désirez que chacun de vos documents soit jugé à sa "juste valeur"...

Voici, pour terminer, une suite d'articles intéressants (pas si nombreux que cela, finalement, sur le Web), parlant des thèmes évoqués dans cet article, qui vous permettront certainement de creuser ces problématiques (rappelons que notre premier article, le mois dernier, proposait également bon nombre de liens sur le sujet) :

How to Deal with Pagination & Duplicate Content Issues (SEOMoz)
http://www.seomoz.org/blog/how-to-deal-with-pagination-duplicate-content-issues

Pagination and Duplicate Content Issues (Search Engine Journal)
http://www.searchenginejournal.com/pagination-and-duplicate-content-issues/7204/

Lutter contre le duplicate content (Référencement, Design et Cie)
http://s.billard.free.fr/referencement/?2008/04/24/477-lutter-contre-le-duplicate-content

Liste d'erreurs classiques de duplicate content (Webrankinfo)
http://www.webrankinfo.com/actualites/200703-erreurs-de-duplicate-content.htm

Olivier Andrieu
Abondance.com

Réagissez à cet article sur le blog des abonnés d'Abondance :
http://abonnes.abondance.com/blogpro/2008/09/duplicate-content-et-rfrencement-2me.html