Duplicate Content et Référencement (3ème partie)

Retour au sommaire de la lettre

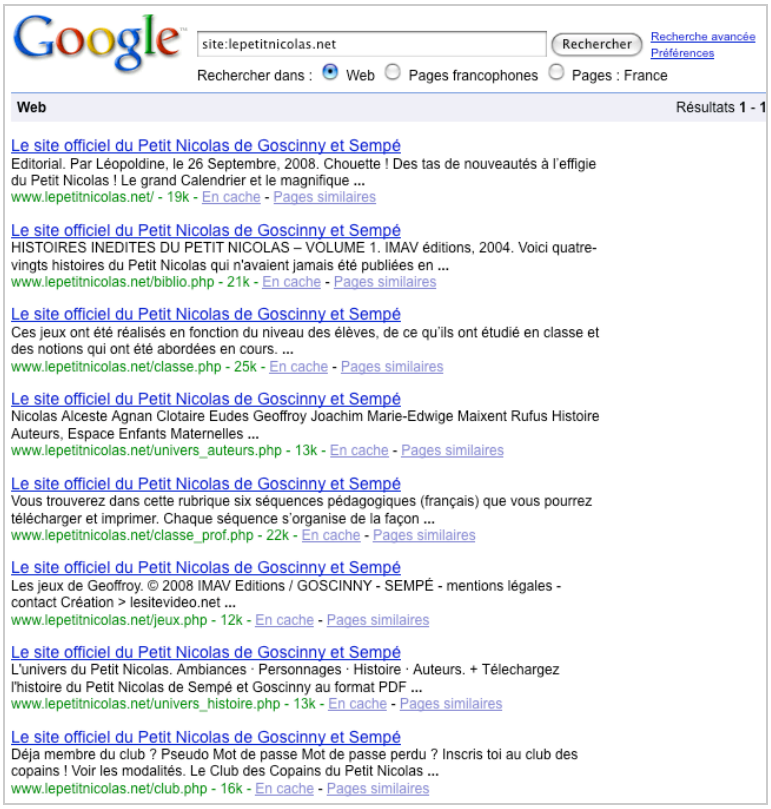
Tout éditeur de site web de contenu a eu, a, ou aura un jour à faire avec la notion de "duplicate content" sur les moteurs de recherche. En d'autres termes, la prise en compte par Google et consorts d'une seule version d'une page proposant un contenu qui est dupliqué à l'identique - ou presque - dans 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 il y a deux mois de cela... Ce mois-ci, nous tentons de comprendre pourquoi certaines pages, au contenu éditorial pourtant différent, sont parfois interprétées comme du "duplicate content" par les moteurs de recherche...

Duplicate content, acte 3

En août 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. Le mois dernier, c'était le fait qu'une même page web soit accessible par des urls différentes sur un même site qui était évoqué...

Pour terminer cette série d'articles, nous allons voir ce mois-ci le cas de pages web proposant un contenu éditorial différent les unes des autres mais pouvant cependant tomber dans les affres des filtres de "Duplicate content", chez Google et ses concurrents.

Commençons par un exemple. La plupart du temps, on comprend vite les risques que l'on court en tapant sur Google (et les autres moteurs) la requête "site:" suivie du nom de domaine de son site. Si les résultats ressemblent à ce que l'on peut voir ci-dessous, vous pouvez commencer à vous faire un peu de mourron :



Dans cet exemple (site:lepetitnicolas.net), chaque page du site a le même contenu pour sa balise TITLE ("*Le site officiel du Petit Nicolas de Goscinny et Sempé*"). Pire encore avec cet exemple (site:chateaudemontvillargenne.fr) :

Dans ce cas, les balises TITLE sont identiques sur de nombreuses pages, mais également les balises meta "description" (reprises dans le "snippet" ou résumé fourni par Google) en-dessous. Faites le test sur votre site pour voir ce qu'il en est...

On pourrait multiplier ce type d'exemple à l'envi... Ils illustrent bien une problématique (souvent présente sur des forums non optimisés par exemple) de "Duplicate content" que l'on trouve finalement assez souvent : les moteurs de recherche, au moment de "filtrer" les contenus identifiés sur le Web, trouvent trop de similitudes dans le code HTML des pages et les classent en "Duplicate content" même si leur contenu éditorial est différent. Google gère le plus souvent assez bien cette problématique, mais il n'en est pas toujours de même de ses principaux concurrents...

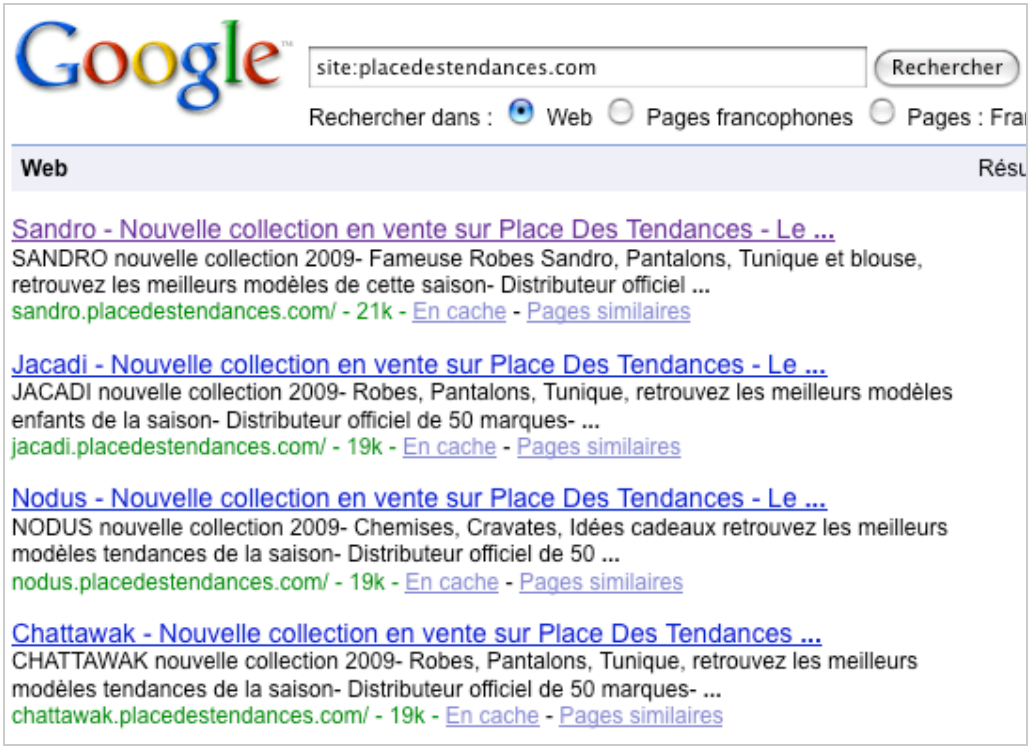
La problématique peut également parfois être similaire mais légèrement différente : vous disposez d'un contenu (par exemple des recettes de cuisine) que vous souhaitez proposer sur plusieurs de vos sites. Comment faire pour que la même recette ne soit pas considérée comme du "Duplicate content" d'un site sur l'autre et que chaque version soit prise en compte par Google ? On le voit, la problématique est ici l'inverse de celle étudiée dans le premier volet de ces articles...

Contenu éditorial différent, structures de page semblables : clairement différencier les codes HTML de chaque page...

Prenons le premier cas : vos pages proposent un contenu textuel (éditorial) différent mais un code HTML trop proche. Comment faire ?

Tout d'abord, il sera essentiel de faire en sorte que le début du code HTML de chacune de vos pages (la partie HEAD) soit clairement différent d'une page à l'autre. Proposez des balises TITLE, Meta description et même Meta Keywords (pour Yahoo!) très descriptives et différentes d'une page à l'autre, que ce soit au niveau de la taille, de la structure et du contenu.

Evitez notamment, si possible, les structures trop redondantes, trop "reconnaissables" et "automatisées" comme ici (site:placedestendances.com) :



Seule une faible portion du contenu des TITLE et balises Meta Description sont ici différentes d'une page à l'autre (des paramètres en faible nombre sont en fait changés sur chaque page). N'hésitez pas à vous différencier plus que cela... Bref, faites en sorte que la partie HEAD de vos pages soit très différente d'une page à l'autre et tout devrait bien se passer...

Ensuite, attaquez le corps de la page (partie BODY du code HTML). Toute la partie "*header*", "*footer*" et "liens de navigation interne" ne devraient pas poser de problème, Google sait différencier ce contenu de la partie plus strictement éditoriale (si Google classait en "Duplicate content" toutes les pages qui ont le même *header* et le même *footer*, il n'aurait plus beaucoup de documents dans son index...)..

N'oubliez pas de donner du contenu en quantité suffisante (100 à 200 mots descriptifs au minimum) aux moteurs. Le titre (H1) et le chapo (premier paragraphe, trois à quatre premières phrases) doivent également être bien différenciés d'une page à l'autre.

Si vous suivez ces conseils, vous ne devriez pas avoir de trop gros problèmes du type de ceux évoqués au début de ce chapitre...

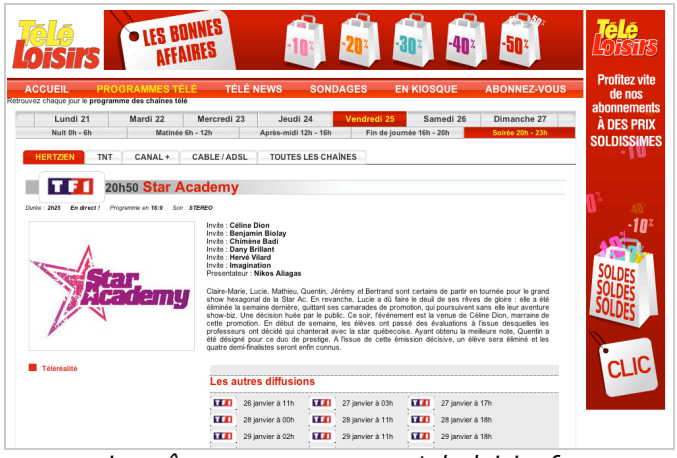
Cas du contenu identique à sauvegarder des filtres de Duplicate Content

Si vous désirez qu'un même contenu sur des pages différentes (ou des sites différents la plupart du temps) ne soit pas considéré comme du "Duplicate content", n'hésitez pas, sur ces pages, à modifier le plus possible ce qui existe "autour du texte" : images, vidéos, liens, encadrés, etc.

Certains changent, lorsque c'est possible, l'ordre des paragraphes (exemples de fiches produits où l'ordre n'est pas essentiel), mais Google a appris également à déceler le "Duplicate content" lorsque l'ordre des informations change, ce n'est donc pas une stratégie gagnante à 100%..

Changez également vos initulés d'url d'un site à l'autre et, ici, le plus possible, les *footer, header* et liens internes (si les contenus similaires se toruvent sur des sites distincts) pour proposer sur chaque site un "environnement" le plus différent possible...

Exemple de programme télé sur le site programme-tv.net



Le même programme sur tele-loisirs.fr.
Comment faire pour que l'un ne "phagocyte" pas l'autre ?

Essayez, le plus possible, de modifier les codes HTML proposés aux moteurs :
- Inversez l'ordre des balises TITLE, META, etc.
- Ajoutez ou supprimez des balises meta peu importantes ("Classification", etc.).
- Ajoutez ou supprimez des commentaires (même si on sait que leur contenu n'est pas lu par les moteurs, ils peuvent changer les séquences linéaires de lecture du code).
- Codez vos pages en UTF-8 ou en ISO-8859-1 selon le cas...
- Proposez des attributs ALT avec des contenus différents pour chaque image.
- La structure des pages (en tableaux ou via des CSS) peut également être totalement différente d'un site à l'autre.
- Etc.

Vous pouvez également ajouter du contenu à l'une ou l'autre page, sur une base commune (des encadrés différents par exemple, ou des infos connexes sur le même sujet). Ce sera autant de travail qui différenciera chaque page... Variez également, si cela est techniquement possible pour vous, les hébergeurs et les adresses IP de vos serveurs d'un site à l'autre.

Attention cependant de faire en sorte que l'une des pages ne pointe pas vers l'autre, ce qui signifierait que l'article qui reçoit le "backlink" est le contenu canonique pour Google (voir la première partie de notre série d'articles).

Il existe en fait des dizaines de façons pour arriver à différencier le plus fortement possible vos contenus et vos pages. A vous de voir, en fonction de vos possibilités, notamment techniques, lesquelles vous pouvez mettre en œuvre...

Pour résumer, il est nécessaire de jouer sur plusieurs niveaux de réflexion pour lutter au mieux contre l'hypothèse du "Duplicate content" dans le cas de contenus similaires sur des sites différents :

- **Contenus éditoriaux** différents : varier le pourcentage entre la partie "texte éditorial" proprement dite et les informations connexes (articles similaires, suggestions, photos, vidéos, fonctions communautaires) d'un site à l'autre.

- **Structure des pages** différentes : code HTML, emplacement des blocs de code, urls, etc.

- **Chartes graphiques** différentes : par exemple ne pas proposer les mêmes liens, mais différemment présentés, etc.

L'évangile selon Saint Google...

Enfin, pour terminer cette série d'articles, n'hésitez pas à lire ce que dit Google dans son centre d'aide pour webmaster (http://www.google.fr/support/webmasters/bin/answer.py?answer=66359&query=dupliqu%C3%A9&topic=&type=) au sujet du "Duplicate content". En voici les extraits qui nous ont semblé les plus importants et intéressants :

Contenu en double

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. À l'origine, la plupart de ces contenus ne sont pas malveillants. Exemples de contenu non malveillant :
** forums de discussion pouvant générer à la fois des pages normales et des pages "raccourcies" associées aux mobiles ;*
** articles en vente affichés ou liés via plusieurs URL distinctes ;*
** versions imprimables uniquement de pages Web.*

Dans certains cas cependant, le contenu est délibérément dupliqué entre les domaines afin de manipuler le classement du site par les moteurs de recherche ou d'augmenter le trafic. Ce type de pratique trompeuse peut affecter négativement la navigation de l'internaute qui voit quasiment le même contenu se répéter dans un ensemble de résultats de recherche.

Google s'efforce d'indexer et d'afficher des pages contenant des informations distinctes. [...] Les mesures suivantes vous permettent de résoudre les problèmes de contenu en double de manière proactive et de vous assurer que les visiteurs accèdent au contenu que vous souhaitez leur présenter.
** Bloquez l'indexation des pages : plutôt que de laisser les algorithmes Google déterminer la "meilleure" version d'un document, vous pouvez nous indiquer votre version favorite. Par exemple, si vous ne souhaitez pas indexer les versions imprimables des articles de votre site, désactivez ces répertoires ou utilisez des expressions littérales dans votre fichier robots.txt.*
** Utilisez des redirections 301 : si vous avez restructuré votre site, utilisez des redirections 301 ("RedirectPermanent") dans votre fichier .htaccess pour rediriger efficacement les internautes, Googlebot et autres robots d'exploration. [...]*
** Soyez cohérent : assurez la cohérence dans vos liens internes. Par exemple, n'établissez pas de lien vers http://www.exemple.fr/page/, http://www.exemple.fr/page et http://www.exemple.fr/page/index.htm.*
** Utilisez des domaines de premier niveau : pour nous aider à présenter la version la plus appropriée d'un document, utilisez dans la mesure du possible des domaines de premier niveau pour gérer du contenu propre à un pays. Nous sommes plus enclins à penser que le site www.exemple.de contient du contenu destiné à l'Allemagne, que www.exemple.com/de ou de.exemple.com.*
** Diffusez du contenu avec prudence : si vous diffusez votre contenu sur d'autres sites, Google affichera systématiquement la version jugée la plus appropriée pour les internautes dans chaque recherche donnée, qui pourra être ou non celle que vous préférez. Cependant, il est utile de s'assurer que chaque site sur lequel votre contenu est diffusé inclut un lien renvoyant vers votre*

article original. Vous pouvez également demander à ceux qui utilisent votre contenu diffusé de bloquer la version sur leur site avec leur fichier robots.txt.

** Utilisez nos outils pour les webmasters afin de nous indiquer votre méthode d'indexation de site favorite : vous pouvez indiquer votre domaine favori à Google (par exemple, www.exemple.fr ou http://exemple.fr).*

** Limitez les répétitions : par exemple, au lieu d'inclure une longue mention de copyright au bas de chaque page, insérez un récapitulatif très bref, puis établissez un lien vers une page plus détaillée.*

** Évitez la publication de pages incomplètes : les internautes n'appréciant pas les pages "vides", évitez dans la mesure du possible les espaces réservés. [...]*

** Apprenez à maîtriser votre système de gestion de contenu : vérifiez que vous maîtrisez l'affichage du contenu de votre site Web. Les blogs, forums et systèmes associés affichent souvent le même contenu dans des formats divers. [...]*

** Limitez les contenus similaires : si de nombreuses pages de votre site sont similaires, développez chacune d'entre elles afin de les rendre uniques ou consolidez-les en une seule. [...]*

On notera également pour terminer un article, sur le blog pour webmasters de Google (intitulé "*Demystifying the "duplicate content penalty*"" et disponible à l'adresse : http://googlewebmastercentral.blogspot.com/2008/09/demystifying-duplicate-content-penalty.html) qui explique, avec raison, que le "Duplicate content" - par exemple le fait d'avoir un même contenu accessible par des adresses différentes - ne génère pas de "pénalités" au sens où on l'entend souvent sur le Web, même si cela peut être "pénalisant" (la nuance est importante...) pour votre visibilité sur les moteurs.

Si avec tout ça, vous vous laissez encore happer par les pièges du "Duplicate content" sur les moteurs de recherche, c'est à désespérer de tout... :-)

Olivier Andrieu
Abondance.com

Réagissez à cet article sur le blog des abonnés d'Abondance :
http://abonnes.abondance.com/blogpro/2008/10/duplicate-content-et-rfrencement-3me.html