Duplicate Content et Référencement (1ère partie)

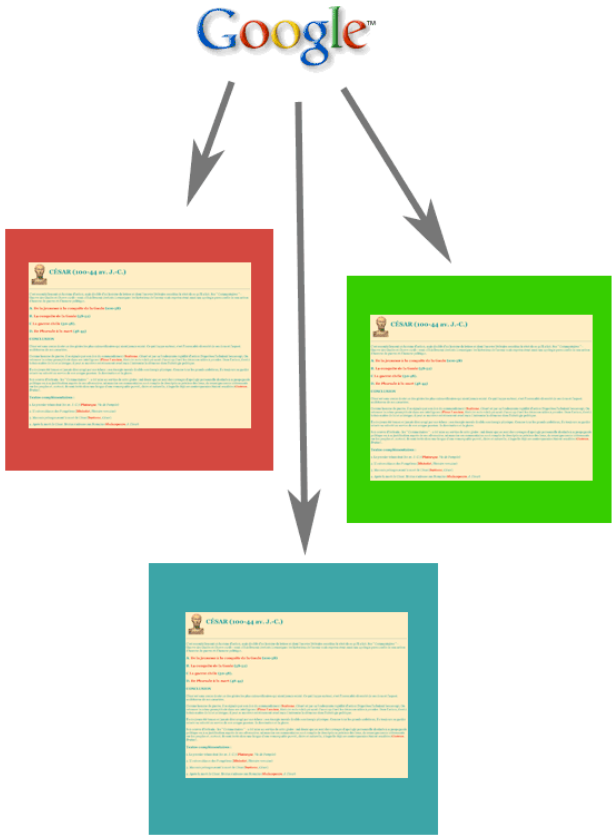
Retour au sommaire de la lettre

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 en commençant par une explication générale du phénomène de "duplicate content" et par l'exploration d'un premier volet sur le contenu "canonique" dupliqué par des sites partenaires ou "pirates"...

Introduction : le Duplicate Content, c'est quoi ?

Depuis de nombreuses années, on entend parler, en termes de référencement, de problèmes dûs au concept de "*Duplicate Content*". De quoi s'agit-il exactement ? Quels types de problèmes ce phénomène pose-t-il et quels sont les remèdes possibles ? Nous allons essayer, dans cet article, de répondre à toutes ces questions et de voir comment faire en sorte que le *duplicate content* ne soit plus qu'un mauvais souvenir pour vous si vous souffrez actuellement de ce type de problème...

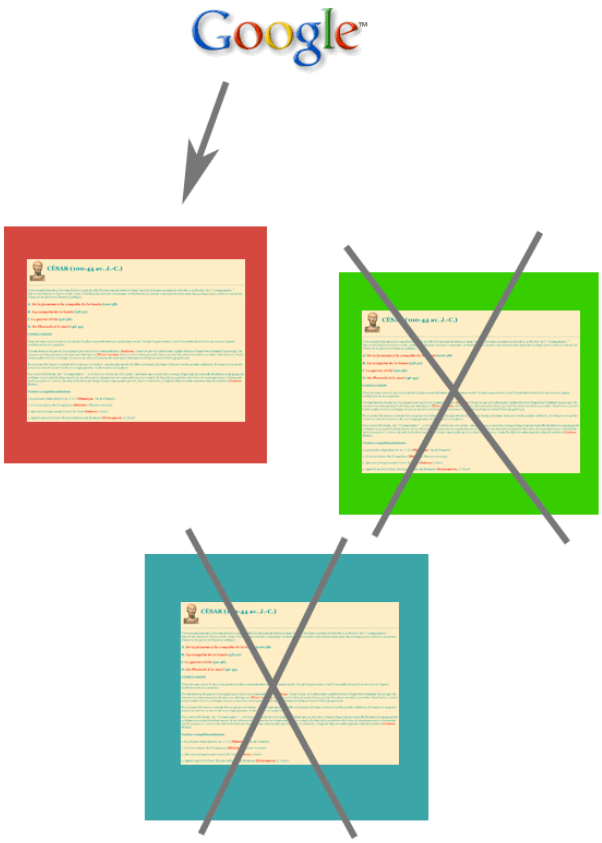
Mais tout d'abord, qu'est-ce que le *duplicate content* ? En fait, il s'agit d'une situation assez simple en soi : imaginons que Google (et les autres moteurs, bien sûr) ait, à un moment donné, indexé une ou plusieurs pages (sur le même site ou sur des sites différents) qui, selon lui, proposent un contenu identique ou, tout du moins, très proche, très similaire :



Google trouve sur le Web trois pages aux contenus éditoriaux très similaires, même si la mise en page / charte graphique est différente dans chacun des trois cas.

Il ne désire pas garder dans son index toutes ces pages trop proches les unes des autres et il décide donc de n'en garder qu'une seule. Ce sera celle qui, selon lui, propose le contenu "original", qui a

donc été "copié" par les autres documents. Il prend en compte ce contenu "canonique" et délaisse les autres pages :



Google choisit le contenu canonique pour ses pages de résultats

Notons bien qu'il ne supprime pas les pages contenant le contenu dupliqué mais qu'il les met dans un index secondaire (concept dont nous parlerons dans un prochain article) et y donne accès au moyen d'un lien en bas de page s'il détecte un phénomène de *duplicate content* :

> *Pour limiter les résultats aux pages les plus pertinentes (total : 5), Google a ignoré certaines pages à contenu similaire. Si vous le souhaitez, vous pouvez relancer la recherche en incluant les pages ignorées.*

Ce traitement, finalement assez logique, permet à Google de ne pas avoir trop de "doublons" dans son index et de fournir à ses utilisateurs des résultats plus pertinents. Cependant, il existe bon nombre de cas où cette notion de *duplicate content* (contenu dupliqué, donc), peut poser des problèmes aux éditeurs de sites web. C'est ce que nous allons étudier maintenant...

Problème numéro 1 : contenu dupliqué sur des sites partenaires

Le problème du *duplicate content* arrive très rapidement lorsqu'un même contenu se trouve sur des sites différents. Exemple-type : une dépêche AFP qui va se trouver sur le site de l'agence de presse, mais également sur de nombreux sites web "officiels" qui la reprennent.

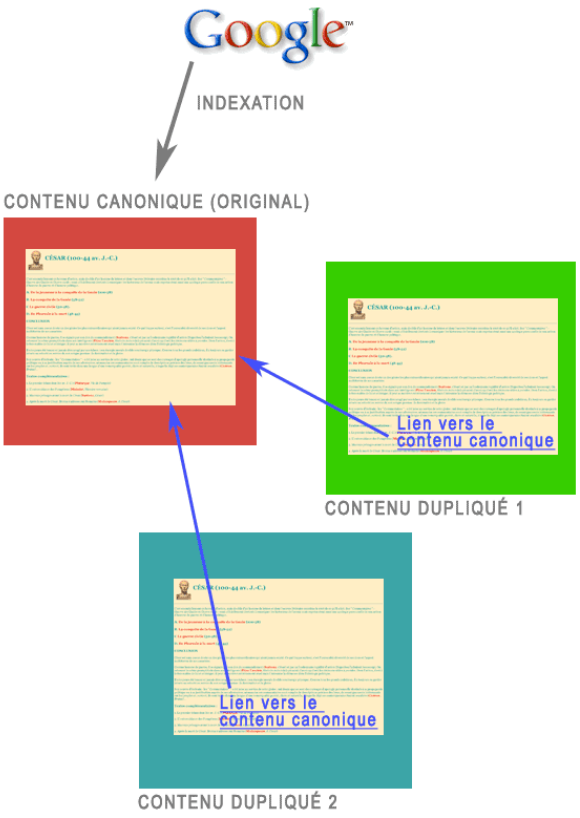
Autre exemple : un site web de contenu propose un article en ligne sur un sujet donné (mode, tourisme, sport, etc.) et cet article est repris par un site web partenaire, qui a signé un contrat pour avoir le droit de reprendre ce contenu.

Google sait très bien aujourd'hui, "extraire" le contenu réel, éditorial, d'une page web et laisser "de côté" toute la partie "navigation / charte graphique" du code HTML. Quand il aura fait ce travail sur les deux pages contenant l'article en question, il sera en possession de deux textes strictement

identiques. Dans ce cas, quelle version va-t-il prendre en compte ? La question n'est pas si simple et le choix risque d'être cornélien pour lui... Il faudra alors que vous l'aidiez dans son choix...

Exemple d'article repris à l'identique sur deux sites web différents

En effet, Google doit ici reconnaître quel est le contenu "canonique" (original) et quel est celui qui est dupliqué. Pour cela, il existe une façon de faire (recommandée d'ailleurs par Google) en demandant à vos partenaires - si vous êtes le propriétaire du contenu canonique - de mettre (si ce n'est déjà fait) un lien sur leur page vers votre page canonique. **Attention** : pas un lien vers la page d'accueil de votre site. Non ! Chaque page reprenant un de vos contenus doit "pointer" vers la page de votre site affichant le contenu original. Ceci est extrêmement important !



Ce lien vers le contenu canonique sera détecté par Google qui comprendra ainsi qu'un contenu est issu d'un autre et pourra "se faire son idée" sur la provenance originale du texte éditorial découvert. C'est également important pour Google News, outil sur lequel Google utilise fortement ses filtres de *duplicate content* car il est alors confronté quotidiennement à ce type de problème.

Sur cet outil, le lien "Trier par date et afficher les doubles" permet ainsi d'afficher les pages en *duplicate content*, triées et éliminées par défaut :

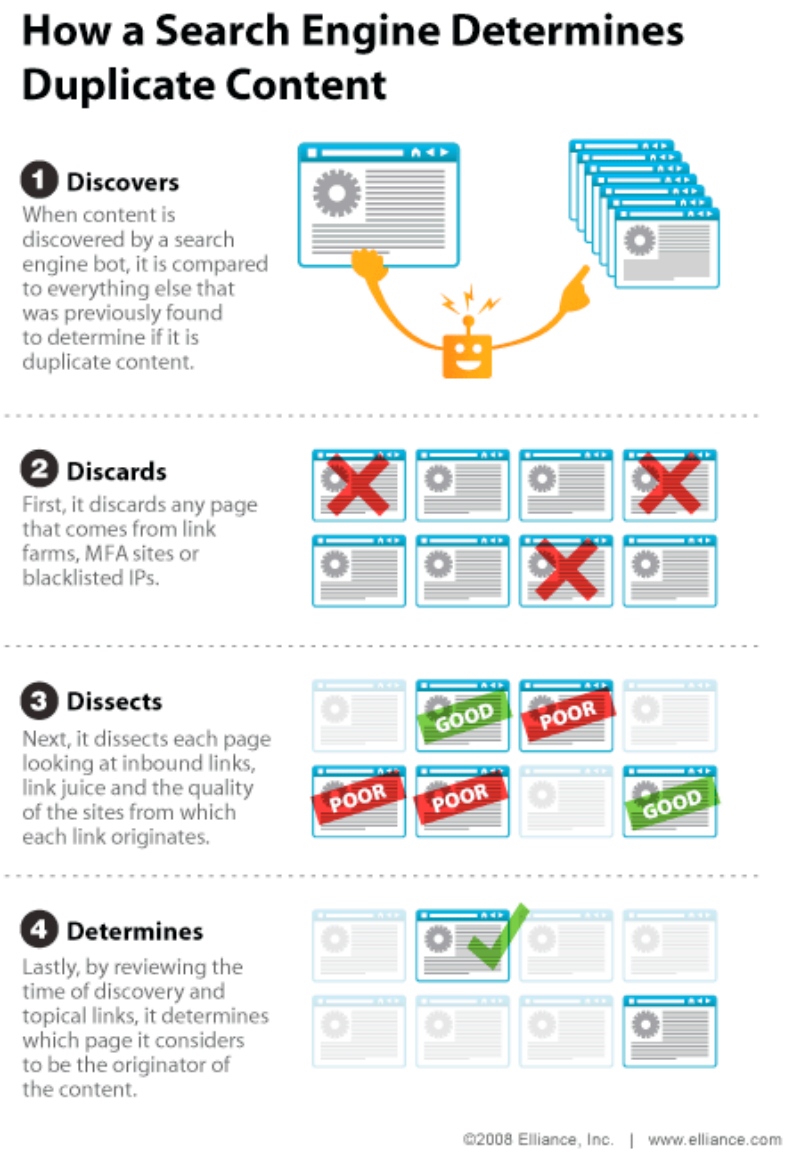
Notez que, même si vos pages ont 90% de leur code HTML (représentant toute la partie "navigation, etc.") identique, **seul le contenu réel - éditorial - sera pris en compte** par Google dans la détection du *duplicate content* (DC). Deux pages peuvent donc avoir un code HTML très différent mais un contenu éditorial identique. Cela ne leur évitera pas de tomber dans les filtres de "DC". Ne l'oubliez pas !!

On peut penser que le "TrustRank" ou indice de confiance des différents sites entrant en ligne de compte ici joue également son rôle, Google octroyant plus de confiance au site web disposant du TrustRank le plus élevé. De même, la date de publication (officielle dans le Sitemap spécifique à Google News) ou la date de découverte de l'article par le moteur a bien entendu son importance dans la somme de critères qui lui permettent de définir le contenu canonique. D'autres critères comme l'univers sémantiques des liens de la page, peuvent entrer en ligne de compte.

Autre solution pour éviter que votre contenu se trouve "dans la charette" au profit de celui d'un de vos partenaires (qui aura mieux optimisé ses pages que vous...) : prévoir, dès le début du partenariat, que ses pages ne doivent pas être référencées. Par exemple, par l'ajout d'une balise meta "Robots" avec valeur "noindex" ou *via* un fichier robots.txt adéquat. Le partenaire a ainsi le droit de reprendre votre contenu sur son site, mais il doit "barrer le passage" aux spiders des moteurs.

Bien entendu, cette solution est beaucoup plus facile à mettre en place avant négociation et signature du contrat qu'après... Cette situation est également valable pour le "rétrolien" vu auparavant. Obliger vos partenaires à insérer un lien vers votre contenu canonique doit être inclu dans le contrat que vous signerez avec lui. Il vous faudra ensuite bien vérifier que ce rétrolien est présent dans ses pages. De même, si vous mettez en ligne un blog ou, plus simplement, du contenu sous la forme d'articles, etc., proposez une "charte de reprise du contenu" dans laquelle vous indiquez l'obligation de ce lien vers la page canonique, et ce même si c'est le fil RSS (titre + résumé) qui est repris... On n'est jamais trop prudent...

Cette illustration, (© Elliance / Search Engine Land) explique bien comment le *duplicate content* est appliqué par le smoteurs de recherche :



Source : http://searchengineland.com/080513-080033.php

Problème numéro 2 : contenu dupliqué sur des sites "pirates"

Le problème explicité ci-dessus risque également de se poser de façon plus accrue si votre contenu est repris... par des sites qui ne sont pas vos partenaires... Dans ce cas, il sera encore plus énervant de voir un de ces contenus s'afficher en bonne position sur Google alors que le vôtre est passé dans les affres des filtres du *duplicate content*.

Bien entendu, il sera difficile de demander à un site web avec qui vous n'êtes pas "en affaires" de mettre en place un lien vers vous... S'il a envie de le faire, il le fera, s'il n'a pas envie, (et il y a de fortes chances pour que cela soit le cas)... Il ne le fera pas.

Quelle est la solution dans ce cas ? Premièrement, il vous faudra privilégier l'approche "amiable" en trouvant l'adresse e-mail du responsable du site "pirate" (sur ses pages ou via une fonction de Whois qui vous indiquera à qui appartient le nom de domaine) et lui indiquer que votre contenu est soumis à *copyright* et qu'il n'a pas le droit de le reprendre ainsi. Dans certains cas, l'éditeur du site distant sera de bonne foi et il stoppera ses activités illicites. Parions cependant que cette première approche ne donnera pas toujours des résultats positifs... Mais elle doit cependant être tentée.

Dans le cas où l'approche "en douceur" ne donne pas de résultats, il vous faudra alors durcir le ton, faire constater (avocat / huissier) la fraude, envoyer un courrier recommandé avec accusé de réception et demander à un avocat ou à votre service juridique qu'il brandisse la menace d'un procès si ce type de pratique ne cesse pas immédiatement. Si le site copieur est situé en France, le problème peut être réglé assez rapidement. S'il se trouve à l'étranger, les problèmes risquent de s'accumuler assez rapidement pour vous car la situation sera complexe...

Dans ce cas, il vous faudra certainement lacher prise et tenter de recevoir "le plus de rétroliens possible" de la part des éditeurs de sites web reprenant vos contenus. Les moteurs vont trouver et identifier ces liens et comprendre ainsi que c'est le vôtre qui est canonique, pas celui des pirates qui, eux, n'auront pas de rétrolien à proposer... C'est donc la page qui aura la plus forte popularité (PageRank), notamment par l'analyse des liens émanant des pages dupliquées, qui sera retenue par le moteur.

Une solution peut également être d'insérer des liens internes (vers d'autres pages de votre site) dans votre contenu rédactionnel (par exemple des tags sur certains mots). Si le site "pirate" reprend votre contenu, il reprendra (peut-être...) ces liens internes (qu'il faudra donc indiquer en absolu : www.votresite.com/tags/mot.html - et pas en relatif : ../tags/mot.html - dans votre code HTML) et cela sera une indication intéressante pour le moteur lorsqu'il analysera les textes à filtrer : un lien non pas vers votre page canonique mais au moins vers votre site, c'est toujours ça de pris.

Il faut également noter que Google, si l'on en croit les brevets qu'il a déposé à ce sujet - voir adresses en fin d'article -, arrive aujourd'hui non seulement à détecter les pages "globalement semblables" mais également à **identifier des parties de contenus ("snippets")** qui seraient repris dans d'autres pages... Le fait de reprendre un contenu et, par exemple, de modifier l'ordre de ses paragraphes, ne sera peut-être pas suffisant, selon les cas, pour éviter les filtres de Google...

Conclusion

Dans cet article, nous avons abordé la première forme de *duplicate content*, à savoir la recopie de contenu textuel "canonique" sur un autre site (il existe cependant d'autres formes de *duplicate content* que nous étudierons le mois prochain). Dès lors, il est important, lorsque vous mettez en place un projet de reprise de vos contenus par un site tiers, de suivre ces quelques conseils :

- Pensez à cette problématique au moment de la mise en place du partenariat pour éviter tout soucis par la suite : déréférencement des pages du site partenaire, mise en place d'un rétrolien, etc. Tout est important et doit être prévu à l'avance.

- Multipliez les liens externes vers votre contenu canonique.

- Créez éventuellement deux versions de votre contenu : l'un destiné à votre site, l'autre, moins riche, pour vos partenaires (par exemple pour un descriptif produit)...

- Affichez sur votre site une "charte de reprise du contenu" si vous autorisez ce type de pratique (notamment via des fils RSS).

- Insérez des liens internes (en adressage absolu) dans vos contenus.

- Faites une veille pour savoir qui reprend vos contenus, soit en saisissant comme requête sur un moteur de recherche une ou deux phrases de vos articles entre guillemets soit en utilisant des outils comme CopyScape (http://www.copyscape.com/), Compilatio (http://www.compilatio.net/duplicate-content.php), CopyTracker (http://copytracker.org/), Noplagia (http://code.google.com/p/noplagia/) ou Tineye (http://tineye.com/) pour les images.

Et pourquoi Google ne proposerait-il pas un tel service, notamment dans ses "Webmaster Tools" puisqu'il dispose de toutes les infos nécessaires à cela ?

Quelques liens sur la notion de "duplicate content"

Voici quelques liens qui nous ont semblé intéressants dans le cadre d'une stratégie de lutte contre le *duplicate content* (de nombreux articles émanent des blogs officiels de Google). Notez bien que certains d'entre eux abordent des sujets qui seront plus spécifiquement abordés dans les prochains articles de cette série :

Detecting duplicate and near-duplicate files (brevet de Google)
http://www.seoguide.org/google-patent-6658423.htm

Detecting query-specific duplicate documents (brevet de Google)
http://www.seoguide.org/google-patent-6615209.htm

Understanding SEO issues related to Duplicate Content (SEO Guide)
http://www.seoguide.org/seo201-duplicate-content.htm

Duplicate content (Google - Centre d'aide Administrateur Web)
http://www.google.com/support/webmasters/bin/answer.py?hlrm=fr&answer=66359

Deftly dealing with duplicate content (Google)
http://googlewebmastercentral.blogspot.com/2006/12/deftly-dealing-with-duplicate-content.html

Duplicate content due to scrapers (Google)
http://googlewebmastercentral.blogspot.com/2008/06/duplicate-content-due-to-scrapers.html

Ranking As The Original Source For Content You Syndicate (Vanessa Fox)
http://www.vanessafoxnude.com/2008/05/14/ranking-as-the-original-source-for-content-you-syndicate/

Duplicate content summit at SMX Advanced (Google)
http://googlewebmastercentral.blogspot.com/2007/06/duplicate-content-summit-at-smx.html

The Illustrated Guide to Duplicate Content in the Search Engines (SEOMoz)
http://www.seomoz.org/blog/the-illustrated-guide-to-duplicate-content-in-the-search-engines

Rewriting the Beginner's Guide Part IV Continued - Canonical and Duplicate Versions of Content (SEOMoz)
http://www.seomoz.org/blog/rewriting-the-beginners-guide-part-iv-continued-canonical-and-duplicate-versions-of-content

Faut-il avoir peur du Duplicate Content ? (RankSpiit)
http://www.rankspirit.com/duplicate-content.php

L'URL canonique, selon Google (Annuaire Info)
http://www.annuaire-info.com/google-url-canonique.html

Compléments de Matt Cutts sur le Duplicate Content (WordPress Tuto)
http://wordpress-tuto.fr/complements-de-matt-cutts-sur-le-duplicate-content-307

Olivier Andrieu
Abondance.com

Réagissez à cet article sur le blog des abonnés d'Abondance :
http://abonnes.abondance.com/blogpro/2008/07/duplicate-content-et-rfrencement-1re.html