Un moteur de recherche pour mon blog

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 7 novembre 2006. Dernière mise à jour le 2 avril 2009

https://www.bortzmeyer.org/recherche-blog.html

C'est évidemment très bien qu'il y aie un moteur de recherche sur ce blog. Pour cela, il y a deux solutions :

- Installer un logiciel libre comme Mnogosearch ou ht ://Dig sur le serveur. Cela m'obligerait à surveiller cet engin, l'indexation quotidienne, et d'une manière générale à abandonner la douce quiétude d'un site Web entièrement statique. Mon expérience avec ces outils, dès que le site Web a une taille non triviale, est... pas toujours rose. Par exemple, l'indexation nocturne plante souvent (disque plein, données corrompues...) nécessitant une intervention manuelle de l'ingénieur système.
- Faire héberger le moteur à l'extérieur. Plusieurs moteurs de recherche publics offrent cette possibilité. J'y perds du contrôle, je dois accepter leurs conditions d'utilisation et ce n'est pas du logiciel libre. mais ça marche nettement mieux et c'est surtout moins de travail.

J'ai finalement quand même choisi la première solution, un logiciel local <https://www.bortzmeyer.org/moteur-recherche.html> car la seconde solution, que j'explore ici, ne m'a pas laissé satisfait, notamment pour des problèmes de licence.

Pour chaque moteur possible, je vais regarder s'il offre cette possibilité, si les conditions d'utilisation ne sont pas trop léonines et si le service fonctionne bien.

Logiquement, on commence par Google qui offre depuis peu un service nommé "*Custom Search Engine*" <http://www.google.com/coop/cse/>. Sur mon blog, cela donne (Javascript obligatoire, par décision de Google) :

Text added to workaround a Firefox 3 stupid bug with empty XML elements. Should not be displayed.

La licence d'utilisation <http://www.google.com/coop/docs/cse/tos.html> est sévère. Par exemple "*You may not in any way frame or cache the Results*". Ou bien "*The Search Box shall conspicuously display a graphic (available at ...) that indicates that the Service is provided by Google.*". Ou encore "*Google may modify the Terms of Use at any time with or without notice [...] If You continue to use the Code and/or the*

Search Box on any Site, You will be deemed to have accepted the modifications.". Pas question de réarranger les résultats : "*You shall not, and shall not allow any third party to : (a) edit, modify, truncate, filter or change the order of the information contained in any Results (either individually or collectively), including, without limitation, by way of commingling Results with non-Google provided search results or advertising*". Et je dois accepter de faire de la pub à Google : "*You hereby grant to Google a nontransferable, nonexclusive license during the Term to use Your Brand Features to advertise that You are using the Service.*" Cependant, il faut noter que l'absence de certaines de ces clauses dans les autres licences ne signifie pas que tout soit permis : si on considère que la boîte de recherche ou que la présentation des résultats est une œuvre de l'auteur du moteur, alors, même en l'absence de clause interdisant toute modification, le droit d'auteur suffit à limiter mes possibilités d'adaptation. Par exemple, remplacer "Search" par "Recherche" dans les codes HTML proposés par les moteurs était limite.

À noter que le fait de proposer Google sur cette page avec les autres viole la licence "*You agree that, during the Term, Google will be the exclusive provider of Internet search services on the Site.*".

Une autre possibilité est d'utiliser Exalead qui avait également un tel service (il n'est plus documenté mais il marche encore ; ce manque de stabilité et de sérieux est d'ailleurs un problème fréquent avec certaines sociétés). Cela donne :

Mais Exalead passe bien plus rarement sur mon blog et de nombreuses pages ne sont pas indexées, problème que je n'ai pas vu avec Google. En outre, dans la page de résultat, les bonnes réponses sont mêlées aux publicités.

J'ai également essayé avec Yahoo, qui a aussi un service pour les auteurs de sites Web `<http://tools.search.yahoo.com/about/forsiteowners.html>`. Voici comment il se présente :

Je n'ai pas encore trouvé leur licence, je n'ai vu que la licence générale `<http://docs.yahoo.com/info/terms/>`.

Pour Yahoo, une autre solution pourrait être d'utiliser leur API, documentée ici `<http://developer.yahoo.com/>`.

Gigablast a aussi un service de recherche `<http://www.gigablast.com/sitesearch.html>`.

Pas de licence trouvée mais le code fourni ne comprend pas de logo, c'est déjà ça. On doit pouvoir arranger les résultats (ce que la licence de Google interdit formellement) puisqu'ils peuvent être fournis en XML.

C'est essentiellement à cause des contraintes de licence que j'ai finalement renoncé à cette solution au profit d'un moteur de recherche local `<https://www.bortzmeyer.org/moteur-recherche.html>`.