Outils de recherche et Open Source

Retour au sommaire de la lettre

De plus en plus de technologies Open Source sont utilisées pour développer des outils de recherche (annuaires et moteurs). Nous vous proposons ici un état des lieux de l'utilisation des technologies "libres" par les moteurs, avec une présentation des technologies les plus répandues et des outils qui les ont adoptées.

L'Open Source est apparu en 1984 avec le lancement par Richard Stallman du projet GNU (acronyme de "GNU's Not UNIX) qui visait à créer un système d'exploitation "libre" et gratuit. De très nombreuses communautés de développeurs "bénévoles" ont ensuite essaimé dans le monde, évoluant pour certaines en communautés de contributeurs constituées en société.

Les applications en Open Source renvoient à des solutions dont l'accès au code source est autorisé par leurs auteurs afin de faciliter le développement de logiciels dérivés. Ces solutions sont le plus souvent développées selon un mode de travail collaboratif, une équipe "pilote" étant chargée de superviser le projet et la qualité des développements. Cette organisation permet de produire des logiciels dont le coût est limité et la technicité élevée.

Chacun est libre d'utiliser une solution en Open Source et de partager avec la communauté les améliorations qu'il apporte au produit. Le support technique est en général assuré par les utilisateurs eux-mêmes ou par les développeurs (par le biais de listes de discussion notamment).

Annuaire Open Directory Project (DMOZ)

L'Open Directory Project ou DMOZ (http://dmoz.org/) est le plus important annuaire de sites Web édités par des êtres humains bénévoles. Cet annuaire a été créé en 1998 dans l'esprit du mouvement Open Source, le nom DMOZ étant un raccourci pour Directory Mozilla.
La consultation de l'annuaire ou l'utilisation de son répertoire par les autres outils de recherche sont entièrement gratuites. L'ODP fournit son contenu au format RDF (Resources Description Framework), un format qui est une variante du XML. Les utilisateurs doivent ensuite utiliser leur propre script pour pouvoir exploiter ces données.
Le contenu de l'annuaire est "ouvert". On utilise d'ailleurs l'expression "Open Content" pour définir l'ODP et non "Open Source" car si le contenu est "ouvert", la base de données reste un système propriétaire.

A noter l'existence d'un autre annuaire appelé Zeal (http://www.zeal.com) qui s'inspire du principe de l'Open Directory puisqu'il est fondé sur le bénévolat des éditeurs. Racheté en 2000 par Looksmart, cet annuaire n'est toutefois pas un outil "Open Content" puisque l'utilisation des données par des tiers n'est pas autorisée.

Moteurs de recherche "complets" en Open Source

Panorama des solutions disponibles

Solution (URL)	Pays - Année	Auteur(s)	Langage
ASPseek (http://www.aspseek.org/)	US	SWsoft	C++
DataParkSearch (http://www.dataparksearch.org/)	US - 2003	Maxim Zakharov	C
Egothor (http://www.egothor.org/)	Rép. Tchèque - 1997	Leo Galambos	Java
Glimpse / WebGlimpse (http://webglimpse.net/)	US - 1997	Internet WorkShop	C / Perl
ht://Dig (http://www.htdig.org/)	US - 1995	San Diego State University	C++

Isearch (http://www.etymon.com/tr.html)	US - 1994	Nassib Nassar	C++
MnoGoSearch (http://mnogosearch.org/)	Russie - 1998	Lavtech	C
Namazu (http://www.namazu.org/)	Japon - 2000	Namazu Project	C / Perl
Nutch (http://www.nutch.org/)	US - 2003	Doug Cutting	Java
Perlfect Search (http://perlfect.com/freescripts/search/)	UK - 1997	Perlfect Solutions	Perl
PHPDig (http://www.phpdig.net/)	US - 2001	Jelsoft Enterprises	PHP
phpMySearch (http://web4.hm/phpmysearch/)	Allemagne	Webagentur web4.hm	PHP
Zebra (http://indexdata.dk/zebra/)	Danemark 1994	Index Data	XML

Nutch

Lancé en 2003 aux Etats-Unis par Doug Cutting, l'un des anciens architectes du moteur Excite, Nutch est un moteur de recherche en Open Source. Ce moteur, qui est l'un des plus connus actuellement, s'est fixé pour objectif de contrecarrer deux "défauts" du marché des outils de recherche "commerciaux", à savoir la domination sur le marché de trois acteurs (MSN, Yahoo et Google) et le manque de transparence des critères de classification des résultats.

A noter en particulier la possibilité que peut offrir l'outil d'indiquer aux utilisateurs les critères utilisés pour le positionnement / *ranking* des sites dans la liste de résultats, en cliquant sur les onglets "Explication" ou "Explain". Cette fonctionnalité est notamment offerte par les moteurs de recherche Mozdex et Objects Search qui sont basés sur la technologie Nutch.

Pour assurer son développement, les responsables du projet Nutch recherchent des sponsors. Ils bénéficient déjà du soutien financier de Yahoo Search Marketing (anciennement Overture).

Moteurs de recherche utilisant la technologie Nutch

Mozdex	*Creative Commons*	*Objects Search*
(http://www.mozdex.com/)	*(http://search.creativecommons.org/)*	*(http://www.objectssearch.com/)*
Lancé fin 2004, Mozdex a été construit par la société américaine Small Productions sur la base de Nutch et il utilise également la technologie Lucene. Son index a été initialisé avec DMOZ, annuaire collaboratif de plus de 4 millions de pages. Enfin, le moteur dispose de son propre système de liens contextuels : MozAds.	Lancé en 2004, Creativecommons.org est un moteur sur les contenus Open Source (disposant de licences Creative Commons). L'outil, basé sur la technologie Nutch, permet de limiter sa recherche à un format de document ou d'exclure les supports dont les licences sont incompatibles avec une utilisation commerciale.	Lancé en 2004 et basé sur la technologie Nutch, Objects Search est un moteur de recherche Web offrant la possibilité d'effectuer des recherches sur les news et les weblogs. Ce moteur offre en outre une version en cache des résultats, ainsi que la possibilité de rechercher par "clusters" (comme Mooter et Kartoo).

Lucene

Egalement créée par Doug Cutting (voir plus haut), Lucene (http://lucene.apache.org/) est une "bibliothèque" de moteurs de recherche écrite en Java. Il ne s'agit pas d'une application mais d'une technologie pouvant être intégrée à d'autres applications. Cette technologie est très répandue (voir la liste complète des utilisateurs sur http://wiki.apache.org/jakarta-lucene/PoweredBy). Elle est notamment utilisée par Arisem (groupe Thales) pour sa solution entreprise de recherche "Kaliwatch

Professional 2.0", une application qui permet d'indexer 225 formats différents de documents (http://www.01net.com/article/251419.html).

ASPseek

Développé par la société américaine SWsoft, ASPseek est une solution Open Source (sous licence GPL) comprenant un moteur d'indexation et une interface de recherche. La recherche avancée permet de limiter sa requête à une période donnée, à un site Web ou à un ensemble de sites. Enfin, les résultats peuvent être triés par pertinence ou par date et le moteur propose une fonction de cache.

Moteurs de recherche utilisant la technologie ASPseek

DeepIndex (*http://www.deepindex.com/*)	*LaBanquise.Org* (*http://www.labanquise.org/*)
Lancé en juin 2002 par la société franco-tunisienne VirtuelPub, Deepindex est un moteur de recherche francophone initialement basé sur l'application Open Source ASPseek.	"La Banquise" est un moteur de recherche sur les sites Web francophones ayant pour sujet le "libre" au sens large. 235 sites sont pour l'instant recensés et 627 903 pages sont indexées. Il est possible de soumettre un site en ligne.

mnoGoSearch

mnoGoSearch était connu sous le nom de UDMSearch jusqu'en Octobre 2000, date à laquelle la société russe Lavtech a racheté UDMSearch et rebaptisé la solution mnoGoSearch. "Mnogo" signifie "beaucoup" en russe.
Lancé en 1998. L'outil est compatible avec les protocoles HTTP, HTTPS, FTP et NTTP, ainsi qu'avec les fichiers locaux.

DataParkSearch

DataparkSearch (http://www.dataparksearch.org/) est une solution Open Source basée sur la version 3.2.16 CVS de monoGoSearch. Sous licence GNU, cette solution permet d'effectuer des recherches sur un site, un groupe de sites ou un Intranet. La solution comprend :
- un spider et indexeur (indexer) collectant les informations
- une interface de recherche dans la base de données

Moteur de recherche utilisant la technologie DataParkSearch

NewsLookUp (http://www.newslookup.com/)
NewsLookUp est un moteur de recherche de news permettant de sélectionner des sources d'information par thèmes (radio, télévision…) et d'effectuer des recherches ciblées par pays ou continent. Il est possible de rechercher uniquement dans le titre, le corps du texte ou les balises META et d'effectuer un tri des résultats par pertinence, par date ou par source d'information.

PHPDig

PhpDig (http://www.phpdig.net/) est un moteur de recherche plein texte basé sur les technologies PHP et nécessitant une base de données. Cet outil permet l'indexation et la recherche de fichiers Web, Office et PDF sur un ou plusieurs sites.

Pour le lancement de son moteur de recherche Numika (http://www.numika.com/) fin 2004, la société française Addonis a utilisé "une base légèrement modifiée de PHPDIG qui s'est avérée très rapidement très insuffisante techniquement". Cette société a "donc décidé peu après de créer un nouveau noyau logiciel utilisant des méthodes simples mais plus efficaces. [Elle utilise désormais] un méta-moteur sur chaque nouvelle requête dont les résultats seront indexés dans [ses] propres bases de données (ceci dans le but de limiter [ses] besoins machines pour constituer [ses] bases)" (Denis Labarre – Addonis).

Egothor

Egothor est un moteur de recherche "modulaire" plein texte en Java compatible avec plusieurs plateformes proposé par des développeurs de République Tchèque. L'outil (qui est compatible avec plusieurs formats de fichiers Web, PDF et Office) est pour l'instant surtout utilisé pour des projets de "petite" échelle dans des bibliothèques. Il comprend un crawler, capable d'indexer 50 pages à la seconde et compatible avec les fichiers "robots.txt" (CAPEK : http://www.egothor.org/api/robot/).

Métamoteurs

JXTA Search

Le moteur de recherche distribué JXTA (http://www.jxta.org/) a vu le jour en juin 2000 lorsque deux employés de la société américaine Infrasearch (Gene Kan et Yaroslav Faybishenko) ont développé un premier moteur de recherche connecté à différents serveurs Web reliés entre eux via un protocole "peer to peer".

Infrasearch a ensuite été racheté par Sun Microsystems en 2001 et JXTA Search est désormais distribué gratuitement en Open Source.

JXTA Search distribue les requêtes vers les serveurs Web du réseau qui sont les plus à même d'y répondre. Cet outil n'utilise pas de spider et ne dispose pas d'index pour répondre aux questions d'utilisateurs. Il utilise en revanche un protocole XML appelé QRP (Query Routing Protocol) définissant la manière dont les fournisseurs de contenus doivent transmettre et répondre aux requêtes sur le réseau JXTA.

Helios

Créé par Antonio Gulli (récemment recruté comme Directeur technologique par le bureau italien de Pise de Ask Jeeves) et Alessio Signorini (étudiant à l'Université d'Iowa), Helios est un nouveau métamoteur en Open Source (voir http://www.cs.uiowa.edu/%7Easignori/helios/). Ce métamoteur, qui est actuellement utilisé par plusieurs groupes de recherches universitaires, est compatible avec 15 moteurs (A9, About, AllTheWeb, Altavista, AOL Search, eSpotting, FindWhat, Gigablast, Google, LookSmart, Mozdex, MSN, Overture, Ask/Teoma et and Yahoo).
Une interface Web permet aux utilisateurs d'effectuer une recherche sur une sélection de moteurs. La requête est ensuite interprétée et réécrite en XML par l'outil pour être transmise dans un format compatible avec chaque moteur.

Magellan Metasearch

Lancé dernièrement sous licence Open Source (GPL), la plateforme de veille collaborative Magellan Metasearch (http://sourceforge.net/projects/magellan2) est destinée à être installée sur un Intranet pour automatiser la veille sur les moteurs de recherche. Cet outil a été conçu afin que chacun puisse personnaliser les aspects de son fonctionnement et lui ajouter des fonctionnalités (filtres d'analyse...). Ses caractéristiques sont les suivantes :
- Interrogation de dix moteurs de recherche dont Google, Teoma, Feedster et Gigablast...
- Exhaustivité des résultats traités.
- Dédoublonnage des résultats de recherche.
- Modules d'analyse et de filtrage des résultats.
- Planification des requêtes dans le temps afin de détecter les nouveaux résultats.
- Exportation des résultats sous Excel ou en HTML pour interfacer un crawler externe.
- Diffusion en temps réel d'alertes par e-mail et par RSS.

Une première démo de l'outil vient d'être mise en ligne sur le site de l'association des anciens de l'Ecole de Guerre Economique (http://www.associationege.com/cgi-bin/search.pl).

Composants technologiques de moteurs en Open Source

Panorama des solutions disponibles

Solution (URL)	Pays - Année	Auteur(s)	Langage
Spiders			
Arachnid (http://arachnid.sourceforge.net/)	2002	Robert Platt	Java
Combine Harvester (http://www.lub.lu.se/combine/)	Suède - 1998	Lund University Libraries NetLab	Perl
Grub (http://www.grub.org/)	US -	Grub (Looksmart)	-
Heritrix (http://crawler.archive.org/)	US - 2004	Internet Archive	Java
Jspider (http://sourceforge.net/projects/j-spider/)	2002	Günther Van Roey	Java
Larbin (http://larbin.sourceforge.net/)	France	Sébastien Ailleret	C++
WebBase (http://www.nongnu.org/webbase/)	France - 2000	Loïc Dachary	C
WebSPHINK (http://www-2.cs.cmu.edu/~rcm/websphinx/)	US	Rob Miller	Java
Moteurs d'indexation			
MG (Managing Gigabytes) (http://www.mds.rmit.edu.au/mg/)	Australie	Ian H. Witten...	ANSII C
Swish-e (http://swish-e.org/)	US - 1995	Kevin Hughes	C / Perl
Swish++ (http://homepage.mac.com/pauljlucas/software/swish/)	US	Paul J. Lucas	C++
Moteur d'interrogation			
Mifluz (http://www.gnu.org/software/mifluz/)	France	Loïc Dachary	C++
XML Query Engine (http://sourceforge.net/projects/xqengine/)	2003	-	Java

Crawlers / Spiders

Grub

Racheté par Looksmart en 2003, Grub est un spider gratuit (sous licence GPL) distribué sur les machines d'internautes "consentants". Ce crawler, qui nécessite le téléchargement par chaque utilisateur volontaire d'une application client (économiseur d'écran), s'appuie sur les grilles de calcul pour indexer le Web. Son fonctionnement est "identique" à celui d'un moteur "classique", si ce n'est que les robots sont lancés depuis les PC des utilisateurs.

Grub offre plusieurs avantages par rapport à un spider classique. Il permet d'économiser la bande passante utilisée par les spiders classiques des moteurs. Il offre en outre une méthode de crawling "local" permettant aux Webmasters de crawler uniquement leurs propres sites et d'envoyer à Grub quotidiennement les données actualisées dans un format compressé.

Le moteur Wisenut (http://www.wisenut.com/) de Looksmart utilise le spider distribué Grub pour améliorer ses capacités de crawl.

Heritrix

Heritrix (http://crawler.archive.org/index.html) est un crawler Open Source développé par l'Internet Archive pour sa collecte du Web. Il respecte les fichiers

"robots.txt" et analyse les balises META de chaque page.
Brewster Kahle, fondateur de l'outil d'archivage du Web Internet Archive (http://www.archive.org/), est à l'origine de l'un des premier spiders du Web en 1991 : WAIS (http://www.infomotions.com/musings/opensource-indexers/).

WebBase
Anciennement utilisé par Ecila (moteur qui a aujourd'hui disparu), Webbase (http://www.nongnu.org/webbase/) est un spider utilisant une base de données MySQL. Il ne comprend pas d'outil d'indexation.

Moteurs d'indexation

Swish-E
Développé en C et en Perl, Swish-E (acronyme de "Simple Web Indexing System for Humans-Enhanced" : http://swish-e.org/) est un moteur d'indexation de pages Web en HTML ou en XML, de fichiers textuels ou de bases de données. Une première version, dénommée Swish, a été lancée en 1995 par Kevin Hughes et l'outil a ensuite été repris et amélioré par l'université de Berkeley en 1996 qui l'a renommé Swish-E et distribué sous licence GPL.

Le système privilégie les balises META ou les titres des documents pour son indexation.
Il est notamment utilisé par la fondation Apache (http://search.apache.org/) et par l'université Laval au Canada (http://www.ulaval.ca/Al/cherche/swishEAide.html).

Swish++
Swish++ est une version en C++ de Swish-E, développée par Paul J. Lucas en 1998 (http://homepage.mac.com/pauljlucas/software/swish/).

Terrier
Terrier (TERa REtrIEver : http://ir.dcs.gla.ac.uk/terrier/about.html) est une plate-forme d'indexation et de récupération données développée par des chercheurs de l'université écossaise de Glasgow en Java et en Perl. Elle utilise un spider appelé Labrador.
Ce projet a bénéficié d'un financement de 30 mois de l'institution britannique EPSRC (Engineering and Physical Sciences Research Council).

Système Carrot2 de ranking des résultats

A noter l'existence de Carrot2 (http://sourceforge.net/projects/carrot2/), un projet Open Source mené par des chercheurs polonais ayant pour ambition de mettre en place un système de "clustering" de l'information sur les moteurs de recherche. Le "clustering" consiste à regrouper en dossiers thématiques créés instantanément les résultats de recherche d'un moteur, Vivisimo et Exalead étant les moteurs les plus connus proposant ce type de restitution des résultats.
Le projet Carrot2 comprend à la fois des éléments de clustering et un métamoteur.
Une version "commerciale" de cette technologie est également offerte : http://www.carrot-search.com/.
Pour plus d'informations sur ce projet, voir l'entretien avec David Weiss dans cette lettre.

Système GATE d'extraction de données

GATE (General Architecture for Text Engineering : http://gate.ac.uk/ie/) est un projet d'extraction de données mené par une quinzaine de chercheurs spécialistes du traitement en langage naturel à l'université de Sheffield au Royaume-Uni. Ce projet est coordonné par Hamish Cunningham.

Cet outil d'extraction d'information (*information extraction*) permet d'analyser des données textuelles pour en extraire des informations importantes sur des sujets prédéfinis.
Il peut être utilisé, par exemple, pour identifier des faits "saillants" (et non des documents) sur des événements, des entités ou des relations. Ces faits marquants sont ensuite en général sauvegardés

dans une base de données qui peut être exploitée pour identifier des tendances ou pour fournir un résumé des informations en langage naturel.

GATE est distribué avec un module d'extraction de données appelé ANNIE (acronyme de "A Nearly-New IE system") utilisable avec toutes sortes de formats de documents, qu'il s'agisse d'emails ou de fils d'actualités...

Initiatives autour de l'Open Source

ALVIS Superpeer Semantic Search Engine

ALVIS (http://www.alvis.info/) est un projet européen qui vise à développer un nouveau moteur distribué sémantique en Open Source. Cet outil "peer to peer" ne sera pas positionné en concurrence avec les grands moteurs mais il sera destiné à des cibles très spécifiques telles que les "minorités" linguistiques ou certains groupes d'universitaires.

Ce projet d'une durée de trois ans a débuté en janvier 2004 et il est coordonné par l'université d'Helsinki. Dix autres partenaires collaborent au projet, parmi lesquels Exalead, l'Institut National de la Recherche Agronomique, l'Ecole Polytechnique Fédérale de Lausanne et l'Université Paris-Nord.

Frutch

Jérôme Charron (cf. entretien dans la lettre R&R de mai 2005), initiateur notamment de la liste Motrech sur les moteurs de recherche http://motrech.free.fr/), a lancé le projet Frutch (http://frutch.free.fr/) en février 2005. Cette initiative ambitionne de proposer une alternative aux moteurs de recherche commerciaux tels que Google, Yahoo et MSN en créant des liens entre les francophones s'intéressant au projet Nutch. Un groupe de travail a notamment été mis en place pour participer à l'effort de développement collaboratif de Nutch. Il est composé à la fois de développeurs et de professionnels ayant travaillé ou travaillant toujours pour des moteurs commerciaux.

Frutch souhaite à terme devenir un représentant francophone de Nutch et peut être développer un nouveau moteur de recherche alternatif.

Sourceforge

Lancé en 1999, Sourceforge (http://sourceforge.net/) propose un hébergement gratuit pour différents projets et programmes en Open Source qu'ils soient développées par des développeurs "lambdas" ou par des groupes comme IBM ou Google. Le site a annoncé en mai 2005 héberger actuellement plus de 100 000 projets hébergés sur ses serveurs et il revendique plus de 950 000 abonnés.

SourceForge.net appartient à la société américaine OSTG (éditeur de différents sites Web d'actualités technologiques, parmi lesquels Slashdot.org et Thinkgeek.com) qui est elle-même une filiale à part entière de la SSII américaine VA Software.

Google Code

Depuis mars 2005, Google propose également un portail dédié Open Source (http://code.google.com/) menant à différents projets en cours. Les solutions API de Google autour de son moteur, de ses liens sponsorisés et de son produit de recherche sur PC sont tout d'abord proposées. Différents projets Open Source "externes" sont également présentés, parmi lesquels PyGoogle, Core Dumper, Sparse Hashtable, Goopy/Functional et Perftools (http://code.google.com/projects.html). Certains développeurs

regrettent toutefois qu'il n'y ait pas réellement d'incitation du moteur à la participation dans des projets.

Opensearch

A9, moteur de recherche du site Amazon, a annoncé dernièrement l'initiative "OpenSearch" (http://opensearch.a9.com/) avec pour objectif de proposer des technologies basées sur des standards ouverts, au travers desquelles tout fournisseur de contenu pourrait publier des données dans un format propre à la syndication au format XML.

Perspectives des technologies en Open Source pour les moteurs de recherche

Les moteurs de recherche Open Source offrent plusieurs avantages :
- L'Open Source permet de mutualiser les compétences en matière de développement, notamment pour la création et la modification des algorithmes.
- Les règles d'indexation et de classement de l'information sont en général transparentes.

A l'instar de Google avec son projet "Google Code", d'autres acteurs "majeurs" du search réfléchiraient actuellement à des projets Open Source, y voyant un moyen de stimuler l'innovation pour enrichir leurs produits. Ask Jeeves est notamment actuellement en discussion avec la fondation Mozilla en vue de mettre sur le "marché Open Source" son produit de "desktop search" (http://actu.abondance.com/2005-07/askjeeves-mozilla.php). Dernièrement, eBay vient également d'annoncer un projet "eBay Community Codebase" (https://www.codebase.ebay.com/) avec la publication d'une partie des codes sources de ses applications de recherche pour stimuler la création par des développeurs externes d'applications compatibles avec son service.

Points de vue de deux acteurs de la recherche :
"Quels sont selon vous les avantages et les inconvénients de l'Open Source pour les moteurs de recherche ?"

Denis Labarre – Addonis (http://www.numika.com/)
Travaillant beaucoup avec de l'Open Source, je suis assez mal placé pour ne pas y adhérer. Le monde de l'Open Source est une petite révolution qui fait son bout de chemin depuis quelques années dans le monde logiciel. Mais à terme, toutes les possibilités que l'Open Source nous offre ne représentent t'elles pas une perspective d'aseptisation d'Internet ? Aujourd'hui, Google et son secret bien gardé fait l'unanimité dans le monde des recherches Internet. Imaginez vous ayant la possibilité d'avoir quatre cents moteurs variantes plus ou moins importantes de Google ! Certes le monopole des recherches serait cassé mais quel en serait l'intérêt pour les internautes ? La richesse d'Internet aujourd'hui est la diversité de ce que l'on peut y trouver. Nous manquons encore de visibilité sur les conséquences économiques que l'Open Source pourrait engendrer. Peut être est-ce là la "contre mesure" des anti-monopoles pouvant amener à un équilibre sur le long terme ??? Ou seulement le reflet d'un désir inconscient de tous de vivre dans un monde monochrome. L'inconvénient majeur à mon goût de l'Open Source est peut être la monotonie que cela pourrait instaurer sur Internet, mais après tout, nous achetons tous des voitures de couleur grise, noire ou blanche alors que de somptueuses nuances de couleurs existent ...

Gilbert Wayenborgh – DeepIndex (http://www.deepindex.fr/)
L'Open Source nous a permis de démarrer un moteur de recherche. Les avantages sont donc importants et permettent de démarrer un tel outil rapidement sans avoir des coûts exorbitants en terme de développement. Néanmoins l'Open Source dispose aussi d'inconvénients majeurs en terme d'évolution. Par exemple, avec Aspseek, outil que nous avions à l'origine de Deepindex, il n'y a plus d'évolution possible aujourd'hui. Les développeurs ont virtuellement disparus ...
Cependant, cela n'a pas empêché Deepindex d'évoluer et une mise à jour importante aura lieu à la rentrée avec en *background* l'appui d'un logiciel propriétaire.

Nous nous sommes concentrés ici sur le volet "outil de recherche" mais les référenceurs disposent également de plusieurs outils Open Source pouvant les aider. Par exemple, l'application RobotStats (http://www.robotstats.com/), écrite en PHP, permet d'analyser les visites de spiders sur votre site, chaque passage d'un crawler étant enregistré dans une base de données MySQL. Un outil Open Source de test de positionnement de site Web sur les moteurs de recherche existe également : PhpSera (http://phpsera.sourceforge.net/). Enfin, un autre projet baptisé "Open Rank" (http://openrank.org/) vient de voir le jour à l'initiative de référenceurs américains. Il a pour objectif de créer à terme un index (indépendant des grands moteurs de recherche) qui servira à analyser la structure du Web et à étudier les différentes méthodes de ranking des résultats.

Quelques liens...

Outils de recherche d'applications en Open Source pour les moteurs de recherche

Sourceforge (Annuaire de projets de moteurs de recherche en Open Source)
http://sourceforge.net/softwaremap/trove_list.php?form_cat=93

Freshmeat (Annuaire d'applications, principalement Open Source, avec descriptifs)
http://freshmeat.net/

Koders (Moteur de recherche spécifique de codes Open Source dans 16 langages de programmation différents)
http://www.koders.com/)

Sites sur l'Open Source

LogicielLibre.Net (Site communautaire sur l'Open Source)
http://www.logiciellibre.net/

Openindex (Site communautaire permettant portant sur les moteurs de recherche conjugués en Open Source)
http://www.openindex.org/