Le Fonds Européen de Développement Régional et la Région wallonne investissent dans votre avenir

[ Rencontres Mondiales du Logiciel Libre 2011 – Lundi 11 juillet 2011 ]

# Développement d'un moteur de recherche avec Zend Search

Auteur : Dr Ir Robert Viseur

www.cetic.be

Your connection to ICT research

CETIC a.s.b.l. • Bâtiment Eole • Rue des Frères Wright, 29/3 • B-6041 Charleroi (Belgique) • T. +32 71 490 700 • F. +32 71 490 799 • info@cetic.be

# Qui suis-je ?

- Robert Viseur
  - Ingénieur Civil, Mastère en Management de l'Innovation, Docteur en Sciences Appliquées.
  - Spécialisé dans les questions relatives à l'économie des logiciels libres et aux pratiques de co-création, ainsi que dans les technologies de recherche et de traitement de l'information (outils d'indexation, API,...).
  - Responsable de l'annuaire de prestataires *logiciellibre.com*.
  - Assistant à la Faculté Polytechnique de l'Université de Mons (www.umons.ac.be).
  - Conseiller technologique au CETIC (www.cetic.be).



CETIC a.s.b.l. • Bâtiment Eole • Rue des Frères Wright, 29/3 • B-6041 Charleroi (Belgique) • T. +32 71 490 700 • F. +32 71 490 799 • info@cetic.be

# Qu'est-ce que le CETIC ?

- Centre d'Excellence en Technologies de l'Information et de la Communication basé à Charleroi (Belgique).
- Trois départements (et types de services) :
  - Software & System Engineering : qualité logicielle (fiabilité, sécurité, respect des normes internationales, processus,...).
  - Software & Services Technologies : architectures orientées services et sémantique.
  - Embedded & Communication Systems : prototypage de systèmes embarqués communicants et nouvelles technologies électroniques.



CETIC a.s.b.l. • Bâtiment Eole • Rue des Frères Wright, 29/3 • B-6041 Charleroi (Belgique) • T. +32 71 490 700 • F. +32 71 490 799 • info@cetic.be

# Introduction



CETIC a.s.b.l. • Bâtiment Eole • Rue des Frères Wright, 29/3 • B-6041 Charleroi (Belgique) • T. +32 71 490 700 • F. +32 71 490 799 • info@cetic.be

# De quoi allons-nous parler ?

- Sujet : création d'un moteur de recherche utilisant les technologies « wget » et « Zend Search » (version PHP de Lucene).

- Plan :
  - Présentation des outils (wget, Lucene, Zend Search).
  - Mise en œuvre (encodage UTF-8 sous PHP, Zend Search en pratique,...).
  - Quelques exemples.
  - Conclusion.



CETIC a.s.b.l. • Bâtiment Eole • Rue des Frères Wright, 29/3 • B-6041 Charleroi (Belgique) • T. +32 71 490 700 • F. +32 71 490 799 • info@cetic.be

# Outils



CETIC a.s.b.l. • Bâtiment Eole • Rue des Frères Wright, 29/3 • B-6041 Charleroi (Belgique) • T. +32 71 490 700 • F. +32 71 490 799 • info@cetic.be

# Wget (1)

- Utilitaire GNU en ligne de commande, compatible Linux ou Cygwin, permettant de récupérer des fichiers en utilisant HTTP, HTTPS et FTP.
- Commande de base :
  - `wget www.cetic.be`
    - Stocke localement la page située à l'adresse « www.cetic.be ».
- Commande pour un crawl :
  - `wget -r -l2 -P www -R jpg,gif,png 'http://www.cetic.be'`
    - Crawl récursif de profondeur 2 pour le site « http://www.cetic.be » et résultats du crawl dans le répertoire « www » (+ rejet des photos).



CETIC a.s.b.l. • Bâtiment Eole • Rue des Frères Wright, 29/3 • B-6041 Charleroi (Belgique) • T. +32 71 490 700 • F. +32 71 490 799 • info@cetic.be

# Wget (2)

- Multiples options :
  - -r (crawl récursif)
    - Par défaut : respect de la convention « *norobots* »
  - -l (profondeur de récursion)
  - -P (répertoire cible pour le stockage)
  - -A et -R (filtrage des URLs par pattern)
  - --user-agent (« user agent » imposé)
  - ...
- Plus d'infos : http://www.gnu.org/software/wget/manual/ .
- Alternative : curl (puissant mais... pas de crawl récursif).

# Lucene (1)

- Outil d'indexation supporté par la fondation Apache (lucene.apache.org).
- Ecosystème étendu :
  - Utilisé dans Alfresco, Jahia, Liferay,...
  - Extension au cloud (ex. : CouchDB-lucene).
  - Beaucoup d'outils tiers : Luke (lecture d'un index), Solr (serveur de recherche ; sans crawler), Nutch (moteur de recherche avec crawler), Carrot2 (interface de recherche compatible OpenSearch et Solr),...
  - Cf. http://wiki.apache.org/lucene-java/PoweredBy .

# Lucene (2)

- Le format d'index est devenu une sorte de standard.
  - Nombreux portages : Lucene.Net (. Net), PyLucene (Python), CLucene (C++) Plucene (Perl), Zend Search (PHP),...
  - Différents types de portages : par traduction littérale (compatibilité d'API d'abord), par traduction optimisée pour le langage cible (performances d'abord) et par binding (Python).
  - Points à surveiller : couverture fonctionnelle, version de l'index,...



CETIC a.s.b.l. • Bâtiment Eole • Rue des Frères Wright, 29/3 • B-6041 Charleroi (Belgique) • T. +32 71 490 700 • F. +32 71 490 799 • info@cetic.be

# Zend Search (1)

- Portage de Lucene en PHP.
  - API spécifique.
  - Support de la version d'index 2.3 (depuis Zend Framework 1.6).
  - Support de l'UTF-8 en interne.
  - Intégré au framework Zend mais utilisable séparément (taille sur disque : 734,3 ko).
  - Taille d'index théorique maximum = 2GB (système 32 bits).
  - Facilement hébergeable (installable sur un hébergement mutualisé type OVH ou Lost Oasis).

# Zend Search (2)

- Plusieurs types de champs supportés :
  « Keyword », « UnIndexed », « Binary », « Text » et « UnStored ».

| Type de champ | Stocké | Indexé | Tokenisé | Binaire |
|---|---|---|---|---|
| *Keyword* | x | x | | |
| *UnIndexed* | x | | | |
| *Binary* | x | | | x |
| *Text* | x | x | x | |
| *UnStored* | | x | x | |



CETIC a.s.b.l. • Bâtiment Eole • Rue des Frères Wright, 29/3 • B-6041 Charleroi (Belgique) • T. +32 71 490 700 • F. +32 71 490 799 • info@cetic.be

# Zend Search (3)

- Syntaxe supportée lors des recherches :
  - Opérateurs booléens (« OR » ou « || », « AND » ou « && », « NOT » ou «  ! », « + », « - »),
  - indicateur de champs (« title: »),
  - jokers (« ? » ou « * »),
  - recherche par intervalle (dates ou chaines),
  - recherche floue (« ~ »),
  - recherches de proximité (« ~ »),
  - facteur de boost (« ^ »).
- Possibilité de trier par champs.
- Cf. http://framework.zend.com/manual/fr/zend.search.lucene.searching.html et http://framework.zend.com/manual/fr/zend.search.lucene.query-language.html .

# Zend Search (4)

- Possibilité de personnaliser l'analyse des documents :
  - Utilisation d'un analyseur par défaut (Zend_Search_Lucene_Analysis_Analyzer_Common_Text_CaseInsensitive).
  - Possibilité de choisir un analyseur (compatible UTF-8, compatible avec les nombres,...).
  - Possibilité de configurer des filtres (« lowercase », « stop words », « short words »).
  - Possibilité de créer son propre analyseur.
  - Cf. http://framework.zend.com/manual/fr/zend.search.lucene.extending.html .
- Possibilité de chargement direct de documents : formats HTML, MS Word / Powerpoint / Excel,...



CETIC a.s.b.l. • Bâtiment Eole • Rue des Frères Wright, 29/3 • B-6041 Charleroi (Belgique) • T. +32 71 490 700 • F. +32 71 490 799 • info@cetic.be

# Divers

- Autres outils utiles (pour un moteur de recherche) :
  - En CLI :
    - PDFtoText : conversion d'un document PDF en texte brut.
  - En PHP :
    - SimplePie (simplepie.org) : lecteur RSS robuste.
    - GeoIP PHP API (www.maxmind.com): géolocalisation d'adresses IP (cf. « pure PHP module »).
    - Dans PEAR :
      - PEAR Text_LanguageDetect (pear.php.net) : détection de la langue d'une phrase.
    - Dans Zend :
      - Zend_Paginator : gestion des pages (collection de données).
      - Zend_Tag : création d'un nuage de tags.
      - Zend_Service : passerelles vers plusieurs API populaires (Delicious, Twitter,...).
  - ...

# Mise en oeuvre



CETIC a.s.b.l. • Bâtiment Eole • Rue des Frères Wright, 29/3 • B-6041 Charleroi (Belgique) • T. +32 71 490 700 • F. +32 71 490 799 • info@cetic.be

# Encodage UTF-8 sous PHP (1)

- PHP 5 (et <5) travaille en ISO-8859-1.
- Problème ?
  - L'UTF-8 permet de présenter davantage de jeux de caractères que l'ISO 8859-1 mais...
  - L'UTF-8 stocke les caractères sur 1 ou plusieurs octets (1 seul en ISO-8859-1).
- Donc ?
  - L'UTF-8 est mieux adapté à l'Internet.
  - La chaîne d'outils utilisés (éditeur, langage de script, base de données, navigateur,...) doit connaître l'encodage utilisé pour comprendre les chaînes.
    - Exemple de problème : « Archive — Little I dreamed of being like LÂ©guman ... now I eat carrots ».
- Cf. http://openweb.eu.org/articles/jeux_caracteres .

# Encodage UTF-8 sous PHP (2)

- Dans PHP :
  - La logique est différente de celle de Python, qui propose un type « string » et un type « Unicode ». PHP est très faiblement typé et travaille uniquement avec des chaines.
  - La conversion entre ISO et UTF-8 se fait à l'aide des fonctions « utf8_encode() » et « utf8_decode() ». Les autres conversions se font via « iconv ».
  - La manipulation des chaines en UTF-8 se fait à l'aide de la bibliothèque « mbstring ».
- La détermination de l'encodage des caractères en entrée n'est pas triviale (headers HTTP parfois erronés, métadonnées HTML parfois absentes ou erronées, outils de détection pas toujours fiables,...).

# Zend Search (1)

- Création d'un index et insertion :

```
$index = Zend_Search_Lucene::create('/data/my-index');
$doc = new Zend_Search_Lucene_Document();
$doc->addField(Zend_Search_Lucene_Field::Text('title', $docTitle));
$doc->addField(Zend_Search_Lucene_Field::UnIndexed('url', $docUrl));
$doc->addField(Zend_Search_Lucene_Field::UnStored('content',
$docContent));
$index->addDocument($doc);
$index->commit() ;
```

- Optimisation de l'index :

```
$index->optimize();
```



CETIC a.s.b.l. • Bâtiment Eole • Rue des Frères Wright, 29/3 • B-6041 Charleroi (Belgique) • T. +32 71 490 700 • F. +32 71 490 799 • info@cetic.be

# Zend Search (2)

- Ouverture d'un index et recherche :

```
$index = Zend_Search_Lucene::open('/data/my-index') ;
$query = Zend_Search_Lucene_Search_QueryParser::parse($input);
$hits  = $index->find($query);
foreach ($hits as $hit) {
  echo $hit->score;
  echo $hit->title;
  echo $hit->url;
}
```

- Paramètres : opérateur par défaut, encodage des données,...



CETIC a.s.b.l. • Bâtiment Eole • Rue des Frères Wright, 29/3 • B-6041 Charleroi (Belgique) • T. +32 71 490 700 • F. +32 71 490 799 • info@cetic.be

# Divers

- Attention : wget affecte conventionnellement le nom « index.html » à l'adresse « / ».
- Analyse des documents HTML :
  - native (cf. « Zend_Search_Lucene_Document_Html »),
  - manuelle non structurée (extraction des métadonnées -title, description,...- par expressions régulières et nettoyage du <body> via « strip_tags ») ou...
  - manuelle structurée (extraction structurée par expression régulière ou Xpath).
- Lucene ne gère pas les contraintes d'intégrité.
  - Les doublons ne peuvent donc pas être évités via un champ « UNIQUE ».
  - Or, *hash* utile sur l'URL, voire sur le contenu (*duplicate content*).
  - Solutions possibles : test sur un champ unique dans l'index ou test sur base d'une table externe (ex. : *hash* dans SQLite).

# Exemple : indexation de flux RSS

- Indexation de flux RSS (retronimo.com).
  - Collecte des URLs des flux RSS et Atom via un crawler multithread Python.
  - Lecture, indexation et interface de recherche en PHP.
    - Lecture des flux RSS avec SimplePie, localisation du serveur avec GeoIP et détection de la langue sur base de liste de « stop words ».

# Exemple : pages HTML (1)

Requête:

15 entité(s) trouvée(s) pour la requête '*python*'.

- Sur *www.emencia.fr* : Prestataire Paris - Zope, Python, Zwook, Plone, CMS, E-Commerce, Travail collaboratif - Intranet / Extranet (43 pages(s) référencée(s) pour cette entité).
- Sur *www.camptocamp.com* : Développeur Python - Camptocamp (1 pages(s) référencée(s) pour cette entité).
- Sur *www.makina-corpus.com* : Python | Makina Corpus (2 pages(s) référencée(s) pour cette entité).
- Sur *www.anaska.com* : Formation zope, formation plone, formation python (1 pages(s) référencée(s) pour cette entité).
- Sur *www.syleam.fr* : Python & XML / formations / Accueil - Syleam (1 pages(s) référencée(s) pour cette entité).
- Sur *www.2le.net* : Python | 2le - Logiciels libres pour l'entreprise (1 pages(s) référencée(s) pour cette entité).
- Sur *www.easter-eggs.com* : CRM et GPAO sur mesure en Python/GTK - Easter-eggs - Spécialiste GNU/Linux (1 pages(s) référencée(s) pour cette entité).
- Sur *www.sednacom.fr* : Progiciels (1 pages(s) référencée(s) pour cette entité).
- Sur *www.gnurandal.com* : Gnurandal – Areas of expertise (1 pages(s) référencée(s) pour cette entité).
- Sur *www.novelys.com* : No title (1 pages(s) référencée(s) pour cette entité).
- Sur *free-electrons.com* : Buildroot 2011.02 released, with many interesting updates and commercial support! - Free Electrons blog (7 pages(s) référencée(s) pour cette entité).
- Sur *www.entrouvert.com* : Entr'ouvert - Entr'ouvert - Lasso (3 pages(s) référencée(s) pour cette entité).
- Sur *www.openwide.fr* : No title (2 pages(s) référencée(s) pour cette entité).
- Sur *www.codelutin.com* : No title (1 pages(s) référencée(s) pour cette entité).
- Sur *www.internethic.com* : Recrutement / Société / Internethic - création site web Open Source eZ publish Open Erp (3 pages(s) référencée(s) pour cette entité).

Index total: 4299 document(s).

# Exemple : pages HTML (2)

- Étape 1 : constituer une base de données d'URLs (basé sur logiciellibre.com).
- Étape 2 :
  - Détecter les éventuelles redirections, sites morts, etc (automatisable sous PHP avec « get_headers »).
  - Générer les requêtes wget correspondantes.
  - Lancer le crawl avec « wget ».
- Étape 3 : lancer l'indexation des pages collectées par « wget ».

- Utilité : identifier les prestataires actifs sur une ou plusieurs technologies particulières (cf. thème « Entreprises »).



CETIC a.s.b.l. • Bâtiment Eole • Rue des Frères Wright, 29/3 • B-6041 Charleroi (Belgique) • T. +32 71 490 700 • F. +32 71 490 799 • info@cetic.be

# Exemple : pages HTML (3)

- Performances :

```
Active index: index-fr:
  Create index:
    From 30-06-2011 13:39:16 to 30-06-2011 13:44:28.
    Size of the index: 4299 document(s) and 35,5Mo.
    Time: 258,957s. (60,237ms./doc.).
  Optimization time: 49,154s.
  Tests:
    Search (test): 25 result(s) (max.: 25) in 10,130ms. for 'python'.
    Search (test): 69 result(s) (max.: 250) in 4,425ms. for 'python'.
    Search (test): 25 result(s) (max.: 25) in 11,798ms. for 'python AND plone'.
    Search (test): 27 result(s) (max.: 250) in 7,111ms. for 'python AND NOT plone'.
    Search (test): 102 result(s) (max.: 250) in 8,569ms. for 'python OR plone^4'.
```



CETIC a.s.b.l. • Bâtiment Eole • Rue des Frères Wright, 29/3 • B-6041 Charleroi (Belgique) • T. +32 71 490 700 • F. +32 71 490 799 • info@cetic.be

# Conclusion



CETIC a.s.b.l. • Bâtiment Eole • Rue des Frères Wright, 29/3 • B-6041 Charleroi (Belgique) • T. +32 71 490 700 • F. +32 71 490 799 • info@cetic.be

# Conclusion (1)

- Grande facilité d'intégration dans un programme PHP. Supporté sur la plupart des hébergements LAMP, même mutualisés.
- Réponses pertinentes. Grande richesse du langage d'interrogation. Interprétation des requêtes.
- Possibilité de choisir et de personnaliser les analyseurs de texte.
- Fonctionnement correct pour des index de taille limitée :
  - Fragilité de l'index lors d'une forte sollicitation en insertion (requêtes simultanées) depuis une application Web (risque d'index corrompu à partir de 5000 documents environ) mais...
  - Problème constaté sous Windows XP et pas sous Ubuntu (?).

# Conclusion (2)

- Deux tests réalisés :
  - Insertions simultanées depuis une application Web (requêtes sur « `localhost` »).
    - Index : 12.065 documents (contenus RSS) et 22 Mo.
  - Insertions séquentielles depuis la ligne de commande (« `php monscript.php` »).
    - Index : 15.790 documents (pages HTML) et 167,5 Mo.
    - Dans ce cas : lancement sur serveur si accès « administrateur » ou synchronisation de l'index depuis un poste local (par exemple via synchronisation FTP, cf. LFTP).

# Merci!

Merci pour votre attention.

Des questions ?



CETIC a.s.b.l. • Bâtiment Eole • Rue des Frères Wright, 29/3 • B-6041 Charleroi (Belgique) • T. +32 71 490 700 • F. +32 71 490 799 • info@cetic.be

# Quelques ressources



CETIC a.s.b.l. • Bâtiment Eole • Rue des Frères Wright, 29/3 • B-6041 Charleroi (Belgique) • T. +32 71 490 700 • F. +32 71 490 799 • info@cetic.be

# Outils

- SQLite (www.sqlite.org).
- WampServer (www.wampserver.com).
- Lucene (lucene.apache.org).
- Zend framework (framework.zend.com).
- SolR (lucene.apache.org/solr/).
- Carrot² (project.carrot2.org).
- Luke (www.getopt.org/luke/).
- Nutch (nutch.apache.org).
- Tesseract (tesseract-ocr.googlecode.com).



CETIC a.s.b.l. • Bâtiment Eole • Rue des Frères Wright, 29/3 • B-6041 Charleroi (Belgique) • T. +32 71 490 700 • F. +32 71 490 799 • info@cetic.be

# Ressources et liens utiles


- Robert Viseur (2010). "Introduction to libre « fulltext » technology". RMLL 2010. URL : http://www.robertviseur.be/news-20100710.php .
- Erik Hatcher et Otis Gospodnetić (2004). "Lucene in Action". Manning Publications Co.
- Moteur de recherche avec Zend Search Lucene. URL : http://www.libre-a-vous.fr/moteur-recherche-zend-search-lucene/ .
- GNU Wget. URL: http://www.gnu.org/software/wget/manual/ .
- Introduction aux jeux de caractères. URL: http://openweb.eu.org/articles/jeux_caracteres .

# Contact


- Dr Ir Robert Viseur.
- Email : robert.viseur@cetic.be
- Phone : 0032 (0) 479 66 08 76