
	Titre du document : Paramétrage de SOLR pour les besoins spécifiques des forges
	Référence : L3.4.1
	Version du 0.02

Paramétrage de SOLR pour les besoins spécifiques des forges


Livrable du au titre du projet	COCLICO
Lot	
Tache	
Livrable	L3.4.1

Rédacteur(s)	Vérificateur(s)	Approbateur(s)
Hratchia Pélibossian Celi-France		

	Titre du document : Paramétrage de SOLR pour les besoins spécifiques des forges
	Référence : L3.4.1
	Version du 0.02


Documents applicables	
Annexe technique au projet COCLICO	
Livrable avec Solr paramétré	https://forge.projet-coclico.org/frs/?group_id=9 wp3 solr-celi_0.02
Code-Sources Celi extractor	https://forge.projet-coclico.org/scm/loggerhead/wp3/celi/files

Documents de références (pour information)
http://lucene.apache.org/solr/
http://lucene.apache.org/solr/tutorial.html
https://forge.projet-coclico.org/projects/wp3/
http://www.projet-coclico.org/index.php/DESCRIPTION_TECHNIQUE_DU_PROJET

	Titre du document : Paramétrage de SOLR pour les besoins spécifiques des forges
	Référence : L3.4.1
	Version du 0.02


Gestion des versions

N° de version	Date	Auteurs	Modification apportées
0.01	03/03/10	Pélibossian	Première version
0.02	04/04/10	Pélibossian	Paramétrage de la langue française, Gestion des permissions pour accessibilité des documents

	Titre du document : Paramétrage de SOLR pour les besoins spécifiques des forges
	Référence : L3.4.1
	Version du 0.02

Sommaire


Objectif	5
1. Solr	6
1.1. Pré-requis	6
1.2. Solr Installation	6
1.3. Configurer Solr pour un fonctionnement avec Jetty	6
1.4. Les fichiers de configuration	6
2. Celi extraction	7
2.1. Installation	7
2.2. Configuration de schema.xml pour les méta-données des forges	7
3. Client Php Installation	12

	Titre du document : Paramétrage de SOLR pour les besoins spécifiques des forges
	Référence : L3.4.1
	Version du 0.02

Objectif

L'objectif du projet est de fournir les Forges avec des fonctionnalités de recherche plein texte sur les méta-données du forge et sur les différents types d'information stockés. Il doit obéir aux contraintes suivantes:

- Une répercussion minimale sur l'utilisation standard du Forge;
- Une charge administrative minimale;
- Une facilité de la configuration.

	Titre du document : Paramétrage de SOLR pour les besoins spécifiques des forges
	Référence : L3.4.1
	Version du 0.02

1 Solr

1.1 Pré-requis

- Installer Java 1.5 ou version supérieur. Téléchargement de Java SE est disponible ici.
- Installer Tomcat 5.5. Téléchargement est disponible ici.
- Récupérer une distribution de Solr. C'est ici pour la version officielle et ici et pour les derniers builds nocturnes.

1.2 Solr Installation

Décompresser l'archive dans un répertoire de travail. Automatiquement un sous-répertoire est créé. Selon la version de Solr que vous aurez récupéré, ce répertoire peut s'appeler « apache-solr-nightly » ou « apache-solr-1.4.0 ». Je suggère donc pour simplifier de renommer ce répertoire en « apache-solr ».

1.3 Configurer Solr pour un fonctionnement avec Jetty

Pour faire fonctionner **Solr** avec **Jetty**, il n'y a pas besoin de configuration. Pour démarrer **Jetty**, il faut ouvrir une console et se placer dans le répertoire `d:\solr\apache-solr\example` et exécuter la commande suivante :
java -jar start.jar

Une fois que vous voyez apparaître la ligne suivante, le serveur est démarré :
 INFO: [] Registered new searcher Searcher@d642fd main

On accède alors à l'administration de **Solr** avec l'url suivante :
<http://localhost:8983/solr/admin>

Plus d'information sur l'installation de **Solr** est également disponible ici
<http://www.wiizio.com/2009/10/02/introduction-a-solr-installation-et-configuration-1/>.

1.4 Les fichiers de configuration

Les fichiers de configuration sont localisés dans le répertoire `/conf`. Il s'agit principalement de `solrconfig.xml` et `schema.xml`.


solrconfig.xml

Il s'agit du fichier qui contient l'essentiel des paramètres liés au fonctionnement de Solr et plus particulièrement des paramètres de l'API Lucene qu'utilise Solr (longueur maximale d'un champ, taille des buffer mémoire, fréquence de commit, etc) . Ce fichier est très bien auto-documenté et fait l'objet d'un document a part entière dans le wiki Solr.

schema.xml

Il s'agit du fichier qui décrit comment seront indexées les données dans Solr. Il définit les types de données, les champs, les manipulations sur les données lors de l'indexation, les champs obligatoires, le champ de requête par défaut, ...

Ce fichier est très bien auto-documenté et fait l'objet d'un document a part entière dans le wiki Solr
<http://wiki.apache.org/solr/SchemaXml>.

	Titre du document : Paramétrage de SOLR pour les besoins spécifiques des forges
	Référence : L3.4.1
	Version du 0.02

2 Celi extractor

Celi extractor est une librairie java qui est sur le site de **Coclico**, précisément dans la partie appelée Source du projet de WP3 https://forge.projet-coclico.org/scm/?group_id=9/celiextractor/lib/celi-apache-solr-cell-1.5-dev.jar.

Celi extractor permet d'indexer des documents de différents format avec les metadonnées supplémentaires sélectionnés par l'administration de forge.

2.1 Installation

- Mettre la librairie `celi-apache-solr-cell-1.5-dev.jar` dans le dossier de dist (par exemple `solr/dist/`) et définir son chemins dans le fichier `solrconf.xml` :

```
<lib dir="../../dist/" regex="celi-apache-solr-cell-.*\.jar" />
```

- Définir dans `solrconf.xml` le **celi/update** request handler:

```
<requestHandler name="/update/celi" class="org.apache.solr.handler.extraction.CeliUpdateRequestHandler" />
```

- Redémarrer **Solr** serveur et notre librairie et request handler sera prise en compte par **Solr**.

2.2 Configuration de schema.xml pour les métadonnées des forges

Les metadonnée que nous disposons sont suivants:


Documents

Title
Description
Owner
Language
Create Date
Update Date
Extra Fields

créer pour chaque métadonnées du document un champs dans [schema.xml](#)

```
<field name="file_address" type="string" indexed="true" stored="true" />
<field name="external_file" type="file_content" indexed="true" stored="false" multiValued="true" />
<field name="doc_title" type="text" indexed="true" stored="true" multiValued="true" />
<field name="doc_description" type="text" indexed="true" stored="true" multiValued="true" />
<field name="doc_owner" type="text" indexed="true" stored="true" multiValued="true" />
<field name="doc_create_date" type="date" indexed="true" stored="true" />
<field name="doc_update_date" type="date" indexed="true" stored="true" />
<field name="doc_language" type="text" indexed="true" stored="true" multiValued="true" />
<field name="doc_extra_field" type="text" indexed="true" stored="false" multiValued="true" />
```

Les champs **file_adresse** et **external_file** font partie d'une déclaration d'un champ qui utilise **CeliExternalFileField** type et qui est défini dans librairie `celi-apache-solr-cell-1.5-dev.jar`. Pour pouvoir utiliser ce type de champs il faut le définir dans [schema.xml](#):

	Titre du document : Paramétrage de SOLR pour les besoins spécifiques des forges
	Référence : L3.4.1
	Version du 0.02

```
<fieldType name="file_content" keyField="file_address"
class="org.apache.solr.schema.CeliExternalFileField" valType="text">
```

```
.....
</ fieldType>
```

Ce type permet de trouver le fichier grâce à l'adresse **file_address** , extraire sont contenu et de l'indexer avec les metadonnées fournit par forge.

Mailing List

Subject
Sender
Date

```
<field name="mail_subject" type="text" indexed="true" stored="true" multiValued="true" />
<field name="mail_sender" type="text" indexed="true" stored="true" multiValued="true" />
<field name="mail_create_date" type="date" indexed="true" stored="true" />
```

Forums

Forum name
Subject
Body
Author
Date

```
<field name="forum_name" type="text" indexed="true" stored="true" multiValued="true" />
<field name="forum_subject" type="text" indexed="true" stored="true" multiValued="true" />
<field name="forum_subject_address" type="string" indexed="true" stored="true" />
<field name="forum_body" type="text" indexed="true" stored="false" multiValued="true" />
<field name="forum_author" type="text" indexed="true" stored="true" multiValued="true" />
<field name="forum_date" type="date" indexed="true" stored="true" />
```

News Voir Forums


Source Code

File name
Creation date
Update date
Owner
Last committer

```
<field name="source_file_name" type="text" indexed="true" stored="true" multiValued="true" />
<field name="source_owner" type="text" indexed="true" stored="true" multiValued="true" />
<field name="source_create_date" type="date" indexed="true" stored="true" />
<field name="source_update_date" type="date" indexed="true" stored="true" />
<field name="source_last_committer" type="text" indexed="true" stored="true" multiValued="true" />
<field name="source_file_name_address" type="string" indexed="true" stored="true" />
<field name="source_external_file" type="source_file_content" indexed="true" stored="false"
multiValued="true" />
```

Les champs **source_file_name_address** et **source_external_file** sert pour l'indexation du fichier source via mécanisme de celi-extractor.

File release system

	Titre du document : Paramétrage de SOLR pour les besoins spécifiques des forges
	Référence : L3.4.1
	Version du 0.02

Package name
Release name
Release date
Owner

```
<field name="release_package_name" type="text" indexed="true" stored="true" multiValued="true" />
<field name="release_release_name" type="text" indexed="true" stored="true" multiValued="true" />
<field name="release_release_name_address" type="string" indexed="true" stored="true" />
<field name="release_owner" type="text" indexed="true" stored="true" multiValued="true" />
<field name="release_date" type="date" indexed="true" stored="true" />
```

Trackers

Trackers ont beaucoup de données semi-structuré. Il y a plusieurs ontologies rédigés pour modéliser ces données, mais un des plus intéressant est le OSLC-CM V2 standart, voire ici <http://open-services.net/bin/view/Main/CmResourceDefinitionsV2>.

```
<field name="tracker_title" type="text" indexed="true" stored="true" multiValued="true" />
<field name="tracker_identifi er" type="text" indexed="true" stored="true" multiValued="true" />
<field name="tracker_identifi er_address" type="string" indexed="true" stored="true" />
<field name="tracker_type" type="text" indexed="true" stored="true" multiValued="true" />
<field name="tracker_description" type="text" indexed="true" stored="false" multiValued="true" />
<field name="tracker_subject" type="text" indexed="true" stored="false" multiValued="true" />
<field name="tracker_creator" type="text" indexed="true" stored="true" multiValued="true" />
<field name="tracker_modified" type="date" indexed="true" stored="true" />
<field name="tracker_name" type="text" indexed="true" stored="true" multiValued="true" />
<field name="tracker_created" type="date" indexed="true" stored="true" />
<field name="tracker_project" type="string" indexed="true" stored="true" multiValued="true" />
<field name="tracker_component" type="string" indexed="true" stored="true" multiValued="true" />
<field name="tracker_status" type="string" indexed="true" stored="true" multiValued="true" />
<field name="tracker_owner" type="text" indexed="true" stored="true" multiValued="true" />
<field name="tracker_priority" type="string" indexed="true" stored="true" multiValued="true" />
<field name="tracker_severity" type="string" indexed="true" stored="true" multiValued="true" />
<field name="tracker_relatedChangeRequests" type="text" indexed="true" stored="true"
multiValued="true" />
<field name="tracker_changeSets" type="text" indexed="true" stored="true" multiValued="true" />
<field name="tracker_comments" type="text" indexed="true" stored="false" multiValued="true" />
<!-- use celi extractor mecanisme -->
<field name="tracker_attachments_address" type="string" indexed="true" stored="true" />
<field name="tracker_external_file" type="tracker_file_content" indexed="true" stored="false"
multiValued="true" />
```

Multi Langue


Actuellement nous proposons le paramétrage de langues anglaise et française.

Pour indexer les données et les documents français nous proposons d'utiliser le mécanisme dynamique field de SOLR que genre dynamiquement des champs dont les noms correspondents aux certains règle de l'expression régulier.

```
<dynamicField name="*_sm_fr" type="text_fr" indexed="true" stored="true" multiValued="true"/>
<dynamicField name="*_s_fr" type="text_fr" indexed="true" stored="true"/>
<dynamicField name="*_m_fr" type="text_fr" indexed="true" stored="false" multiValued="true"/>
<dynamicField name="*_fr" type="text_fr" indexed="true" stored="false"/>
```

Tout élément du datagramme XML dont le nom se termine par **_sm_fr**, est considéré comme un élément du type **text_fr** qui sont stockés et multivalués. Par exemple équivalent du champs

```
<field name="doc_title" type="text" indexed="true" stored="true" multiValued="true" />
```

	Titre du document : Paramétrage de SOLR pour les besoins spécifiques des forges
	Référence : L3.4.1
	Version du 0.02

pour les documents français sera

```
<field name="doc_title_sm_fr" type="text_fr" indexed="true" stored="true" multiValued="true" />
```

Nous devons également déclarer le type **text_fr** dans schema.xml pour pouvoir appliquer les traitements spécifiques pour la langue française, ce sont les traitements de lexémisation, des accents et autres.

```
<fieldType name="text_fr" class="solr.TextField" positionIncrementGap="100">
  <analyzer>
    <charFilter class="solr.MappingCharFilterFactory" mapping="mapping-ISOLatin1Accent.txt"/>
    <!-- this is alternativ accent char treatment <filter class="solr.ASCIIFoldingFilterFactory"/> -->
    <tokenizer class="solr.WhitespaceTokenizerFactory"/>
    <filter class="solr.StopFilterFactory"
      ignoreCase="true"
      words="stopwords_fr.txt"
      enablePositionIncrements="true"
    />
    <filter class="solr.WordDelimiterFilterFactory"
      generateWordParts="1" generateNumberParts="1" catenateWords="1"
      catenateNumbers="1" catenateAll="0" splitOnCaseChange="1"/>
    <filter class="solr.LowerCaseFilterFactory"/>
    <filter class="solr.SnowballPorterFilterFactory" language="French"/>
    <filter class="solr.RemoveDuplicatesTokenFilterFactory"/>
  </analyzer>
</fieldType>
```


Groupe de permission

Nous avons déclaré dans schema.xml le champ

```
<field name="forge_permission_group" type="text" indexed="true" stored="true" multiValued="true"/>
```

pour proposer aux utilisateurs seulement les résultats de recherche auxquels ils ont le droit d'accéder.

Le paramétrage de **Solr** pour les forges est terminé.

	Titre du document : Paramétrage de SOLR pour les besoins spécifiques des forges
	Référence : L3.4.1
	Version du 0.02

3 Client Php Installation.

Récupérer le répertoire de **SolrPhpClient** sur le site de coclico précisément dans la partie Source du projet de WP3 https://forge.projet-coclico.org/scm/?group_id=9. Mettre avec vos autres fichiers PHP de votre Forge.

Indexation: **Exécution de ces fichiers test:**

Accéder aux fichiers doc_celitest.php ou source_celitest.php ou tracker_test.php via un internet browser .
Automatiquement le code php de ces fichier sera exécuté. Voici le code de doc_celitest.php

```
<?php
require_once('pre.php');

echo "<html><body>Test Solr..<br> </body></html>";
//CeliService.php est créer ppar celi pour pouvoir utiliser la librairie celi-apache-solr-cell-1.5-dev.jar
require_once( './SolrPhpClient/Apache/Solr/CeliService.php' );

//connection au serveur solr
$solr = new Celi_Apache_Solr_Service( 'localhost', '8983', '/solr' );


if ( ! $solr->ping() ) {
    echo "<html><body>Test do not work Solr with php client.<br></body></html>";
}

if ( $solr->ping() ) {
    echo "<html><body>Test Work Solr with php client.<br> </body></html>";
}

//
//
// Create a solr document
//
$document = new Apache_Solr_Document();
$document->id = '0002_doc_php';
$document->comments= 'Test with PHP';
// adresse de fichier dont le contenu sera indexer via celi extractor mechanisme
$document->file_address = '/home/solr/apache-solr-1.4.0/example/exampledocs/Doc-Word.doc';

$document->doc_title = 'Document Word to add Solr Index';
$document->doc_description = 'Document Word to add Solr Index';
$document->doc_owner = '4-101';
$document->doc_create_date = '2005-12-31T23:59:59Z';
$document->doc_update_date = '2007-01-22T23:01:02Z';
$document->doc_language = 'english';
$document->doc_extra_field = 'personale note: this test is simple';

$solr->addDocument($document);
$solr->commit(); //commits to see the deletes and the document
$solr->optimize(); //merges multiple segments into one
```

	Titre du document : Paramétrage de SOLR pour les besoins spécifiques des forges
	Référence : L3.4.1
	Version du 0.02

```
echo "<html><body>END <br> </body></html>";
```

```
?>
```

Ce teste connecte a **Solr** server , crée un document **Solr** via celi handler et indexe le document.

Recherche de documents:

Nous allons réaliser une recherche d'un document a l'aide de code PHP suivante:

```
$offset = 0;
$limit = 10; // nombre de résultat à retourner

// composition de la requête
$queryes = array( external_file: word add Solr Index, doc_language: english,);

foreach ( $queryes as $query ) {
    $response = $solr->search( $query, $offset, $limit );

    if ( $response->getHttpStatus() == 200 ) {
        // print_r( $response->getRawResponse() );

        if ( $response->response->numFound > 0 ) {
            echo "$query <br />";


            foreach ( $response->response->docs as $doc ) {
                echo "$doc->external_file $doc->doc_language <br />";
            }

            echo '<br />';
        }
    }
    else {
        echo $response->getHttpStatusMessage();
    }
}
```

le résultat de la recherche est:

```
<response>
<lst name="responseHeader">
  <int name="status">0</int>
  <int name="QTime">1</int>
  <lst name="params">
    <str name="indent">on</str>
    <str name="start">0</str>
    <str name="q">external_file: word add Solr Index doc_language: english doc_extra_field:
      "personale note: this test is simple"</str>
    <str name="version">2.2</str>
    <str name="rows">10</str>
  </lst>
</lst>

<result name="response" numFound="1" start="0">
```

	Titre du document : Paramétrage de SOLR pour les besoins spécifiques des forges
	Référence : L3.4.1
	Version du 0.02

```

<doc>
  <str name="comments">Test with PHP</str>
  <date name="doc_create_date">2005-12-31T23:59:59Z</date>
  <arr name="doc_description">
    <str>Document Word to add Solr Index</str>
  </arr>
  <arr name="doc_language">
    <str>english</str>
  </arr>
  <arr name="doc_owner">
    <str>4-101</str>
  </arr>
  <arr name="doc_title">
    <str>Document Word to add Solr Index</str>
  </arr>
  <date name="doc_update_date">2007-01-22T23:01:02Z</date>
  <str name="file_address">/home/solr/apache-solr-1.4.0/example/exampledocs/Doc-Word.doc</str>
  <str name="id">0002_doc_php</str>
</doc>
</result>
</response>

```