Lucene

0.1 Documentation

by Michael Wechner

Table of contents

1. Generic Search

URL:

```
/lenya/$PUB_ID/search-$AREA/lucene
```

Indices and Excerpts:

```
src/webapp/lenya/pubs/$PUB_ID/work/search/index/$AREA/index
src/webapp/lenya/pubs/$PUB_ID/work/search/htdocs_dump/$AREA
```

Configuration:

```
src/webapp/global-sitemap.xmap
src/webapp/lenya/lucene.xmap
```

2. Customizing/Overwriting Generic Search Interface

XSLT:

```
src/webapp/lenya/pubs/$PUB_ID/lenya/xslt/search/search-and-results.xsl
```

URL:

```
/lenya/$PUB_ID/search-$AREA/lucene
```

3. Crawling a website

Crawl a website by running

```
ant -f src/webapp/lenya/bin/crawl_and_index.xml crawl
-Dcrawler.xconf=/home/username/src/cocoon-lenya/src/webapp/lenya/pubs/default/config/search/crawler-li
```

whereas the crawler.xconf has the following elements

```
<crawler>
  <user-agent>lenya</user-agent>

  <base-url href="http://cocoon.apache.org/lenya/index.html"/>
  <scope-url href="http://cocoon.apache.org/lenya/"/>

  <uri-list src="work/search/lucene/uris.txt"/>
  <htdocs-dump-dir src="work/search/lucene/htdocs_dump/cocoon.apache.org"/>

  <!-- <robots src="robots.txt" domain="cocoon.apache.org"/> -->
</crawler>
```

where the element robots is optional.

In case you don't have access to the server and want to disallow certain URLs from being crawled, then you can also define a "robots.txt" on the crawler side, e.g.

```
# cocoon.apache.org

User-agent: *
Disallow: /there_seems_to_be_a_bug_within_websphinx_Robot_Exclusion.html
#Disallow:

User-agent: lenya
Disallow: /do/not/crawl/this/page.html
```

4. Creating an index from the command line

```
ant -f src/webapp/lenya/bin/crawl_and_index.xml
-Dlucene.xconf=/home/username/src/cocoon-lenya/src/webapp/lenya/pubs/default/config/search/lucene-live
index
```

whereas the lucene.xconf has the following elements

```
<lucene>
  <update-index type="new"/>
  <!--
  <update-index type="incremental"/>
  -->

  <index-dir src="../../work/search/lucene/index/index"/>
    <htdocs-dump-dir src="../../work/search/lucene/htdocs_dump"/>

    <indexer class="org.apache.lenya.lucene.index.DefaultIndexer"/>
<!--
    <indexer class="org.apache.lenya.lucene.index.ConfigurableIndexer">
      <configuration src="cmfs-luceneDoc.xconf"/>
      <extensions src="xml"/>
    </indexer>
-->
<!--
    <indexer class="org.apache.lenya.lucene.index.ConfigurableIndexer">
      <configuration src="cmfs-luceneDoc.xconf"/>
      <filter class="foo.bar.FileFilter"/>
    </indexer>
-->
</lucene>
```

5. Indexing XML documents

In order to index XML documents one needs to configure the `org.apache.lenya.lucene.index.ConfigurableIndexer` (see above).

With namespaces:

```
<?xml version="1.0"?>

<luc:document xmlns:luc="http://apache.org/cocoon/lenya/lucene/1.0">
  <luc:field name="currwfstate" type="Text" xpath="/wf:history/wf:version[last()]/@state">
    <namespace prefix="wf">http://apache.org/cocoon/lenya/workflow/1.0</namespace>
  </luc:field>
```

```
</luc:document>
```

Concatenating element values and setting default values in case element value doesn't exist:

```
<?xml version="1.0"?>

<luc:document xmlns:luc="http://apache.org/cocoon/lenya/lucene/1.0">
  <luc:field name="title" type="Text" xpath="/article/head/title"/>
  <luc:field name="subtitle" type="Text" xpath="/article/head/subtitle"/>
  <luc:field name="lead" type="UnStored" xpath="/article/head/abstract"/>
  <luc:field name="contents" type="UnStored" xpath="/"/>
  <luc:field name="author" type="UnStored"/>
    <namespace prefix="lenya">http://apache.org/cocoon/lenya/page-envelope/1.0</namespace>
    <namespace prefix="dc">http://purl.org/dc/elements/1.1/</namespace>
    <xpath>/*/lenya:meta/dc:contributor</xpath>
  </luc:field>
  <luc:field name="date" type="Text">
    <namespace prefix="lenya">http://apache.org/cocoon/lenya/page-envelope/1.0</namespace>
    <xpath default="1969">/*/lenya:meta/year</xpath><text>.</text><xpath
default="02">/*/lenya:meta/month</xpath><text>.</text><xpath default="16">/*/lenya:meta/day</xpath>
  </luc:field>
</luc:document>
```

6. Extract text from a PDF document

```
ant -f src/webapp/lenya/bin/crawl_and_index.xml
-Dhtdocs.dump.dir=/home/username/src/cocoon-lenya/src/webapp/lenya/pubs/default/work/search/lucene/htd
xpdf
```

Also see the targets `pdfbox` and `pdfadobe`.