Lucene

0.1 Documentation

by Michael Wechner

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
```

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. 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`.