Lucene

Table of contents

1 Overview..2
2 Crawling a website..2
3 Creating an index from the command line..3
4 Indexing XML documents..3
5 Extract text from a PDF document...3

1. Overview

There are two URL for the search screen relative to your publication: `search-live/lucene` to search the live area, `search-authoring/lucene` to search the authoring area of your publication.

If you want to customize the layout of the search screen for your publication, place a stylesheet at `lenya/xslt/search/search-and-results.xsl` relative to your publication root.

Lucene indices are stored within the `work/search/index/$AREA/index` directory of your publication. The `work/search/htdocs_dump/$AREA` directory holds content from crawling (see below).

The search pipelines are defined within `global-sitemap.xmap` and `lucene.xmap`

2. Crawling a website

Crawl a website by running

```
ant -f build/lenya/webapp/lenya/bin/crawl_and_index.xml
-Dcrawler.xconf=build/lenya/webapp/lenya/pubs/default/config/search/crawler-live.xconf crawl
```

Note that there is a search.properties file in build/lenya/webapp/lenya/bin that you may have to change. crawler.xconf needs to have the following elements:

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

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

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

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

- user-agent is the HTTP user agent that will be used for the crawler
- base-url is the start URL for the crawler
- scope-url limits the scope of the crawl to that site, or subdirectory
- uri-list is a reference to a file that will contain all URLs found during the crawl
- htdocs-dump-dir specifies the directory that will contain the crawled site
- robots specifies an (optional) robots file that follows the Robot Exclusion Standard (http://www.robotstxt.org/wc/norobots.html)

If you want to fine-tune the crawling (and do not have access to the remote server to put a robots.txt there), then you can specify exlusions in a local robots.txt file:

```
# lenya.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
```

3. Creating an index from the command line

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

Note that there is a search.properties file in build/lenya/webapp/lenya/bin that you may have to change. lucene-live.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"/>
</lucene>
```

4. 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>
```

5. Extract text from a PDF document

```
ant -f build/lenya/webapp/lenya/bin/crawl_and_index.xml
-Dhtdocs.dump.dir=build/lenya/webapp/lenya/pubs/default/work/search/lucene/htdocs_dump xpdf
```

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