Searching Publications How-To

Table of contents

1. Introduction

This article has 4 goals:

1. Complete instructions for indexing on Windows.
2. Display search with a publication's layout rather than the global layout.
3. Filter out a Member's Only area from the results.
4. Fix poor design decisions and bugs.

Changes include:

1. Index live XML files.
   The standard method crawls the site, putting the results into the "htdocs_dump" directory. The index is built from there. The index will not include documents not accessible from the start page, such as Members' Only sections. The index also include navigation menus.
   Using the XML files indexes all content, whether accessible or not, and does not index the site architecture. If a visitor searches on "search", they should receive documents including the word "search", not every page with the search function (which should be every page in a well-designed website).
   The index still includes everything in a document including header information such as author. It is easy to limit the index to the content body, but that could cause complications when using non-standard Lenya documents (Custom DocTypes/Resource Types). If Custom Doctypes are used, modify searchfixer.xsl (see below). **Custom Doctypes were not tested with this configuration.**
2. Remove "Members' Only" documents if not authorized. Visitors must be logged in and in one of the specified Goups. It is the reverse of the current Lenya security, since deep URLs must pass the test for all parents. Example: /employees/programmers must pass tests for both the "/employees" and "/employees/programmers" sections.
3. Add language to the index.
4. Limit initial search to current language.
5. Search page: Remove choice of publications. (This is a design decision. One publication = one website. With protected areas, there should be no need for multiple publications.)
6. Search page: Filter by chosen languages.
7. Default to search "Content", not "Title".
8. Increase the default results per page from 3 to 10.

NOTE: Replace {pub} with your publication name in all instructions.

2. Indexing on Windows

This assumes Lenya 1.2.2 was installed to C:\apache-lenya-1.2.2 If your installation is different, adjust the paths. The indexer adds namespaces to the data of Fields in the index. The namespaces are not used (and are annoying), so remove them. An alternative is to fix the XML later, but why bother? File: *C:\apache-lenya-1.2.2\build\lenya\webapp\WEB-INF\classes\org\apache\lenya\lucene\index\configuration2xslt.xsl* Add the following line:

```
<xsl:template match="namespace"/>
```

1. Set the configuration by changing:
   C:\apache-lenya-1.2.2\build\lenya\webapp\lenya\pubs\{pub}\config\search\lucene-live.xconf To:

```
            <?xml version="1.0"?>
             <lucene>
                 <update-index type="new"/>
                 <index-dir src="../../work/search/lucene/index/live/index"/>
                 <htdocs-dump-dir src="../../content/live"/>
                 <indexer
class="org.apache.lenya.lucene.index.ConfigurableIndexer">
                     <configuration src="lenyadocs.xconf"/>
                     <extensions src="xml"/>
                 </indexer>
             </lucene>
```

2. Create a new file in the same directory to tell lucene what fields to index (filename must match the configuration src in lucene-live.xconf):
   C:\apache-lenya-1.2.2\build\lenya\webapp\lenya\pubs\{pub}\config\search\lenyadocs.xconf Add this:

```
<?xml version="1.0"?>
             <luc:document xmlns:luc="http://apache.org/cocoon/lenya/lucene/1.0">
             <luc:field name="title" type="Text">
                 <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:subject</xpath>
             </luc:field>
             <luc:field name="htmltitle" type="Text">
                 <namespace
prefix="xhtml">http://www.w3.org/1999/xhtml</namespace>
                 <xpath>/xhtml:html/xhtml:head/xhtml:title</xpath>
             </luc:field>
             <luc:field name="language" type="Text">
                 <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:language</xpath>
             </luc:field>
             <luc:field name="description" type="Text">
                 <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:description</xpath>
             </luc:field>
             <luc:field name="htmlbody" type="Text">
                 <namespace
prefix="xhtml">http://www.w3.org/1999/xhtml</namespace>
                 <xpath>/xhtml:html/xhtml:body</xpath>
             </luc:field>
             <luc:field name="contents" type="UnStored" xpath="/"/>
             </luc:document>
```

3. Create a batch file: *C:\apache-lenya-1.2.2\tools\bin\Index-{pub}.bat* With this:

```
             SET LENYAPUB={pub}
             SET CLASSPATH=.
             SET ANT_HOME=C:\apache-lenya-1.2.2\tools
             ant -f ../../build/lenya/webapp/lenya/bin/crawl_and_index.xml
-Dlucene.xconf=../../build/lenya/webapp/lenya/pubs/%LENYAPUB%/config/search/lucene-live.xc
            index
```

4. To create a logfile (and avoid some ant errors), create a file:

C:\apache-lenya-1.2.2\tools\bin\log4j.properties With this:

```
            log4j.rootLogger=INFO, lucene
            log4j.appender.lucene = org.apache.log4j.FileAppender
            log4j.appender.lucene.File = lucene.log
            log4j.appender.lucene.Append = false
            log4j.appender.lucene.layout = org.apache.log4j.PatternLayout
            log4j.appender.lucene.layout.ConversionPattern = %d{ABSOLUTE} [%t]
%-5p
          %-30.30c{2} %x - %m %n
```

5. Quit Lenya (to avoid file-locking issues). Run the batch file. Check the log.The index created works, but the results are not formatted properly.
6. "sitemap.xml" may appear.
7. All links are wrong. They have an extra slash and "/index_xx.xml" must be changed to ".html".
8. The excerpt is not available and displays a Java error.
9. These are fixed in the next section.

3. Fix the XML results to be usable.

Copy *C:\apache-lenya-1.2.2\build\lenya\webapp\lenya\xslt\search\sort.xsl* To: *C:\apache-lenya-1.2.2\build\lenya\webapp\lenya\pubs\{pub}\lenya\xslt\search\sort.xsl*. Copy *C:\apache-lenya-1.2.2\build\lenya\webapp\lenya\xslt\navigation\search.xsl* to *C:\apache-lenya-1.2.2\build\lenya\webapp\lenya\pubs\{pub}\lenya\xslt\navigation\search.xsl* After the other params, add this line:

```
<xsl:param name="chosenlanguage"/>
```

Add the usecase and language fields to the form tag:

```
            <form>
                <input type="hidden" name="lenya.usecase" value="search"/>
                <input type="hidden" name="language" value="{$chosenlanguage}"/>
                <input class="searchfield" type="text" name="query" alt="Search
field"/>
                <input class="searchsubmit" type="submit" value="Search"
name="find"/>
            </form>
```

Download (search-and-results.xsp) new file: *C:\apache-lenya-1.2.2\build\lenya\webapp\lenya\pubs\{pub}\lenya\content\search\search-and-results.xsp* Based on *C:\apache-lenya-1.2.2\build\lenya\webapp\lenya\content\search\search-and-results.xsp*

- Removed useless information. (I like dynamic lists better than anyone, but Search is a standard function with standard outputs, so why bother? I only left the <fields> tag to separate our output from lucene's.)
- Added language filter.
- Added protected section filter.
- Hardcoded ProtectedUrls. The default is to require visitors be in an "employee" Group to access "/live/employee". **Configure this for your website.**
- Uses Groups rather than Roles. (Roles are useless as long as "world" inherits "visit" for everything.)
- Fixed counters and total. (Total-hits changed from property to element of results.)
- Other bug fixes

Download (searchfixer.xsl) new file:

C:\apache-lenya-1.2.2\build\lenya\webapp\lenya\pubs\{pub}\lenya\xslt\search\searchfixer.xsl This file converts our poor output to something usable.

- Add languages to configuration
- Move total-hits from element to property of results.
- Choose "title" from "htmltitle" or Lucene's "title".
- Choose "excerpt" from "htmlbody", Lenya's "description", or Lucene's "excerpt"
- Transform URI from lucene's "uri" (/about/jobs/index_en.xml) to Lenya link (about/jobs_en.html).
- The default is to use htmlbody for the excerpt. **This file must be modified when using Custom Doctypes that do not have a /html/body.**

Download (usecase-search.xmap) new file:
C:\apache-lenya-1.2.2\build\lenya\webapp\lenya\pubs\{pub}\usecase-search.xmap

4. Blocking default search.

It is important to block the default search when implementing ProtectedAreas to prevent visitors from typing the URL of the default search (or, if this publication was in production, using a bookmark to the old search) and seeing links to protected documents.

- Create new file: *build\lenya\webapp\lenya\pubs\{pub}\lenya\lucene.xmap*
- Add content:

```
            <?xml version="1.0" encoding="UTF-8"?>
            <map:sitemap xmlns:map="http://apache.org/cocoon/sitemap/1.0">
            <map:pipelines>
                <map:pipeline>
                    <map:match pattern="*/search-*/lucene.xml">
                        <map:redirect-to
uri="/{page-envelope:publication-id}/live/index.html?lenya.usecase=search"/>
                    </map:match>
                    <map:match pattern="*/search-*/lucene*">
                        <map:redirect-to
uri="/{page-envelope:publication-id}/live/index.html?lenya.usecase=search"/>
                    </map:match>
                </map:pipeline>
            </map:pipelines>
            </map:sitemap>
```