Texas Center for Digital Knowledge
University of North Texas
http://www.txcdk.org

A Proof-of-Concept Repository for Learning Objects: Supporting the Reuse and Repurposing of Redesigned Courses and Their Content

DSpace Fulltext Indexing

Serhiy Polyakov
<sp0055@gmail.com >

August 28, 2007
Version 1.1

Version Control Specify the **Version**, **Date and Time of Modification** of the document, **Name of the Modifier**, **Section** of the document where the changed have been made, and **Brief Description of the Changes**.	
Document Title	DSpace Fulltext Indexing
Document Filename	Indexing_Fulltext_sp_09Aug2007.doc
Original Creation Date	August 08, 2007
Original Author	Serhiy Polyakov

Version	Date and Time of Modification	Name of Modifier	Section Modified	Brief Description of the Changes
1.1	August 28	Serhiy Polyakov	crontab	new
1.1	August 28	Serhiy Polyakov	types of searches	new

Table of Contents

DSpace Fulltext Indexing

Introduction

This document describes enabling full text indexing in DSpace.

Configure full text indexing

Media Filters are used to enable fulltext indexing in DSpace. Media Filters are classes used to generate derivative or alternative versions of master bitstreams. For example, the PDF Media Filter will extract textual content from PDF bitstreams, the JPEG Media Filter can create thumbnails from image bitstreams.

Media Filters are configured as a Sequence Plugin, with each filter also having a separate config item indicating which formats it can process. The following media filters are coming with default configuration of DSpace:

- HTMLFilter – extracts the full text of HTML documents for full text indexing.
- PDFFilter – extracts the full text of Adobe PDF documents (only if text-based or OCRed) for full text indexing
- WordFilter – extracts the full text of Microsoft Word or Plain Text documents for full text indexing
- JPEGFilter – creates thumbnail images of GIF, JPEG and PNG files
- BrandedPreviewJPEGFilter – creates a branded preview image for GIF, JPEG and PNG files (disabled by default)

There is no need to change this configuration for the purposed of this project.

Field search.maxfieldlength in dspace.cfg configuration file specifies the maximum number words to index for each document, and by default is set to the first 10,000 words. It can be modified or set to the value -1 to enable unlimited number of words.

Running full text indexing and configuring the script to run it regularly

In DSpace, “media filters” are what control both full-text indexing and automated creation of thumbnail images. Both can be scheduled by calling the filter-media script.

To run the filter-media script execute the following command:

```
[dspace]/bin/filter-media
```

Filter-media shell script may be configured to run regularly by adding a cron entry to the crontab for the user who installed DSpace. To set this up the following command need to be run as the dspace UNIX user:

```
crontab -e
```

Then the following lines need to be added (for example):

```
# Run the media filter at 03:00 every day
0 3 * * * ''[dspace]''/bin/filter-media
```

The above entry would schedule filter-media to run nightly at 3am.

Note: the file containing these entries is:

```
/var/spool/cron/dspace3
```

PostgreSQL also benefits from regular 'vacuuming', which optimizes the indices and clears out any deleted data.

Become the postgres UNIX user, run `crontab -e` and add (for example):

```
# Clean up the database nightly at 4.20am
20 4 * * * vacuumdb --analyze dspace > /dev/null 2>&1
```

Notes

- Lots of little changes that add up over time without a re-indexing can cause DSpace's search function to become erratic. This is another reason re-index DSpace regularly.
- Running filter-media will automatically update the DSpace search index for metadata elements meaning that there is no need to run `index-all` script in addition to `filter-media` script.
- There is no need to restart Tomcat web server after running `filter-media` script.

Search

In the basic search boxes, any terms entered are searched for anywhere within any of the search indices (i.e. any of the search.index.# fields in dspace.cfg), or the full text of the document (if it is full-text indexable).

So, full text searching in DSpace occurs when a user searches via the default search boxes on the home page of LOR, or when a user selects the "Keyword" option from the Advanced Search screen.

Note: In DSpace default configuration search by "keyword" performs search in `dc.subject.*` metadata field(s). However, In THECB project DSpace configuration, search by "keyword" is equivalent to basic search box search.

Types of searches

All searches are case insensitive. All terms are stemmed. Stop words are dropped.

Term search

A term is smallest index piece.
Example:
Search by *constitution* will retrieve all items that containing terms with the steam *constitut* in metadata or text.

Boolean search and grouping

The following operators can be used:
AND, OR, NOT, -, +

Verbose syntax	Shortcut syntax
a AND b	+a +b
a OR b	a b
a AND NOT b	+a -b

Implicit operator between terms is OR

Phrase and proximity search

An index contains positional information of terms. Phrase queries can be executed with slope factor - a number of edit distance needed to match the phrase.

For example, collection includes two documents containing a phrase "custom or political practice".

1) Searching by *"custom or political practice"* would retrieve these documents.
2) Searching by *custom practice* (no quotes) would retrieve these two and other documents.
3) Searching by *"custom practice"* would not retrieve these documents.
4) Searching by *"custom practice" ~1* would retrieve these two documents.
5) Searching by *"practice custom" ~3* would retrieve these two documents.

The slop is indicated after the tilde followed after search phrase.

Wildcard search

(this should be tested, does not work as expected)

Two standard wildcard characters are used:
* for zero or more characters
? for zero ore one character

First character of a wildcarded term may not be a *.

Prefix search (special case of Wildcard search)

Example:
Search by "con*" will retrieve all items containing terms starting with "con" in metadata or text.

Fuzzy search

This type of search matches terms similar to a specified term.
For example, *tree* is similar to *free*. Tilde after a search term enables fuzzy search.

Search by *tree~* will retrieve items containing term *free.*

Boosting queries
A cart (^) followed by a floating-point number sets the boosts factor for the preceding term

Search by *political^4 practice* will retrieve items in a different ranking order then search by *political practice* giving more weight to the term *political*.

References:

Configure full text indexing - DSpace Wiki <http://wiki.dspace.org/index.php/Configure_full_text_indexing>

dspace.org - DSpace System Documentation: Configuration and Customization: Media Filters <http://www.dspace.org/index.php?option=com_content&task=view&id=147#mediafilters>