Infrastructure - Task #2295

Story # 1386 (Closed): Generation of SOLR index for Mercury

Update fullText field in solr schema.xml

2012-02-07 22:08 - Skye Roseboom

Status:	Closed	**Start date:**	2012-02-07
Priority:	Normal	**Due date:**	
Assignee:	Skye Roseboom	**% Done:**	100%
Category:	d1_indexer	**Estimated time:**	0.00 hour
Target version:	Sprint-2012.07-Block.1.4		
Milestone:	CCI-1.0.0	**Story Points:**	
Product Version:	*		

Description

Need to ensure that "fullText" field with all text in the XML document (excluding element names).

copyField destination. replace 'text' field with fullText? Existing fullText used?

History

#1 - 2012-02-08 15:44 - Skye Roseboom

- Status changed from New to In Progress

#2 - 2012-02-08 17:00 - Skye Roseboom

Consider stripping //dataset/dataTable elements from fullText to avoid indexing large data sets.

#3 - 2012-02-08 22:38 - Skye Roseboom

Further clarification from mercury folk (Jim Green):

The fullText field is populated by building a string
which contains all of the contents.
The caveat here is that most of the fgdc files are relatively small,
and these do NOT contain actual data.

As for filtering, not much. Just some cleanup of artifacts from the
harvesting process. We remove all of the tags, and anything which is
part of a "<!\[CDATA\[" block.
This is all done with some low-level java, since we had performance
issues using any of the DOM libraries for manipulating the xml.

#4 - 2012-02-08 22:40 - Skye Roseboom

Since tags and element names are not part of the fullText field, Ive re-implemented this as a copyField that accumulates the text from science metadata fields. (rather than creating and maintaining a class to strip tags, CDATA blocks, and dataTables).

#5 - 2012-02-08 22:40 - Skye Roseboom

- Status changed from In Progress to Closed

#6 - 2012-02-08 22:41 - Skye Roseboom

Skye Roseboom wrote:

> Since tags and element names are not part of the fullText field, Ive re-implemented this as a copyField that accumulates the text from science metadata fields. (rather than creating and maintaining a class to strip tags, CDATA blocks, and dataTables).

Also it appears that mercury actually queries the 'text' field, not 'fullText'.

#7 - 2012-02-09 22:16 - Skye Roseboom

- Parent task changed from #2004 to #1386