Infrastructure - Bug #7881

index processor is failing on documents containing certain characters

2016-09-08 19:21 - Rob Nahf

Status:	Closed	**Start date:**	2016-09-08
Priority:	Normal	**Due date:**	
Assignee:	Rob Nahf	**% Done:**	100%
Category:	d1_indexer	**Estimated time:**	0.00 hour
Target version:	CCI-2.3.0		
Milestone:	None	**Story Points:**	
Product Version:	*		

Description

In cn-stage, there is a document that contains degree symbols in the abstract. The indexer can parse it, but produces an invalid solr Document that fails to parse.

The problem has to do with inconsistent encoding of lack there-of of these characters, resulting in this message:

20160830-18:58:28: [ERROR]: null:[com.ctc.wstx.exc.WstxLazyException] com.ctc.wstx.exc.WstxParsingException: Illegal character entity: expansion character (code 0xdbc0) not a valid XML character
at [row,col {unknown-source}]: [2,7233]
at com.ctc.wstx.exc.WstxLazyException.throwLazily(WstxLazyException.java:45)
at com.ctc.wstx.sr.StreamScanner.throwLazyError(StreamScanner.java:671)

The leading suspect is the use of StringEscapeUtilities.escapeXml() in org.dataone.cn.indexer.solrhttp.SolrElementField.serialize() method (d1_index_processor project). Which has been deprecated in later versions of org.apache.commons.lang. (org.apache.commons.lang3) in favor of escapeXml10() and escapeXml11() methods.

An example document is committed to the project under src/test/resources/org/dataone/cn/indexer/GriidcExample.xml, and a rough test started.

The offending character is in the gmd:supplementalInformation field, just after the text "bath guide.pdf "

Associated revisions

Revision 18296 - 2016-09-08 21:09 - Rob Nahf

fixes #7881: updated use of StringEscapeUtils from 2.6 to 3.4 (in commons-lang3 now), which completely rewrote the escapeXml implementations. New unit tests using the GRIIDC example demonstrate this change causes the test to pass. Note: StringUtils implementation is now also from commons-lang3 package.

History

#1 - 2016-09-08 21:11 - Rob Nahf

- Status changed from New to Closed

- % Done changed from 0 to 100

replaced use of StringEscapeUtilities.escapeXml() with newer version.

Note: The escapeXML method in the new version from commons-lang3 fixes the problem, but is also deprecated in favor of either escapeXml10() or escapeXml11(). I did not know what to choose, so I left it as is for now.