CN Index - Bug #6800

SOLR indexes malformed strings - identifier, id

2015-02-03 17:02 - Mark Servilla

Status:	Closed	**Start date:**	2015-02-03
Priority:	Normal	**Due date:**	
Assignee:	Skye Roseboom	**% Done:**	100%
Category:		**Estimated time:**	0.00 hour
Target version:			
Story Points:			

Description

During maintenance of mnTestLTER content on cn-sandbox-ucsb-1.test.dataone.org, it was found that results from a SOLR query returned records that contained malformed "identifier" and "id" strings. There are a total of 18 of such records.

Query:

curl -s -X GET "https://cn-sandbox-ucsb-1.test.dataone.org/cn/v1/query/solr/?q=datasource:urn\:node\:mnTestLTER"

List of malformed identifiers:

ld>:10.6073/pasta/e2c4d7746fc2dcda6bf62cba99389567

>oi:10.6073/pasta/18e14273378ba371de1833a41d32301d

otrends/2927/2/d5f755665edc9ckage/data/eml/ecotrends/2927/2/d5f755665edc967b5c9195366cbb6c8d

417/2.lternet.edu/package/report/eml/ecotrends/10417/2

rends/8468/2t.edu/package/report/eml/ecotrends/8468/2

85a238d13910aeaf02db6aage/data/eml/ecotrends/3920/2/ea7fd9dfab85a238d13910aeaf02db6a

bff04f425a9409e595b9ckage/data/eml/knb-lter-and/2722/6/924fa927309abff04f425a9409e595b9

071ce2bca9a8f73a007241b96f2240f1ml/ecotrends/1671/2/071ce2bca9a8f73a007241b96f2240f1

age/data/eml/.lternet.edu/package/data/eml/ecotrends/10568/2/a115d2a94070ca2085996903277db48a

34fd95f1581a290a6a35e1ege/data/eml/ecotrends/8459/2/5644b03fb34fd95f1581a290a6a35e1e

ld>ps://pasta.lternet.edu/package/data/eml/ecotrends/4473/2/f0e51aae0a51606233df892959d3a56d

otrends/8606/2/f0b80ea995d9fckage/data/eml/ecotrends/8606/2/f0b80ea995d9f6d585a2c9b2d99a7872

age/report/em.lternet.edu/package/report/eml/knb-lter-arc/10272/3

11085/2/18e3871ec8c06eb1420867cd2c6ata/eml/ecotrends/11085/2/18e3871ec8c06eb1420867cd2c6a0cb6

46a31afb6bfbeb683a12e938e/data/eml/ecotrends/6868/2/c2ca91ad46a31afb6bfbeb683a12e938

f3e297e1dd</pasta/a76c718448bff15dc8b004f3e297e1dd

211835310dbf305f731b9cce54b41bdeml/ecotrends/6584/2/e211835310dbf305f731b9cce54b41bd

d53e0930ernet.edu/package/data/eml/ecotrends/1586/2/c21358a57bffd9e176f634c1d53e0930

It is as if the string is being overwritten by left-over string-buffer content - for example, the malformed identifier "d53e0930ernet.edu/package/data/eml/ecotrends/1586/2/c21358a57bffd9e176f634c1d53e0930" contains the wrong sub-string "d53e0930" in position 0 through 15, thereby replacing the correct sub-string "https://pasta.lt". The same type of malformed strings is also found on cn-stage-ucsb-1.test.dataone.org.

The full output from the SOLR query may be found in the attached text document "cn-sandbox-ucsb-1-SOLR.txt".

Related issues:

Duplicated by ONE Mercury - Bug #6870: Fix handling of identifiers with url-e...	**Rejected**	**2015-02-28**

History

#1 - 2015-02-03 17:50 - Matthew Jones

The MDC project also found similarly malformed strings in the production SOLR index. Peter Slaughter is investigating.

#2 - 2015-02-03 17:51 - Skye Roseboom

- translation missing: en.field_release set to 2

#3 - 2015-03-10 19:14 - Rob Nahf

- Related to Bug #6870: Fix handling of identifiers with url-escaped characters added

#4 - 2015-04-03 21:08 - Lauren Walker

Here is an identifier that is experiencing this bug right now, on sandbox 2:

_NA_v2.0.xmlMAP_PRESENTVEG_C3Grass_RelaFrac_NA_v2.0.xml

#5 - 2015-04-07 16:38 - Skye Roseboom

- Status changed from New to In Progress

- % Done changed from 0 to 30

#6 - 2015-04-07 17:44 - Skye Roseboom

- Target version set to CCI-1.5.1

#7 - 2015-04-07 20:12 - Skye Roseboom

- Related to deleted (Bug #6870: Fix handling of identifiers with url-escaped characters)

#8 - 2015-04-07 20:12 - Skye Roseboom

- Duplicated by Bug #6870: Fix handling of identifiers with url-escaped characters added

#9 - 2015-05-13 16:34 - Skye Roseboom

- Target version deleted (CCI-1.5.1)

#10 - 2015-05-15 16:06 - Skye Roseboom

- Subject changed from SOLR indexes malformed identifier and id strings for mnTestLTER to SOLR indexes malformed strings - identifier, id

Problem effects all environments. It also effects more than the identifier and id fields. Problem has been observed in date fields, full-text field, origin, author, fileID.

Problem is not consistent - in that re-indexing a solr record wit mangled values creates a new solr record without any mangled values.

Debugging indicates that the index-processor is sending proper solr xml messages to the solr server, but the string values are being misinterpreted by the solr server. This is supported by output of the xml payload being generated at d1's index processor and the xml payload received at the solr server - both of which when logged appear to be proper - no mangled string values.

Problem persists and will take more work to discover cause.

Working on some scripts to help detect and clean up solr records and re-index documents which get mangled.

#11 - 2015-06-30 17:38 - Peter Slaughter

- File pidsInSolrNotInObjectStore.txt added

Ran a test on 6/26/2015 on cn-ucsb-1.dataone.org that fetched all pids from solr and then checked in object store (in this case guid field from 'systemmetadata' table). The attached
file 'pidsInSolrNotInObjectStore.txt' shows the pids that were in solr but not in the systemmetadata table.

#12 - 2015-07-31 19:09 - Matthew Jones

This problem also extends to the AuthoritativeMN field. Here's an example of it being mangled:

evolve whRYAD

which can be seen here:

https://cn.dataone.org/cn/v1/query/solr/?fl=identifier,authoritativeMN&q=id:*dryad.gp23s/1%3Fver*

#13 - 2015-10-29 17:06 - Skye Roseboom

Issue is a cause of running solr3.x on a Java7 JDK. Solr3.x line is no longer patched by solr community so issue will not be resolved until deployment of solr5 with the CCI V2.0

#14 - 2015-12-11 21:32 - Skye Roseboom

- % Done changed from 30 to 50

- Status changed from In Progress to Testing

New version of solr is being deployed to production and search index is rebuilding. No mangled strings have been detected in stage or production.

#15 - 2016-01-05 18:08 - Skye Roseboom

- Status changed from Testing to Closed

- % Done changed from 50 to 100

Solr 5 installed into production and index rebuilt. Problem appears to have gone away with solr5

Files

cn-sandbox-ucsb-1-SOLR.txt	19.4 KB	2015-02-03	Mark Servilla
pidsInSolrNotInObjectStore.txt	116 KB	2015-06-30	Peter Slaughter