Metacat - Bug #5599

absence of line feeds in eml causes pathQuery to not find some elements

05/07/2012 03:54 PM - gastil gastil

Status:	New	**Start date:**	05/07/2012
Priority:	Normal	**Due date:**	
Assignee:	ben leinfelder	**% Done:**	0%
Category:	metacat	**Estimated time:**	0.00 hour
Target version:	Unspecified	**Spent time:**	0.00 hour
Bugzilla-Id:	5599		

Description

Presence of line feeds seems to be needed for an eml doc to get loaded properly so pathQuery can find attributeList or attribute. Not just one line feed at the end.
We detected this on metacat 1.9.5 at metacat.lternet but tested it on metacat 2.0 (lava.lternet)

Evidence:
in lava.lternet.edu

knb-lter-kbs.10.19 has no line feeds at all in the document.

revision 20 is same as 19 except stmml-1.1 is spelled right.
revision 21 is same as 20 except it has one line feed at the end of the file.
(so revision 21 has one line)

revisions 19 thru 21, while they were the last revision, did not have their attributeList found by pathQuery.

revision 22, with 165 lines feeds DOES have its attributeList seen by pathQuery.

wc -l knb-lter-kbs.10.*
0 knb-lter-kbs.10.19.xml
0 knb-lter-kbs.10.20.xml
1 knb-lter-kbs.10.21.xml
165 knb-lter-kbs.10.22.xml

pathQuery result snippets from two separate queries (when two different revisions were the last revision):

<document>
<docid>knb-lter-kbs.10.22</docid>
<docname>eml</docname>
<doctype>eml://ecoinformatics.org/eml-2.1.0</doctype>
<createdate>2012-05-07</createdate>
<updatedate>2012-05-07</updatedate>
<param name="attributeList"></param>
<param name="@packageId">knb-lter-kbs.10.22</param>
</document>

older query:

<document>
<docid>knb-lter-kbs.10.21</docid>
<docname>eml</docname>
<doctype>eml://ecoinformatics.org/eml-2.1.0</doctype>
<createdate>2012-05-07</createdate>
<updatedate>2012-05-07</updatedate>
<param name="@packageId">knb-lter-kbs.10.22</param>
</document>

History

#1 - 05/08/2012 12:32 PM - gastil gastil

To diagnose this further:

(1)
I ran pathQuery for returnfield dataset/dataTable/attributeList/, that is, the whole xpath, not just the attributeList element name by itself.
Result:
none found
That is, same result whether attributeList is by itself or a more complete xpath.

(2)
I ran pathQuery for elements under attributeList such as
dataset/dataTable/attributeList/attribute/measurementScale/dateTime/formatString
Result:
Lots of KBS eml docs (the ones with no line feeds) DO return formatString under a path including attributeList, but zero returns for attributeList itself.

(3)
We are guessing maybe the bug relates to the word "attribute" because of its special meaning in xml. I would not think "attributeList" is a reserved keyword though.

#2 - 05/08/2012 12:33 PM - gastil gastil

to clarify, these pathQueries were for
<queryterm casesensitive="false" searchmode="starts-with">
<value>knb-lter-kbs</value>
all of which are known to currently have no line feeds
in metacat.lternet

(not lava)

#3 - 05/08/2012 05:14 PM - ben leinfelder

While this is indeed odd, my hunch is that we get a placeholder leaf node for the line feed that separates <attributeList> from it's first child <attribute> element. In query results you're getting of a quirky response with this blank returned as retrunfield data.

My suggestion - no matter what becomes of this bug - is to use a more definitive xpath to a true leaf node like "attributeList/attribute/attributeName" since this is a required element for any attributeList. This way you will be guaranteed positive/negative results no matter what the whitespace on the document looks like.

#4 - 03/27/2013 02:31 PM - Redmine Admin

Original Bugzilla ID was 5599