Metacat - Bug #5497

metacat accepts eml docs that fail knb parser

09/21/2011 10:05 PM - gastil gastil

Status:	Resolved	**Start date:**	09/21/2011
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:	5497		

Description

This is from nis track ticket #325https://trac.lternet.edu/trac/NIS/ticket/325
Im pasting the below in from that ticket.

docid knb-lter-sev.137.39511 is in the LTER metacat, recently harvested, and does not pass the knb parser. It has some empty elements.

Documents which do not pass the parser should not be harvested into metacat.

This is packageId knb-lter-sev.137.6933 last updated Sept 8, 2011.

I expect there are 43 eml documents in that harvest with this same problem. This one is just an example. knb-lter-sev.153.19842 is another example of the same thing.
Both are invalid with regard to the EML 2.1.0 schema, but have been successfully harvested into Metacat 1.9.3. How many other documents are invalid, but in Metacat?

In addition, a minor concern is that the HTML representation of this EML appears to use the packageId for generating the Metacat URI (http://metacat.lternet.edu:8080/knb/metacat/knb-lter-sev.153.5657/lter) -- the revision value of the packageId (knb-lter-sev.153.5657) is not the same as the revision value of the documentId (knb-lter-sev.153.19842). The generated URI is not correct and results in a Metacat "document not found" error when used in web browser query field.

This issue is primarily the domain of the KNB and the Metacat developers, not necessarily the LTER NIS.

History

#1 - 09/28/2011 07:39 PM - ben leinfelder

This sample EML document is using single quotes around the EML namespace (which should be valid) but Metacat was only looking for double quotes when it parses the namespace to determine what XSD to use to validate the incoming file.

knb-lter-sev.137.39511 has:
xmlns:eml='eml://ecoinformatics.org/eml-2.1.0'

The fix is committed to Metacat's SVN repository and will be included in the 1.10 Metacat release.
In the meantime, you can change how these documents are generated before harvesting so that they only use double quotes for the xmlns attributes.

Unfortunately, I'm not sure how to quickly locate the documents that were allowed to be saved to Metacat unless there's a common attribute in all of them that you can search on.

#2 - 03/27/2013 02:30 PM - Redmine Admin

Original Bugzilla ID was 5497