Metacat - Bug #5198

OAI-PMH: Data provider may fail to trigger reharvest of documents

10/11/2010 09:11 AM - Duane Costa

Status:	New	**Start date:**	10/11/2010
Priority:	Normal	**Due date:**	
Assignee:	Chad Berkley	**% Done:**	0%
Category:	metacat	**Estimated time:**	0.00 hour
Target version:	2.x.y	**Spent time:**	0.00 hour
Bugzilla-Id:	5198		

Description

On Oct 6, 2010, Marco Fahmi (hani.fahmi@qut.edu.au) wrote:

From my tests I found out that the metacat OAI PMH Provider stores something called "refreshDate" which is the last time the xml is generated and compares it to maxDateUpdated (date that the data package records table had last been modified). If the maxDateUpdated is newer than the refreshDate, OAI-PMH XML will be refreshed. I'm sure you can see the problem with this immediately but I'll just elaborate for our documentations about morpho and metacat.

Sample Scenario:

a. 1st October 10 AM - Harvester/User accesses OAI-PMH Provider URL and an XML is returned for the first time, refreshDate is set at 1st October
b. 1st October 11 AM - Researcher delete/update/add a data package through Morpho to metacat, and the maxDateUpdated = '1st October'
c. 1st October 12 AM - Harvester/user accesses the provider URL again and refreshDate is compared with maxDateUpdated. There is no indication of time, so there will be no refreshing of the OAI-PMH XML since the date is the same.
d. 2nd October 12 AM - Harvester/user accesses the provider URL again and refreshDate is compared with maxDateUpdated. There is still no difference between the dates and the bug is not resolved until someone decided to update on this day or future days or tomcat is refreshed.

Conclusion: Any updates on the same day with the refreshDate will not be reflected and may return the NullPointerException error in the case of data package deletion until a data package is modified/added"

"If you change the String comparison in MetacatCatalog.shouldRefeshCatalog() to be:

```
else if (maxDateUpdated.compareTo(refreshDate) >= 0) {
```

Then you will get fresh data when the refreshDate is equal to maxDateUpdated. Not an ideal fix, but should get you past that bug.

--

On October 11, 2010, Duane Costa and Mark Servilla wrote in a reply to Marco Fahmi:

The bug relating to the sample scenario you outline is a known limitation of the OAI-PMH data provider code. It originates from the fact that Metacat stores the 'date_created' and 'date_updated' fields as type 'date', rather than type 'timestamp', in the database table that stores metadata about EML documents. This makes it difficult for the data provider to know when to re-harvest based on anything more fine-grained than the current date. We're not certain that the change you suggest would actually resolve the issue: we think it would re-load the data catalog into memory, but it wouldn't necessarily trigger the re-harvest of a document. We agree that this is a significant limitation of the data provider code and we think further analysis will be needed before a good solution can be implemented.

History

#1 - 10/15/2010 12:41 PM - Duane Costa

On October 12, 2010, Marco Fahmi wrote:

By the way, I forgot to report another (little) bug: When you make an OAI-PMH call to an empty DB (i.e. has zero records), metacat returns a null pointer.

Marco Fahmi | Research Data Manager | Institute for Sustainable Resources | Queensland University of Technology
Address: Level 3, D Block, Gardens Point Campus, 2 George Street, Brisbane QLD 4000 |
Tel (M/T): +61 7 3138 1284 | Tel (W/Th/F): +61 7 3346 3141 | Email : fahmi@qut.edu.au | www.isr.qut.edu.au

#2 - 03/27/2013 02:29 PM - Redmine Admin

Original Bugzilla ID was 5198