Member Nodes - Task #3856

MNDeployment # 3552 (Closed): USGS CSAS

Re-harvest ORE documents from MN

2013-06-28 19:39 - Skye Roseboom

Status:	Closed	**Start date:**	2013-06-28
Priority:	High	**Due date:**	
Assignee:	Robert Waltz	**% Done:**	100%
Category:		**Estimated time:**	0.00 hour
Target version:			
Story Points:			

Description

The content of the ORE documents from USGS CSAS have been changed 'in-place'.

Need to determine how to re-harvest the content to the CN.

Potentially same solution as being investigated for re-harvesting content from ORNL DAAC.

Directly related to issue 3839.

This content would be immediately available through the CN REST API and search index, if new resource maps were generated with new pids/system metadata which obsolete the original/edited versions.

Related issues:

Related to Member Nodes - Task #3839: ORE documents contain references to non...	**Closed**	**2013-06-24**

History

#1 - 2013-07-01 17:48 - Chris Jones

- Assignee changed from Chris Jones to Robert Waltz

- Priority changed from Normal to High

- Status changed from New to In Progress

I'm assigning this to Robert for now, since he's working on the code for a one-time to update the CNs that will allow us to reharvest USGSCSAS content this week before the DUG meeting.

#2 - 2013-07-10 16:35 - Robert Waltz

- Status changed from In Progress to Closed

- translation missing: en.field_remaining_hours set to 0.0

#3 - 2013-07-10 19:05 - Ranjeet Devarakonda

- Estimated time set to 0.00

- Status changed from Closed to In Progress

It is showing only 5 records from the ONEMercury search.
https://cn.dataone.org/onemercury/send/facetsQuerry2?filterForDataHidden=true&term1=*&term1attribute=text&op1=&term3attribute=overlaps&term3=%2C%2C%2C&op3=&term8=collection&pageSize=10&start=0&sortattribute=default&facetattribute=datasource&facet=urn:node:USGSCSAS

#4 - 2013-07-11 14:58 - Skye Roseboom

Hi Robert,

Looking at the indexing output regarding these re-harvested pids -- I am seeing a lot of object path errors. In the index processing log it looks like this:

[INFO] 2013-07-11 14:45:11,573 (IndexTaskProcessor:isObjectPathReady:262) Object path exists for pid: resourceMap_doi_10.5066_F77H1GHV.xml however the file location: /var/metacat/data/autogen.2013042612461679446.1 does not exist. Marking not ready - task will be marked new and retried.

[INFO] 2013-07-11 14:45:11,606 (IndexTaskProcessor:isObjectPathReady:262) Object path exists for pid: resourceMap_doi_10.5066_F7WW7FN6.xml however the file location: /var/metacat/data/autogen.2013042612464924180.1 does not exist. Marking not ready - task will be marked new and retried.

[INFO] 2013-07-11 14:45:11,640 (IndexTaskProcessor:isObjectPathReady:262) Object path exists for pid: resourceMap_doi_10.5066_F7NZ85MB.xml however the file location: /var/metacat/data/autogen.2013042612465153383.1 does not exist. Marking not ready - task will be marked new and retried.

This indicates that the index processing process is attempting to read the contents of the ORE document off the local hard disk at the file path location indicated by the shared hazelcast data structure 'objectPath'. This is the structure in the storage cluster that maps PIDS to file system paths. Indexing does this in order to parse the contents of the ORE document - to derive the information contained by the ORE for the index record. Without a valid object path (file system path), indexing is unable to process the ORE documents. This is the reason these updated documents have not appeared updated in the index.

#5 - 2013-07-17 15:02 - Robert Waltz

- Status changed from In Progress to Closed

added logic in repair scripts to touch the hzObjectPath map when evicting pids.