Member Nodes - Task #3719

MNDeployment # 2564 (Deprecated): ORNL DAAC

Discrepancy in number of objects in DataONE versus DAAC

2013-04-19 18:26 - Matthew Jones

Status:	Closed	**Start date:**	2013-04-19
Priority:	High	**Due date:**	
Assignee:	Robert Waltz	**% Done:**	100%
Category:		**Estimated time:**	0.00 hour
Target version:	Operational		
Story Points:			

Description

Bob Cook writes: "I'm not sure of the reason, but the ORNL DAAC has only 942 data sets in DataONE, but we have metadata records in the DAAC archive for 1,029 data sets as of this week."

We need to resolve this discrepancy, and make it clear to MNs how they can resolve discrepancies like this that show up. Are these sync/validation issues? ORE issues? Other?

To close this ticket, please be sure to report how the issue was resolved here and to Bob Cook.

History

#1 - 2013-04-19 20:34 - Skye Roseboom

postgres -c "psql -d metacat -c \"Select count(*) from systemmetadata where origin_member_node = 'urn:node:ORNLDAAC' AND object_format='http://www.openarchives.org/ore/terms' ;\""
yields:

count

955
(1 row)

http://mercury-ops2.ornl.gov/ornldaac/mn/v1/object?formatId=http://www.openarchives.org/ore/terms
yields:

1011 results.

So it appears we need to synch the difference of 56 ORE docs.

Not sure how to reconcile the difference reported by the MN of 1011 with the 1029 reported. Perhaps the public object list request filtered some objects due to access policy.

#2 - 2013-04-19 20:50 - Skye Roseboom

This is not a synch issue. There are a set of ORE objects which have been assigned formatId type of 'octet stream' this causes DataONE to treat these objects as Data - not as a resource map/ORE. The following query reveals the documents by selected objects with a resourceMap like pid naming convention from ORNLDAAC but have been typed as 'octet-stream' instead of as an ORE document.

https://cn-ucsb-1.dataone.org/cn/v1/query/solr/?q=id:resourceMap*%20AND%20formatType:DATA%20AND%20datasource:urn\:node\:ORNLDAAC&fl=id&rows=2000

These documents need to have their formatId changed to ORE/resourceMap type: http://www.openarchives.org/ore/terms

This doesn't seem like an issue the CN are able to detect - it was simply mis-typed system metadata. Not sure there is a validation that would be able to determine a problem in this case. The records are on the CN and in the index .

#3 - 2013-04-23 19:20 - Skye Roseboom

Discussed resolution with Chris Jones and Ranjeet.

CN will update formatId manually to match what MN has.

MN will eventually want to update its copy of system metadata to reflect the new serial version (what is present on the CN).

#4 - 2013-04-24 20:13 - Skye Roseboom

- File ornl-daac-formatId-update.pids.txt added

- File ornl-daac-formatId-update.pids.txt added

- Assignee changed from Skye Roseboom to Chris Jones

Attaching listing of pids that represent resource maps (ORE) documents but do not currently have the ORE formatId: ' http://www.openarchives.org/ore/terms'. They currently have formatId: 'application/octet-stream' on the CN and ' http://www.openarchives.org/ore/terms' on the MN. We would like to update the formatId on the CN to match that of the MN.

Chris has prepared a script to effect an update on the CN so the formatId's are updated to match the desired formatId.

#5 - 2013-05-24 15:55 - Chris Jones

- translation missing: en.field_remaining_hours set to 0.0

- Status changed from New to Closed

I've run the update script on cn-ucsb-1.dataone.org that set the formatIds for the above identifier list to http://www.openarchives.org/ore/terms.

The ORNLDAAC MN reports:

curl -s -k - o -"http://mercury-ops2.ornl.gov/ornldaac/mn/v1/object?count=0&formatId=http://www.openarchives.org/ore/terms"

The CNs report 1544 science metadata documents through ONEMercury, so we need to investigate this further. We now have more on the CNs than is reported by the MN.

However, when d1_indexer tried to index the content, a number of these ORE documents were not on cn-ucsb-1, as reported in the log:

```
[ INFO] 2013-05-24 15:40:37,379 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_1120.xml is not available.
[ INFO] 2013-05-24 15:40:37,428 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_437.xml is not available.
[ INFO] 2013-05-24 15:40:37,565 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_427.xml is not available.
[ INFO] 2013-05-24 15:40:37,601 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_426.xml is not available.
[ INFO] 2013-05-24 15:40:37,652 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_1109.xml is not available.
[ INFO] 2013-05-24 15:40:37,694 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_1138.xml is not available.
[ INFO] 2013-05-24 15:40:37,731 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_409.xml is not available.
[ INFO] 2013-05-24 15:40:37,771 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_246.xml is not available.
[ INFO] 2013-05-24 15:40:37,877 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_1131.xml is not available.
[ INFO] 2013-05-24 15:40:37,911 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_1129.xml is not available.
[ INFO] 2013-05-24 15:40:38,018 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_524.xml is not available.
[ INFO] 2013-05-24 15:40:38,118 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_1135.xml is not available.
```

```
[ INFO] 2013-05-24 15:40:38,241 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_100566.xml is not available.
[ INFO] 2013-05-24 15:40:38,279 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_1140.xml is not available.
[ INFO] 2013-05-24 15:40:38,379 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_1139.xml is not available.
[ INFO] 2013-05-24 15:40:38,708 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_1107.xml is not available.
[ INFO] 2013-05-24 15:40:38,843 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_1130.xml is not available.
[ INFO] 2013-05-24 15:40:38,882 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_1108.xml is not available.
[ INFO] 2013-05-24 15:40:38,982 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_1097.xml is not available.
[ INFO] 2013-05-24 15:40:39,018 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_1134.xml is not available.
[ INFO] 2013-05-24 15:40:39,054 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_1128.xml is not available.
[ INFO] 2013-05-24 15:40:39,105 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_100582.xml is not available.
[ INFO] 2013-05-24 15:40:39,159 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_1115.xml is not available.
[ INFO] 2013-05-24 15:42:37,420 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_280.xml is not available.
[ INFO] 2013-05-24 15:42:37,453 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_1120.xml is not available.
[ INFO] 2013-05-24 15:42:37,486 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_437.xml is not available.
[ INFO] 2013-05-24 15:42:37,528 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_1127.xml is not available.
[ INFO] 2013-05-24 15:42:37,637 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_427.xml is not available.
[ INFO] 2013-05-24 15:42:37,673 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_426.xml is not available.
[ INFO] 2013-05-24 15:42:37,724 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_1109.xml is not available.
[ INFO] 2013-05-24 15:42:37,762 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_1138.xml is not available.
[ INFO] 2013-05-24 15:42:37,802 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_409.xml is not available.
[ INFO] 2013-05-24 15:42:37,835 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_246.xml is not available.
[ INFO] 2013-05-24 15:42:37,934 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_1131.xml is not available.
[ INFO] 2013-05-24 15:42:37,970 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_1129.xml is not available.
[ INFO] 2013-05-24 15:42:38,068 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_524.xml is not available.
[ INFO] 2013-05-24 15:42:38,167 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_1135.xml is not available.
[ INFO] 2013-05-24 15:42:38,267 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_100566.xml is not available.
[ INFO] 2013-05-24 15:42:38,303 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_1140.xml is not available.
[ INFO] 2013-05-24 15:42:38,401 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_1139.xml is not available.
[ INFO] 2013-05-24 15:42:38,434 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_1107.xml is not available.
[ INFO] 2013-05-24 15:42:38,534 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_1130.xml is not available.
[ INFO] 2013-05-24 15:42:38,569 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_100611.xml is not available.
[ INFO] 2013-05-24 15:42:38,603 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_1108.xml is not available.
[ INFO] 2013-05-24 15:42:38,721 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_1097.xml is not available.
[ INFO] 2013-05-24 15:42:38,754 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_1134.xml is not available.
[ INFO] 2013-05-24 15:42:38,789 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_1128.xml is not available.
[ INFO] 2013-05-24 15:42:38,827 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_100582.xml is not available.
[ INFO] 2013-05-24 15:42:38,864 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_1115.xml is not available.
[ INFO] 2013-05-24 15:42:38,902 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_1104.xml is not available.
[ INFO] 2013-05-24 15:42:38,939 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_1117.xml is not available.
[ INFO] 2013-05-24 15:42:39,041 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_1121.xml is not available.
[ INFO] 2013-05-24 15:42:39,141 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_1112.xml is not available.
[ INFO] 2013-05-24 15:42:39,177 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_1119.xml is not available.
[ INFO] 2013-05-24 15:42:39,280 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_235.xml is not available.
[ INFO] 2013-05-24 15:42:39,319 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_1098.xml is not available.
[ INFO] 2013-05-24 15:42:39,421 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_1137.xml is not available.
[ INFO] 2013-05-24 15:42:39,521 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_1118.xml is not available.
[ INFO] 2013-05-24 15:44:37,403 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_280.xml is not available.
[ INFO] 2013-05-24 15:44:37,438 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_1127.xml is not available.
[ INFO] 2013-05-24 15:44:37,509 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_100611.xml is not available.
[ INFO] 2013-05-24 15:44:37,545 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_1104.xml is not available.
[ INFO] 2013-05-24 15:44:37,583 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_1117.xml is not available.
[ INFO] 2013-05-24 15:44:37,712 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_1121.xml is not available.
[ INFO] 2013-05-24 15:44:37,812 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_1112.xml is not available.
[ INFO] 2013-05-24 15:44:37,850 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_1119.xml is not available.
[ INFO] 2013-05-24 15:44:37,948 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_235.xml is not available.
[ INFO] 2013-05-24 15:44:37,984 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_1098.xml is not available.
[ INFO] 2013-05-24 15:44:38,088 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_1137.xml is not available.
[ INFO] 2013-05-24 15:44:38,189 (IndexTaskProcessor:isObjectPathReady:251) Object path for pid: resourceMap_1118.xml is not available.
```

I'll look into this further with regard to the 3 CNs being in sync with this content (on disk).

#6 - 2013-05-24 16:17 - Chris Jones

- Estimated time set to 0.00

Although ONEMercury reports 1544 results (where the query= * AND (datasource :(urn:node:ORNLDAAC)))

a Solr query of:

https://cn.dataone.org/cn/v1/query/solr/?q=formatId:http\://www.openarchives.org/ore/terms%20AND%20datasource:urn\:node\:ORNLDAAC&rows=0&fl=identifier

returns a total of 955 still. Working on this. I've confirmed that the numbers are the same on all 3 CNs.

#7 - 2013-06-04 19:00 - Matthew Jones

- Status changed from Closed to In Progress

Reopening task -- it seems to have been closed prematurely. Skye indicates that sync has not picked up the formatid change.

#8 - 2013-06-24 22:42 - Chris Jones

- Assignee changed from Chris Jones to Robert Waltz

After discussing this issue in standup, we've decided that the ORNLDAAC content needs to be purged from the CNs and reharvested. I'm reassigning this to Robert since he's writing and testing a delete script that does a full purge. Once this has been completed, all ORNLDAAC content changes need to go through the standard API calls using MN.update(), or the Tier I equivalent on creating the correct obsoletes/obsoletedBy chain in the system metadata for the new and old versions.

#9 - 2013-07-10 17:51 - Robert Waltz

- Status changed from In Progress to Testing

ORNLDAAC pids from ORE documents have been reharvested and the counts are equivalent

#10 - 2013-07-10 17:51 - Robert Waltz

- Status changed from Testing to In Review

#11 - 2013-07-10 17:54 - Robert Waltz

- Status changed from In Review to Closed

Files

ornl-daac-formatId-update.pids.txt	1.14 KB	2013-04-24	Skye Roseboom