Infrastructure - Task #3612

CN SystemMetadata has incomplete obsoletedBy information

2013-02-26 20:25 - Chris Jones

Status:	New	**Start date:**	2013-02-27
Priority:	Normal	**Due date:**	
Assignee:	Ben Leinfelder	**% Done:**	0%
Category:	Environment.Production	**Estimated time:**	0.00 hour
Target version:	2013.6-Block.1.3		
Milestone:	CCI-1.1.2	**Story Points:**	
Product Version:	*		

Description

In a conversation over IRC, Matt pointed out that:

https://cn.dataone.org/cn/v1/meta/doi:10.6085/AA/MLPA_intertidal.51.2 @obsoletes
https://cn.dataone.org/cn/v1/meta/doi:10.6085/AA/MLPA_intertidal.51.1
but
https://cn.dataone.org/cn/v1/meta/doi:10.6085/AA/MLPA_intertidal.51.1 @obsoletedBy
https://cn.dataone.org/cn/v1/meta/doi:10.6085/AA/MLPA_intertidal.51.2

is not set in the system metadata on the CN. We need to understand why this is and address it. I'm assigning it to Ben for the moment since it may be related to

redmine.dataone.org/issues/2938 which seems to describe the same situation, but on MNs rather than CNs. Feel free to reassign, Ben, if this is off the mark.

Subtasks:

Task # 3620: Correct revision history on PISCO node	**Closed**
Task # 3621: Correct revision history on LTER node	**Closed**
Task # 3622: Correct revision history on KNB node	**Closed**
Task # 3623: Correct revision history on SANParks node	**Closed**

Related issues:

Related to Infrastructure - Task #3619: update d1_cn_repair tool to 1.1.0 rel...	**Closed**	**2013-02-27**
Related to Infrastructure - Task #2938: SystemMetadata has incomplete obsolet...	**Closed**	**2012-06-14**

History

#1 - 2013-02-27 17:56 - Ben Leinfelder

There is a CN.setObsoletedBy() method.
http://mule1.dataone.org/ArchitectureDocs-current/apis/CN_APIs.html#CNCore.setObsoletedBy() so I am hopeful that is will work...somehow.

After discussion on CCIT standup, seems like we can locate the incomplete chains on the CN and force reprocessing on them.

#2 - 2013-02-27 18:36 - Ben Leinfelder

There are many CN entries that should have a value for "SystemMetadata.obsoletedBy" (12,553 rows on cn-ucb-1, for instance):

```
-- these are old (obsoleted) entries that are not marked as such
select sm.guid, sm.obsoleted_by, sm.obsoletes, sm_by.guid as should_be_obsoleted_by, sm.authoritive_member_node
from systemmetadata sm, systemmetadata sm_by
where sm.guid = sm_by.obsoletes
and sm.obsoleted_by is null;
```

Spot-checking these on PISCO shows that the PISCO MN does not reflect the revision chain bidirectionally:

https://data.piscoweb.org/catalog/d1/mn/v1/meta/doi:10.6085/AA/LB15XX_015MXTI001R00_20030715.40.1

should be obsoletedBy:

https://data.piscoweb.org/catalog/d1/mn/v1/meta/doi:10.6085/AA/LB15XX_015MXTI001R00_20030715.40.2

but is not.

The LTER MN seems to have its chains correctly represented bidirectionally even if they are not fully represented on the CN.

#3 - 2013-02-27 19:04 - Ben Leinfelder

Only a small number (4 rows) of the opposite case where "obsoletes" is missing:

```
-- these are ones that should be marked as newer revisions
select sm.guid, sm.obsoleted_by, sm.obsoletes, sm_s.guid as should_obsolete
from systemmetadata sm, systemmetadata sm_s
where sm.guid = sm_s.obsoleted_by
and sm.obsoletes is null;
```

These appear to be cases where the revision history started on the KNB but then continued on SANParks and were replicated to KNB so that the most recent version is really on SANParks and KNB is not authoritative for it.

#4 - 2013-02-27 19:08 - Ben Leinfelder

Here is a summary of missing obsoletedBy, grouped by MN

```
select sm.authoritive_member_node, count(sm.*)
from systemmetadata sm, systemmetadata sm_by
where sm.guid = sm_by.obsoletes
and sm.obsoleted_by is null
group by sm.authoritive_member_node;
```

#5 - 2013-02-27 19:16 - Ben Leinfelder

The KNB MN has 18 records that do not correctly have their revision chain set.
SANParks has 17 records.

These can be corrected manually on the MN if the CN will pick up those changes. I do not have access to PISCO or LTER to manually update the revision chains on those MNs. Would have to involve Mike Frenock and Mark Servilla, respectively.

#6 - 2013-02-27 19:21 - Robert Waltz

We should be able to inject the pids as synchronization tasks directly into d1_processing for updates needed for those pids missing obsoletedBy.

We can use the d1_cn_repair tool for fixing these entries.
The d1_cn_repair tool will need a file in the following format:
MemberNode Id followed by a space followed by a pid on a single line, and repeated for the number of synchronization tasks to reprocess

so something like:
urn:node:PISCO doi:10.6085/AA/MLPA_intertidal.51.1
urn:node:PISCO doi:10.6085/x/yz

etc

would work as a file to pass in as a parameter.

On a related note, I have not used d1_cn_repair tool since 1.0.2 of the CN stack and so it will need to be updated.

using the d1_cn_repair tools is my preferred first method of fixing the missing obsoletedBy tasks since it mimics what the proposed audit tool would do in this situation

#7 - 2013-03-02 01:35 - Ben Leinfelder

Robert ran cn-repair-tool for the KNB pids that had unidirectional obsolescence chains. It sounds like there are still some differences between each CN:

there were 46 reported from cn-unm-1; now there are two reported from cn-unm-1

getSystemMetadata on kgordon.31.31, it appears to have obsoleteBy set, but not in postgres table (cn-unm-1)

kgordon.31.31 is fixed on cn-ucsb-1

resourceMap_will.14.1 has an error: "Update sysMeta Not Unique! Checksum is different"

#8 - 2013-03-02 06:00 - Robert Waltz

root@cn-unm-1:~# su postgres -c "psql -d metacat -c \"select sm.authoritive_member_node, count(*) \

from systemmetadata sm, systemmetadata sm_by \
where sm.guid = sm_by.obsoletes \
and sm.obsoleted_by is null group by sm.authoritive_member_node \""
authoritive_member_node | count
---
urn:node:SANPARKS | 115
urn:node:LTER | 1304
urn:node:KNB | 46
urn:node:PISCO | 11274

Ran cn_repair job on cn-unm-1 for 46 KNB pids

Results

```
root@cn-unm-1:~/FixObsoleteChains# su postgres -c "psql -d metacat  -c \"select sm.authoritive_member_node, sm.guid \
from systemmetadata sm, systemmetadata sm_by \
where sm.guid = sm_by.obsoletes \
and sm.obsoleted_by is null and sm.authoritive_member_node like 'urn:node:KNB'\""
authoritive_member_node |         guid

-------------------------+------------------------
urn:node:KNB            | resourceMap_will.14.1
urn:node:KNB            | kgordon.31.31
```

```
root@cn-ucsb-1:/var/log/dataone/synchronize# su postgres -c "psql -d metacat  -c \"select sm.authoritive_member_node, sm.guid \
from systemmetadata sm, systemmetadata sm_by \
where sm.guid = sm_by.obsoletes \
and sm.obsoleted_by is null and sm.authoritive_member_node like 'urn:node:KNB'\""
authoritive_member_node |         guid

-------------------------+------------------------
urn:node:KNB            | resourceMap_will.14.1
urn:node:KNB            | kgordon.14.83
urn:node:KNB            | datastar.14.28
urn:node:KNB            | kgordon.17.32
urn:node:KNB            | df35d.3.1
urn:node:KNB            | KWG.15.1
urn:node:KNB            | cbfs.34.3
urn:node:KNB            | datastar.11.18
```

resourceMap_will.14.1 fails with sysMeta Not Unique! Checksum is different.

#9 - 2013-03-08 06:02 - Robert Waltz

after PISCO updated their systemMetadata, there were still inconsistencies on the CNs.

On UNM the count of incomplete obsoletedBy was 796
On UCSB the count of incomplete obsoletedBy was 614

I ran the cn_repair tool on the intersection of pids between UNM and UCSB totalling 553

Afterwards the inconsistencies were as follows:

On UNM, the count of incomplete obsoletedBy was 250
On UCSB the count of incomplete obsoletedBy was 68

It appears that 7 failed while 546 succeeded.

#10 - 2013-03-08 17:43 - Robert Waltz

After LTER ran the update scripts on the MN, there were still 1304 inconsistencies on the CNs, both UNM and UCSB. A diff of the pids found no differences. So, UNM and UCSB were consistent with their inconsistencies...

I ran the cn_repair tool on the intersection of pids on UNM totalling 1304

Afterwards, there are still 75 inconsistencies. The errors are due to Checksums being different on the CN and MN.