Member Nodes - Bug #7896

Dryad data sets don't mark new versions correctly in SystemMetadata with obsoletes/obsoletedBy

2016-09-30 18:38 - Matthew Jones

Status:	In Progress	**Start date:**	2016-09-30
Priority:	Normal	**Due date:**	
Assignee:	Monica Ihli	**% Done:**	30%
Category:		**Estimated time:**	0.00 hour
Target version:			
Story Points:			

Description

Dryad currently has 74,299 metadata dcouments reported in its DataONE profile, while on their site they indicate having only 14,217 data packages (http://datadryad.org). I believe this is because Dryad registers each new version of their metadata and data files with a new timestamped PID, but has not provided the appropriate 'obsoletes' information in their system metadata to indicate that the newer PID replaces an original PID. Thus, these newer versions are being counted in DataONE as independent data sets, when in fact they are just revisions. For a concrete example, see:

Original metadata:
https://cn.dataone.org/cn/v2/meta/http://dx.doi.org/10.5061/dryad.0t407/1%3Fver=2016-09-30T10:03:37.256-04:00
New version: https://cn.dataone.org/cn/v2/meta/http://dx.doi.org/10.5061/dryad.0t407/1%3Fver=2016-09-30T10:08:08.080-04:00

That new version is not marked as obsoleting the original, and so they both show up as independent data sets. Ideally, the obsoletes field would be set for the newer of the two. Dryad would also probably benefit from using the SID field in system metadata to indicate that the two versions are part of the same series with the given DOI (e.g., in this case, http://dx.doi.org/10.5061/dryad.0t407/1). This lack of obsoletes fields should be easily fixed because the PIDs in Dryad seem to consistently use the underlying DOI with a timestamped version field, so a script could probably be written to update all of the system metadata using a simple timestamp comparison to determine version order for the PIDs.

I believe the same issues exist for resource maps and data files as well, although possibly to a lesser extent for the data files.

Related issues:

Related to Member Nodes - MNDeployment #3118: Dryad Member Node	**Operational**	**2012-08-05**

History

#1 - 2016-09-30 22:32 - Matthew Jones

- Related to MNDeployment #3118: Dryad Member Node added

#2 - 2017-02-13 19:28 - Monica Ihli

- Status changed from New to In Progress

- % Done changed from 0 to 30

I am in the process of writing a program to compare the contents of system metadata as it appears on the MN to the system metadata on the CN, and create a report of discrepancies which will be provided to the MN operators.

We still face the issue of figuring out how to correct discrepancies once they are identified. One solution to consider is upgrading Dryad from v1 to v2, in order to give them the power for MN changes to system metadata to cascade to CN. This possibility should be discussed with MN.

#3 - 2017-02-13 19:53 - Monica Ihli

- Assignee changed from Ryan Scherle to Monica Ihli

#4 - 2017-07-24 06:45 - Matthew Jones

Update: According to our summary page, Dryad now has 101,210 non-obsoleted metadata documents listed in our system, while there self-reported data sets have only grown to 17861 as reported on http://datadryad.org. We are significantly inflating our data set counts by not properly marking

versions in Dryad, and the gap is growing. Any thoughts on when this might get resolved?