Member Nodes - Task #3229

MNDeployment # 3228 (Operational): NCEI - National Centers for Environmental Information

contact NODC to continue discussions about an NODC MN

2012-09-06 21:51 - Matthew Jones

Status:	Closed	**Start date:**	2012-09-06
Priority:	Normal	**Due date:**	
Assignee:	Matthew Jones	**% Done:**	100%
Category:		**Estimated time:**	0.00 hour
Target version:	Operational		
Story Points:			

Description

Matt Jones discussed becoming a Member Node with Ken Casey of the NODC at the summer 2012 ESIP meeting. Conversations are continuing over email.

History

#1 - 2012-09-17 17:09 - Matthew Jones

We are planning a conference call for Oct 5th with Matt Jones (DataONE), Ken Casey (NODC), and Christine White (ESRI) at a minimum.

#2 - 2012-10-05 19:34 - Matthew Jones

Matt, Dave, and Amber had discussions with NODC and ESRI. There is strong interest in NODC becoming a Member Node, and they will be looking at two paths -- one is to adapt GeoServer Portal directly to be a Tier 1 MN, the other is to use GMN to act as an adapter proxy in front of their services. Next call scheduled for November 9 to touch base on progress. Notes from the meeting are below. Timing on implementation is still unclear.

Matt

DataONE/NODC/ESRI Conference Call (http://epad.dataone.org/dataone-nodc-conf-20121005)

DataONE:
Matt Jones: Co-Lead, Core Cyberinfrastructure Team & Leadership Team
Amber Budden: Assitant Director for Community Engagement
Dave Vleglais: Assistant Director for Development and Operations

NODC:
Ken Casey, Technical Director
Yuanjie Li, GIS expert, Technical Development Team Member
John Relph, Technical Development Team Lead

ESRI:
Christine White, Project Manager/Geoportal Product Engineer
Teddy Matinde, Developer & engineer, Esri Geoportal Server
Urban MacGillvray, Development Lead, Esri Geoportal Server

Agenda & Notes

tier infor in box here: http://www.dataone.org/sites/all/documents/DataONEMNFACTSheetFormattedDec1.pdf

A. Plans for NODC becoming a Member Node
-- On the NODC priorities to connect with efforts like DataONE and similar initiatives
-- Not yet clear on what milestones might be needed to make this happen
-- Needs a sense of level of effort and complexity of technical work required
-- Will be deciding on FY13 deliverables over the next month
-- may be able to make progress even if it doesn't make it to the deliverables list

-- Teddy from ESRI has looked over the DataONE API
-- seems feasible for Geoportal to connect to DataONE
-- more difficult for DataONE to connect to Geoportal unless we use an open standard like CSW

-- Ken: would target Tier 1 for now
-- can they host some other people's data as well (mix in Tier 4 as well)

-- CSW:
-- if we are only talking Tier 1, then the level of entry is lower
-- http://mule1.dataone.org/ArchitectureDocs-current/apis/MN_APIs.html
-- CSW can be used to federate other CSW servers

-- Benefits to DataONE to expose CSW as an endpoint
-- would this be done at Coordinating Node or Member Node levels?
-- suggestion that CSW be implemented at the DataONE Coordinating Node level

-- Searching
-- http://mule1.dataone.org/ArchitectureDocs-current/design/SearchMetadata.html

-- Logging
-- http://mule1.dataone.org/ArchitectureDocs-current/design/LoggingSchema.html

-- Discussion of identification, versioning, and mutability issues

-Ken - seems we would have some work to do to support the logging API call, but that there is nothing in the logging requirements that would be a show stopper for us (since we have privacy guidelines and only have limited statistics that are kept... basic, standard web log stuff like URL accessed, accessing IP address, volumes, and numbers of files).

- Ken seems also that we should probably be able to meet the versioning requirements.
- Ken : rough idea for way forward would be to look at using the Generic Member Node software to map some of the API calls to Geoportal calls, and then develop more specific services for the things not handled by the Geoportal.

B. Connection points between DataONE and Geoportal Server

C. Next steps

-- Plan for a monthly call to touch base on progress
-- Unclear currently what the timing is for implementation
-- Ken's team will explore use of GMN as an adapter to front their systems, Christine's team will explore adapting GeoPortal Server
-- Matt: would like to know how DataONE team members can help, and what time period that might represent

-- Next call: November 9, 2012 11:30 - 12:30 PDT

#3 - 2013-01-30 20:14 - Matthew Jones

- Target version set to Deploy by end of Y5Q2

#4 - 2013-08-01 23:10 - John Cobb

- Target version changed from Deploy by end of Y5Q2 to Deferred

#5 - 2014-05-08 21:01 - Matthew Jones

Matt Jones, Ken Casey, and John Relph met today (May 8, 2014) to discuss possible partnerships between NODC and DataONE. In summary, we found that the aims of DataONE and NODC are aligned, and there is interest in working together to both expose NODC data via DataONE over the short term, and over a (possibly much) longer period look into providing Tier 3 access for writing data and Tier 4 access for replicating data to NODC.

There are numerous logistical constraints that prevent a lot of time investment in this at the moment by NODC. The biggest item, which is also relevant to DataONE, is that Congress is likely to administratively merge (via the budget process) the three national NOAA data centers (NODC, NGDC, and NCDC). While this merger will play out over the next few years, and while none of the individual data centers will disappear, it does mean a significant rearrangement of services and staff is imminent. But it also provides an opportunity, as the types of interoperable interfaces that DataONE provides very well could make a merger of systems and services across these centers less painful. Ken will know more when Congress passes a budget.

So, even with that, NODC still wants to make data holdings available via DataONE, but won't have much of an opportunity to work on it. I offered that NCEAS staff on our NOAA-funded project could try to help bootstrap a Tier 1 activity, and they were supportive. A distant second priority was making writing data possible at Tier 3, so that would have to be revisited at a later time.

Past discussions revealed some potential barriers to integration, but in revisiting them today I got the sense that none would be particularly hard to overcome. In particular:

- immutability
  ** NODC assigns an accession # for packages of data on ingest, and internally tracks the revisions of package components carefully, so they have all of the information needed to conform to the DataONE immutability requirements
  ** the NODC accession # corresponds to the DataONE serial identifier (SID), and their internal revision number provides the ability to create the equivalent of DataONE's persistent identifier (PID) by concatenating the accession # and the revision
  ** however, NODC doesn't generally expose the accession # in user facing services, and instead provides various access services that hide the individually accessioned data packages (for example, showing a virtual filesystem broken down by year via THREDDS)
  ** we discussed the possibility of deploying a new GeoPortal that directly exposes accessioned data packages, but it has complications

- logging
  ** NODC has apache logs, but there isn't an obvious link between those log entries and the accession numbers that they use for data packages. They produce aggregate access statistics on a periodic basis
  ** We could consider helping build an apache log parser that would extract logEntry records from apache log urls for use in reporting usage back to DataONE.

  ** In addition, because they are talking about being a Tier 1 node, its not as critical that their log reporting be complete, as they would only be underrepresenting their own data usage.
  ** Ken proposed a staged approach in which log parsing was used to generate records for the more obvious access applications at first, and postpone the more complex mapping to later.

- harvesting
  ** NODC does not have the resources this year to invest in building out the DataONE listObjects() and other REST services needed to be a Tier 1 node
  ** NODC exposes their data through several open protocols that DataONE could harvest today (OAI*PMH, CSW, and OpenDAP)
  ** There would be some issues with mapping those harvest services to the accession # as well
  ** There are issues with ensuring that data linkages in harvested metadata stay current, especially as the data centers merge

Next steps for moving forward:
* NODC has interest in supporting our harvest of their holdings via existing published services.
* The contact point to support such an effort would be: Yuanjie Li Yuanjie.Li@noaa.gov
* Matt will discuss these findings with Dave and others and will come up with a plan to pursue one or more of the following:
** harvest pmh, map to serialid, and ingest into DataONE
** Once Gallagher finishes a new version of Hyrax supporitng DataONE, NODC could look into deploying that
*** NODC could act as a testbed for the DAP/Hyrax member node work
** harvest csw, map to serialid, and ingest into DataONE

Plan to meet at ESIP in July (@DUG?) to demo and discuss progress, and chart a further course forward.

#6 - 2014-05-19 12:33 - Bruce Wilson

- Target version deleted (Deferred)

#7 - 2015-01-02 17:04 - Laura Moyers

- Target version set to In Progress

#8 - 2015-07-09 20:53 - Laura Moyers

- Status changed from In Progress to Closed

- % Done changed from 0 to 100

- translation missing: en.field_remaining_hours set to 0.0

Contact has been made and maintained; we are in the development phase as of 7/9/15. Can close this ticket as further development activities will be tracked in their own tasks.