Member Nodes - MNDeployment #7895

Pangaea

2016-09-26 20:18 - Matthew Jones

Status:	Operational	**Start date:**	2017-12-06
Priority:	Normal	**Due date:**	
Assignee:	Monica Ihli	**% Done:**	58%
Category:		**Estimated time:**	0.00 hour
Target version:	Operational		
Latitude:	53.08	**MN_Date_Online:**	2018-03-20
Longitude:	8.80	**Name:**	PANGAEA Data Publisher
MN Description:	PANGAEA is an open access data publisher for Earth & environmental science supporting the long-term archiving and publication of georeferenced data related to environmental sciences	**Logo URL:**	https://github.com/DataONEorg/member-node-info/blob/master/production/graphics/web/PANGAEA.png
Base URL:	Prod: pangaea-orc-1.dataone.org/mn / Stage: pangaea-dev-orc-1.test.dataone.org	**Date Upcoming:**	2018-02-23
NodeIdentifier:	urn:node:PANGAEA	**Date Deprecated:**	
MN Tier:	Tier 1	**Information URL:**	https://www.pangaea.de/
Software stack:	Slender Node (GMN)	**Version:**	2.4.0

Description

At the 2016 RDA plenary in Denver, I made contact with Markus Stocker markus.stocker@gmail.com who is on the staff at the Pangaea (https://pangaea.de/) repository in Germany about the possibility of bringing Pangaea into DataONE as a MN. Markus was enthusiastic, and said he would bring it to the attention of the leads at Pangaea, namely Michael Diepenbroek mdiepenbroek@pangaea.de and Uwe Schindler uschindler@pangaea.de. He has since replied, saying that there is support for doing so but minimal resources. He suggested that they would join if we 'do the work', which given their existing web service interfaces, probably means some form of Slender Node. They have ElasticSearch, PMH, and other endpoints accessible, as well as data subsetting services. Their metadata format largely appears to be a custom Pangaea schema. Their web services are listed and accessible here: http://ws.pangaea.de/

Markus' response follows:

Hi Matt,

Was great meeting and talking to you at RDA.

As promised, I have raised the idea of a PANGAEA - DataONE integration in internal discussions. There is agreement here that this is something interesting to explore.

We will need to discuss the details but given the little free capacity on our side, I am afraid any integration would have to be "easy on us." As a start, the easiest for us would be if DataONE integrates via our suite of web services [1]. If you take a look at those services, is it possible for you to gauge whether this is a reasonable starting point, and can we draft a plan detailing how an integration could work based on those services?

Cheers, m.

We need to discuss and send a reply.

Subtasks:

Story # 8225: Customize Indexing & View for gmd-pangaea	**In Progress**
Story # 8260: (Pangaea) Move to Production	**Closed**
Task # 8261: Mutual acceptance	**Closed**
Task # 8262: Register in Production	**Closed**
Task # 8263: Formal announcement	**Closed**

Task # 8508: Implement in Production Environment			**Closed**
Bug # 8471: Apache crashed			**Closed**
Task # 8673: Pangaea- GMN upgrade from 2.4.0 to 3.0.4			**New**
Related issues:			
Related to Member Nodes - Task #8137: Pangaea - Adapter Configured in Test	**Closed**		**2017-07-18**
Related to Member Nodes - Task #8138: Pangaea - Register in sandbox	**Closed**		**2017-07-21**
Related to Search UI - Story #8574: PANGAEA Temporary Fix: SID only in Data C...	**New**	**2018-04-30**	

History

#1 - 2016-09-26 20:39 - Matthew Jones

And my response to Markus:

Hi Marcus,

It was great to meet you, and thanks for the information. At DataONE, we are very excited about partnering with you, not least because your topical emphasis is extremely complementary to ours. Like you, however, we have to figure out how we might be able to work with you in terms of developer time. One possibility might be for us to use a 'harvest only' strategy for Pangaea to become a Tier 1 member node, which would be a fairly simple deployment for you. That loses the ability for users to contribute data to Pangaea via DataONE APIs, but that could always be added later when it makes sense. Our member team is looking over the materials you sent, and we'll get back to you with some possibilities. Is this the right list of contacts to include in these discussions? Thanks,

Matt

#2 - 2017-01-31 21:32 - Laura Moyers

Laura to contact Pangaea week of 30 January to set up an exploratory meeting.

#3 - 2017-02-20 23:57 - Laura Moyers

Meeting 20 Feb 2017 at 11am ET. https://epad.dataone.org/pad/p/Pangaea_and_DataONE

DataONE to develop an implementation plan; Pangaea to send examples of metadata and other points of interest. Meet again around 13 March date/time TBD.

#4 - 2017-02-27 18:49 - Laura Moyers

- Status changed from New to Planning

- Target version set to Deploy by end of Y3Q4

- Longitude set to 8.80

- Latitude set to 53.08

- % Done changed from 0 to 10

See https://epad.dataone.org/pad/p/Pangaea_and_DataONE for notes of first meeting 20 February 2017. Next meeting scheduled for 13 March 2017.

#5 - 2017-03-01 23:31 - Laura Moyers

- MN Tier set to Tier 1

- NodeIdentifier set to urn:node:PANGAEA

- Software stack set to Slender Node (GMN)

#6 - 2017-04-11 17:03 - Laura Moyers

- Base URL set to https://www.pangaea.de/

#7 - 2017-06-25 13:49 - Laura Moyers

Meeting 9 June 2017 (https://epad.dataone.org/pad/p/Pangaea_and_DataONE). Monica demo'd parts of the OAI-PMH harvester. Some questions arose which have been addressed, including (below from email from Monica to Pangaea team 17 June):

• Identifiers - The adapter script now strips the oai-pmh prefix from identifiers as they are loaded into the generic Member Node.

• The modification of record updating logic has been implemented. Previously, if a Pangaea identifier showed up in a harvest, and that identifier already existed in the member node data store, then an .update() upon the record would automatically be processed. What has changed is that now the datestamp of the record from the OAI-PMH harvest is compared to the dateUploaded in the Member Node. An update will only be processed if the date has changed. This means that, should some problem be encountered during a harvest, there will be no consequences from simply starting over. Records will not be needlessly updated/versioned.

• The current version of the adapter script has been installed on one of my test installations of the generic member node software. This particular test installation is not registered to a Coordinating Node. It's just a "stand-alone" test instance for monitoring the activity between the Pangaea service provider, the adapter, and the generic Member Node software.

• The adapter is currently scheduled to run every 2 hours. When this is implemented live for Pangaea, the frequency will probably be much less, but the idea will be the same: use cron to run it periodically. I will leave this running over the next week to keep an eye on things. I've setup rsync to copy over the harvest log and the error log to a subdirectory in the site's document root so you can monitor this too. What I'll be watching out for is any actual processing errors that might be encountered on a record by record basis.

o Harvesting Log:
https://centos7-3gmn.kitty.ninja/pangaea/OAI-PMH_harvest.log
o Error Log:
https://centos7-3gmn.kitty.ninja/pangaea/adapter-errors.log

• Here are some endpoints for the test installation for Pangaea content on my server that you will find helpful:
Description End Point
You can see the current number of objects on this member node installation here: https://centos7-3gmn.kitty.ninja/mn/v2/object?start=0&count=0

You can browse through lists of abbreviated system metadata for harvested records with this end point. Just adjust the start and count parameters to suit you: https://centos7-3gmn.kitty.ninja/mn/v2/object?start=0&count=50

You can grab the identifier field from any of the records in the list and use this endpoint to see the complete system metadata record:
https://centos7-3gmn.kitty.ninja/mn/v2/meta/doi:10.1594/PANGAEA.51463_20170617_0026

The science metadata record for an object which was harvested from Pangaea is accessible through this endpoint:
https://centos7-3gmn.kitty.ninja/mn/v2/meta/doi:10.1594/PANGAEA.51463_20170617_0026

At this point we are basically ready to move forward with setting up Pangaea. We understand that Pangaea desires some alterations to how their science metadata is presented to end users through the search interface. However, we can still make progress in setting up the member node software and adapter on a Pangaea controlled server. The process must be operational between Pangaea's service provider and the member node software anyways, before the node can be registered to a Coordinating Node in a testing environment.

It would be good for DataONE to know where Pangaea stands on getting an installation of the Generic Member Node active. The instructions for installation are pretty thorough, but there are a lot of steps. So I would like to volunteer to be present with someone from Pangaea during the installation process. That way you can share your screen, and I can provide real-time support as needed.

Next steps: feedback from Pangaea about GMN installation and assist if necessary (Monica), define desired changes to metadata view in search interface (Laura)

#8 - 2017-08-22 19:25 - Laura Moyers

- Assignee changed from Laura Moyers to Monica Ihli

- Subject changed from Member | Pangaea to Pangaea

Monica has been running the harvester successfully in Sandbox for some time, but there have been roadblocks in testing of the complete MN in the Stage environment. As of 21 August, there are 844 "packages" discoverable via DataONE search in Stage.

There seem to be issues with the metadata as presented by Pangaea (not passing validation? Monica would know). A relatively low percentage of the metadata objects are synchronizing with the CNs, probably due to validation errors.

We have not received the MNDD from Pangaea yet (last contact re: MNDD 3/13/17, then 8/22/17)

#9 - 2017-08-22 19:53 - Laura Moyers

- Related to Task #8137: Pangaea - Adapter Configured in Test added

#10 - 2017-08-22 19:53 - Laura Moyers

- Related to Task #8138: Pangaea - Register in sandbox added

#12 - 2017-08-29 08:48 - Laura Moyers

- % Done changed from 10 to 50

- Status changed from Planning to Testing

#13 - 2017-08-30 20:31 - Laura Moyers

- Target version changed from Deploy by end of Y3Q4 to Deploy by end of Y4Q1

#14 - 2018-01-08 19:17 - Monica Ihli

Summary of Pangaea thus far:

DataONE is hosting both the development and production installations of GMN software for metadata only implementations of Pangaea. Pangaea's metadata is exposed using their OAI-PMH web service and integrated with GMN using a Python 2.7 adapter.

Metadata Format:Pangaea is using a custom formatId implementation of iso-19115 as a result of a mismatch between the version of GMD referenced within their documents and the version actually needed to validate as a dependency. This custom formatId has been installed and tested in Stage, and we are satisfied with testing that proves the efficacy of this formatId for validating Pangaea metadata content.

An additional issue uncovered during testing Pangaea installation in the Stage environment was limitations to the capacity of synchronization to harvest greater than 50,000 records at a time. A solution to this problem has been developed and is currently installed on Sandbox version of CNs.

Testing: The test GMN is installed as 2.3.8 at host pangaea-dev-orc-1.test.dataone.org. Testing actually started off in stage, but the GMN settings are currently pointed switched to sandbox because that's where the synchronization changes were later being tested.

Production: The production GMN is currently installed as 2.4.0 at host pangaea-orc-1.dataone.org. The data is currently being loaded. The CURRENT status is that the prod GMN is unregistered and being loaded in standalone mode.

Indexing: Indexing the data (which affects how the metadata records are presented to end users in the search interface) was discussed at one point. We have considered whether or not to customize how indexing is performed with relation to the custom formatId. However at this point, we are holding off on any changes, and the gmd-pangaea formatId will follow the same indexing rules as gmd-noaa. We can always reindex at a future point in time if needed.

Moving forward:

Synchronization changes put into production: Confirm schedule for synchronization changes being put into production, as Pangaea going into production will depend on that.

Register in Production: Register in production when synchronization changes are in place.

Work with MNC: To communicate size of node, workflow for news release and other administrative tasks for going live.

Version: 2.4.1 is out. Not sure if we should update again before going live.

#15 - 2018-01-08 20:07 - Amy Forrester

Communication Assist about information for testing:
1/8/18: email Michael Diepenbroek mdiepenbroek@pangaea.de to get the approx. order of magnitude of PANGAEA number of metadata records. [some discrepancy between # of objects in our dev instance vs production]

1/9 (Response): We keep around 370 Tsd records
* Total number can be found by search engine: https://www.pangaea.de/?q=*
* All from set “citable”: https://www.pangaea.de/?q=@citable
* All from set “citableWithChilds”: https://www.pangaea.de/?q=@citableWithChilds
Actually, the number is the search engine can be a bit lower, because our search engine does not show our dataset’s older versions (it only shows the newest ones). The total difference is around 1000, as far as I remember.

#16 - 2018-01-09 19:58 - Dave Vieglais

- Sprint set to Q1-2018

#17 - 2018-01-22 01:02 - Monica Ihli

370,000 + objects loaded in production GMN. We will be waiting on the next CI update to push sync changes to production before registering.

One additional important note: Whereas sandbox/stage used a concatenation of native system identifier (which is a doi) + datetimestamp for assigning PID (the native system id is assigned as our SID), a change has been made that in production, the checksum is assigned as pid. This is to avoid confusion when an end user is looking at record details in the search interface, so they are not presented with two doi-looking values (pid and sid), only one of which would actually resolve.

#18 - 2018-02-13 19:01 - Monica Ihli

Update 2.3.7 to CN is deployed to Stage. This update handles the indexing changes that that address synchronization limits that were previously encountered in Stage. We are switching Pangaea's test installation to point back to Stage and will re-enable sync. Last harvest date will be reset and will attempt to allow a complete resync to monitor for success with the new code updates. If all is well, 2.3.7 could be deployed next week to Production.

An additional step is that the Pangaea custom formatID needs to be manually registered in Production.

Once 2.3.7 is is deployed to prod & formatID is registered on the CN, the Pangaea production installation of GMN will be registered in prod. To review: the prod installation of the GMN app is already loaded with 370k plus objects. It's just waiting to be registered. Since we are so close to the finish line, Amy will start getting things in place such as news announcements and other steps to prepare for Pangaea officially going live.

#19 - 2018-02-16 15:08 - Amy Forrester

2/16/18: updated POCs about sync optimization and broad overview of steps that will ensue

#20 - 2018-03-01 18:51 - Amy Forrester

- MN Description set to PANGAEA is an open access data publisher for Earth & environmental science supporting the long-term archiving and publication of georeferenced data related to environmental sciences

#21 - 2018-03-01 18:56 - Amy Forrester

- Logo URL set to https://github.com/DataONEorg/member-node-info/blob/master/production/graphics/web/PANGAEA.png

- Information URL set to https://www.pangaea.de/

- Base URL deleted (https://www.pangaea.de/)

- Name set to PANGAEA

#22 - 2018-03-01 18:57 - Amy Forrester

- % Done changed from 50 to 80

- Name changed from PANGAEA to PANGAEA Data Publisher

- Status changed from Testing to In Review

#23 - 2018-03-01 19:00 - Amy Forrester

- Date Upcoming set to 2018-02-23

#24 - 2018-03-01 19:03 - Amy Forrester

- Version set to 2.4.0

#25 - 2018-03-20 18:13 - Amy Forrester

- MN_Date_Online set to 2018-03-20

#26 - 2018-05-02 17:51 - Monica Ihli

- Related to Story #8574: PANGAEA Temporary Fix: SID only in Data Citation added

#27 - 2018-05-08 14:48 - Monica Ihli

- Base URL set to Prod: pangaea-orc-1.dataone.org/mn / Stage: pangaea-dev-orc-1.test.dataone.org

#28 - 2018-05-24 15:37 - Amy Forrester

- Status changed from In Review to Operational

#29 - 2018-12-04 14:58 - Amy Forrester

- Target version changed from Deploy by end of Y4Q1 to Operational