Member Nodes - MNDeployment #4730

Global Biodiversity Information Facility

2014-04-15 01:41 - Bruce Wilson

Status:	Testing	**Start date:**	2014-07-17
Priority:	Normal	**Due date:**	
Assignee:	Dave Vieglais	**% Done:**	50%
Category:		**Estimated time:**	0.00 hour
Target version:	Deploy by end of Y4Q1		
Latitude:	55.70	**MN_Date_Online:**	
Longitude:	12.56	**Name:**	Global Biodiversity Information Facility
MN Description:	GBIF—the Global Biodiversity Information Facility—is an open-data research infrastructure funded by the world's governments and aimed at providing anyone, anywhere access to data about all types of life on Earth.	**Logo URL:**	https://github.com/DataONEorg/member-node-info/blob/master/production/graphics/web/GBIF.png
Base URL:	https://labs.gbif.org:7005/mn	**Date Upcoming:**	2014-07-17
NodeIdentifier:	urn:node:GBIF	**Date Deprecated:**	
MN Tier:	Tier 4	**Information URL:**	http://gbif.org/
Software stack:	Hadoop	**Version:**	

Description

Tim Robertson at GBIF is working on a Hadoop-based MN stack which would allow GBIF to come on board as a MN.

Subtasks:

Task # 5823: GBIF: Initial Preparation	**Closed**
Task # 5824: GBIF: Establish contact with potential MN organization	**Closed**
Task # 5825: GBIF: Determine feasibility	**Closed**
Story # 5826: GBIF: Planning	**Closed**
Task # 5827: GBIF: Join DataONE	**Closed**
Task # 5828: GBIF: Integrate MN representatives with community	**Closed**
Task # 5830: GBIF: Create MN description document	**Closed**
Task # 5831: GBIF: Scope the implementation	**Closed**
Task # 5832: GBIF: Plan the implementation	**Closed**
Story # 5833: GBIF: Developing	**In Progress**
Task # 5835: GBIF: Local Testing	**Ready**
Task # 5836: GBIF: Verify that MN passes the Web Tester	**Testing**
Story # 5838: GBIF: Testing	**New**
Task # 5839: GBIF: Registration in environment	**New**
Task # 5840: GBIF: Register new Science Metadata formats	**Rejected**
Task # 5841: GBIF: SSL Certificates	**New**
Task # 5845: GBIF: Register MN	**Closed**
Task # 5846: GBIF: Set up Node document	**Closed**
Task # 5847: GBIF: Set the node status to approved (start synchronization)	**Closed**
Task # 5848: GBIF: Synchronization	**New**
Task # 5849: GBIF: Set up synchronization of the MN	**New**
Task # 5850: GBIF: Content Review	**New**
Task # 5851: GBIF: Verify Science Data	**New**
Task # 5852: GBIF: Verify Science Data content	**New**
Task # 5853: GBIF: Verify that the Science Data is returned with the correct HTTP Conte...	**New**
Task # 5854: GBIF: Verify Science Metadata	**New**

Task # 5855: GBIF: Verify Science Metadata content	**New**
Task # 5856: GBIF: Verify Science Metadata is returned with the correct HTTP Content-Type	**New**
Task # 5857: GBIF: Verify that the Science Metadata is correctly processed by CNs	**New**
Task # 5858: GBIF: Verify Resource Maps	**New**
Task # 5859: GBIF: Verify Resource Map content	**New**
Task # 5860: GBIF: Verify Resource Map is returned with the correct HTTP Content-Type	**New**
Task # 5861: GBIF: Verify that Resource Maps are correctly processed by CNs.	**New**
Task # 5862: GBIF: Authentication and Authorization	**Rejected**
Task # 5863: GBIF: Science Data access	**Rejected**
Task # 5864: GBIF: Science Metadata access	**Rejected**
Task # 5865: GBIF: Log record access	**New**
Task # 5866: GBIF: Replication testing (if Tier 4)	**Rejected**
Story # 5867: GBIF: Registration in Production environment	**New**
Task # 5868: GBIF: Register new Science Metadata formats	**Rejected**
Task # 5869: GBIF: SSL Certificates	**New**
Task # 5870: GBIF: Generate client certificate.	**New**
Task # 5871: GBIF: Verify successful installation of client side certificate	**New**
Task # 5872: GBIF: Verify successful installation of server side certificate	**New**
Task # 5873: GBIF: Register MN	**New**
Task # 5874: GBIF: Set up Node document	**New**
Task # 5875: GBIF: Set the node status to approved (start synchronization)	**New**
Task # 5876: GBIF: Synchronization	**New**
Task # 5877: GBIF: Set up synchronization of the MN	**New**
Task # 5878: GBIF: Content Review	**New**
Task # 5879: GBIF: Verify Science Data	**New**
Task # 5880: GBIF: Verify Science Data content	**New**
Task # 5881: GBIF: Verify that the Science Data is returned with the correct HTTP Conte...	**New**
Task # 5882: GBIF: Verify Science Metadata	**New**
Task # 5883: GBIF: Verify Science Metadata content	**New**
Task # 5884: GBIF: Verify Science Metadata is returned with the correct HTTP Content-Type	**New**
Task # 5885: GBIF: Verify that the Science Metadata is correctly processed by CNs	**New**
Task # 5886: GBIF: Verify Resource Maps	**New**
Task # 5887: GBIF: Verify Resource Map content	**New**
Task # 5888: GBIF: Verify Resource Map is returned with the correct HTTP Content-Type	**New**
Task # 5889: GBIF: Verify that Resource Maps are correctly processed by CNs.	**New**
Task # 5890: GBIF: Authentication and Authorization	**New**
Task # 5891: GBIF: Science Data access	**New**
Task # 5892: GBIF: Science Metadata access	**New**
Task # 5893: GBIF: Log record access	**New**
Task # 5894: GBIF: Replication testing (if Tier 4)	**Rejected**
Story # 5895: GBIF: Transition to production	**New**
Task # 5896: GBIF: Mutual acceptance	**New**
Task # 5897: GBIF: Verify content available for Current MNs web page	**New**
Task # 5898: GBIF: Create legal documents	**Closed**
Task # 5899: GBIF: Create news item	**New**
Task # 5900: GBIF: Formal announcement	**New**

Related issues:

Related to Member Nodes - MNDeployment #4296: Sierra Nevada Global Change Obs...	**New**	**2014-06-20**	**2015-01-31**

History

#1 - 2014-04-15 01:42 - Bruce Wilson

Ongoing discussions between Tim and Dave V during late March and early April 2014; some discussion with Bruce Wilson during EU BON meeting in Crete, early April 2014.

#2 - 2014-04-15 04:21 - Chris Jones

- Subject changed from Global Biodiversity Information Framework to Global Biodiversity Information Facility

#3 - 2014-05-29 18:58 - Bruce Wilson

Update from Tim Robertson: Has certificates working, but had to put this on hold for some other work. Expects to have significant progress before CCIT,

#4 - 2014-08-06 18:24 - Laura Moyers

- Target version changed from Deploy by end of Y5Q4 to Deploy by end of Y1Q2

- Due date changed from 2014-07-31 to 2015-01-31

#5 - 2014-08-18 14:37 - Laura Moyers

- Longitude changed from 12.57 to 12.56

- Latitude changed from 55.68 to 55.70

- NodeIdentifier set to urn:node:GBIF

#6 - 2014-09-15 13:56 - Laura Moyers

- Due date changed from 2015-01-31 to 2015-07-31

- Target version changed from Deploy by end of Y1Q2 to Deploy by end of Y1Q4

Based on current information from Tim Robertson, development on the GBIF MN has been moved out several months.

#7 - 2014-09-15 18:28 - Laura Moyers

GBIF's announcement of the MOC with DataONE: http://www.gbif.org/page/8199

#8 - 2015-04-09 20:56 - Laura Moyers

- Target version changed from Deploy by end of Y1Q4 to Deploy by end of Y2Q1

GBIF anticipates a delay in significant effort on developing their MN.

#10 - 2015-10-12 19:56 - Laura Moyers

- Target version changed from Deploy by end of Y2Q1 to Deploy by end of Y2Q2

#11 - 2015-11-19 17:49 - Laura Moyers

- Related to MNDeployment #4296: Sierra Nevada Global Change Observatory Member Node added

#12 - 2015-12-14 18:33 - Laura Moyers

- Target version changed from Deploy by end of Y2Q2 to Deploy by end of Y2Q4

Latest (12/9/15) from Tim is:

I'm afraid we put it [DataONE MN development] on hold some time ago, and have not yet picked it up again. We still fully intend to, but when we got into it, found that building out a member node stack is actually quite a big undertaking for what appears at first glance to be relatively simple – it is literally several months of work. Should DataONE continue it would be really useful to have a solid software stack that makes it easy for people to connect datastore (in e.g. Java). Currently implementers have to understand all manner of details that should be internal to DataONE, such as error codes, and certificate structures etc. I'd like to see a Java interface with 7-10 methods or so to implement. Our implementation will provide this, and we hope that it will then be of use to others.

The GBIF informatics team have 7 items to fulfil for 2016, one of which is the DataONE member node implementation and a web site for people to upload into the repository. I'd expect we will have a pilot up and running by April but I don't anticipate it will be fully functional. Part of this is due to me taking extended unpaid leave (10 weeks) from the end of next week, having recently had a baby. While I am away, Christian (CC'ed) will likely be picking up the DataONE stuff, but I expect it will start in earnest in March now.

If DataONE were interested in putting development resources on this stack it would be welcome, but you have your own Java stack so I would not

expect you would want to. I explained to the D1 dev group in some peer reviews why we would be reluctant to use those DataONE libraries (thread safety issues, unwanted dependencies). The stack we are building out is started on https://github.com/timrobertson100/dataone – it is in a shoddy state right now, but reasonably well advanced, having tackled the more time consuming things like the authentication and software architecture and build process (e.g. building out code from XSDs etc).

#13 - 2016-07-02 12:45 - Laura Moyers

- Target version changed from Deploy by end of Y2Q4 to Deploy by end of Y3Q1

#14 - 2016-11-21 17:58 - Laura Moyers

- Target version changed from Deploy by end of Y3Q1 to Deploy by end of Y3Q2

Moving out in schedule.

#15 - 2017-01-31 21:04 - Laura Moyers

- Target version changed from Deploy by end of Y3Q2 to Deploy by end of Y3Q4

Latest conversation with Tim in December 2016:

> Does anyone know of any effort to build a module for Invenio [1] for DataONE?
> Invenio is the digital library from CERN in Europe, with https://zenodo.org being the most high profile installation I'm aware of.
>
> We're evaluating a few options here for the GBIF repository that we have been postponing and a requirement is to integrate with DataONE. We've got a mostly working D1 Java stack [2] but are pondering if we really do want to build out a repository from scratch.
>
> Thanks,
> Tim
>
> [1] http://invenio-software.org/
> [2] https://github.com/gbif/dataone

Dave replied:

> I'm not aware of anyone working on DataONE support for Invenio, though the architecture seems quite approachable for such an implementation. Are you considering using that stack within GBIF as a collection repository?

Tim:

> Yes, we're currently exploring a few options and that¹s one for a general purpose repository - custom exports, shapefiles, the quarterly GBIF index snapshots, people who¹ve created derived datasets and want to cite using DOIs in a publication etc. It¹s not really for data that is well structured (like collections / observational datasets) for which we¹ve got the IPT based repositories that do the classic mapping to DwC. Zenodo is often cited as the one to emulate in terms of usability and with CERN backing, a fairly modern tech stack and DataCite integration it seems like a strong candidate for our needs. We also have some TBs in HDFS we¹d like to expose too which is the reason we have the standalone Java stack.
>
> Did you ever adopt HDFS / Hadoop / Spark or anything in that space? I seem to recall you were exploring getting a Cloudera Enterprise license - what did you use that for if anything in the end?

Dave:

> We decided not to pursue HDFS, Hadoop, Spark at this stage, mostly because of resource limitations (issues with hardware vendors). Our plans for that technology centered on the content indexing pipeline, though we have been quite happy with SolrCloud and it's underlying zookeeper coordination without the need to involve additional infrastructure at this stage.
>
> Let us know if you decide to deploy Invenio. I'd like to explore Invenio a bit, to get a better idea of effort required to implement a plugin that will support the MN API. We have a good set of Python libs that will help with implementation, though of course there's a lot of low level detail that might throw in a few blockers.

#16 - 2017-06-25 19:00 - Laura Moyers

GBIF is in a position now to proceed with MN development. We would like to have a GBIF MN in place before the RDA meeting in September. Notes from 6/23/2017:

timrobertson100 [8:02 AM]Hi Folks. GBIF (fmendez leading) will stand up an installation of our D1 MN stack next week to start some tests with you. This is a java software stack built primarily around Dropwizard. We designed it to work with any backend, but we'll be backing ours by Hadoop HDFS. Our first objective will be to get the 28 snapshot views of all GBIF.org data into D1 (i.e. views of GBIF dating back to 2008). The second objective will be to enable full tier 4 replication. Next week we'll just run it with an in-memory backend, and we might have some questions on how to run conformance tests. Our codebase is https://github.com/gbif/dataone (edited)

davev [8:11 AM] That's great Tim - looking forward to seeing GBIF as a participant in DataONE. Approximately how large are the snapshot views?

timrobertson100 [8:29 AM]not entirely sure how we will format this… Worst case: 14:28:10 c5n1 ~ $ hdfs dfs -du -h -s /c4-backup/user/hive/warehouse/snapshot.db 5.8 T 17.3 T /c4-backup/user/hive/warehouse/snapshot.db shows 5.8TB of compressed data (17.3TB because we have 3x replication) but that is probably a fair few more views of the data than we'll prepare. I expect we'll store the raw views, and then every quarter we'll store each view processed to the latest QA routines. So maybe growing at 0.5TB a quarter or so (edited)

davev [8:30 AM] is that 5.8TB / snapshot?

timrobertson100 [8:30 AM] no - that would be all

davev [8:32 AM] would each snapshot be treated as a single blob or file?

timrobertson100 [8:33 AM] We've just set up a new HDFS cluster, which is showing 153TB used with 365TB free. I'd expect we'll be offering 50-100TB of replication space, and would target ecological datasets if at all possible. That way we can start trying to data mine for tabular data with species scientific names, and try and build the taxa+space+time indexes of tabular data

davev [8:33 AM] Nice

timrobertson100 [8:33 AM]each snapshot would be a compressed CSV of verbatim data and another of the same data after QA

davev [8:34 AM]what metadata format is used to describe the snapshots?

timrobertson100 [8:34 AM]Specifically each is about 30 columns of Darwin Core. We'd craft that metadata - Dublin Core, EML or probably more applicable would be DataCite. We'll DOI everything, so DataCite for sure. But your guidance appreciated there.

davev [8:36 AM]so DataCite is fairly light on the details. It's ok to assist with discovery but falls short of a rich description of a dataset

timrobertson100 [8:38 AM] The issue though, is that these represent a derived few of 10,000+ datasets. Each of those has EML though

davev [8:38 AM] yeah, I was just pondering that

timrobertson100 [8:38 AM] we could archive each dataset independently as well, but I had thought we might start with the snapshots first (to enable people to do large scale analysis and change in GBIF.org across time) (edited)

[8:39]
https://api.gbif.org/v1/image/unsafe/http://www.gbif.org/sites/default/files/gbif_analytics/global/figure/occ_kingdom.png (125kB)
shows the data growth on those snapshots

davev [8:40 AM]
an EML description of a snapshot would at least enable description of the individual columns

timrobertson100 [8:41 AM]
Yeah, we could do that.

davev [8:41 AM]
that's a lot of growth and not slowing down

timrobertson100 [8:42 AM]
The DataCite kernel though, allows us to link the DOIs of each individual contributing dataset using the (or something like this) so you can trace those back to each source and it's EML. we're anticipating 30% growth this year. The citations are growing quickly too. you can find those here [[https://demo.gbif.org/resource/search?contentType=literature]] - I think it was 78 peer reviewed citations last month (edited)

davev [8:44 AM] that is awesome.so each collection has a DOI, and each snapshot, which is a collection of collections, has a DOI as well, right?

timrobertson100 [8:45 AM] We'd like to chat with you and DataCite at the RDA meeting about this. We're finding our way a bit here, and would like to align what we do with you. Yes. and then we crawl for papers. So paperDOI -> downloadDOI -> "many source dataset DOIs". This allows us to show publishers which papers made use of parts of their data. We offer filtered search across all records, so normally a use spans parts of many datasets. https://demo.gbif.org/occurrence/search is a SOLR Cloud index across 782M records (download temporarily disabled as we move datacenters today)

#17 - 2017-08-25 00:12 - Laura Moyers

- MN Description set to GBIF—the Global Biodiversity Information Facility—is an open-data research infrastructure funded by the world's governments and aimed at providing anyone, anywhere access to data about all types of life on Earth.

- MN Tier set to Tier 4

GBIF is currently testing with the WebTester.

#18 - 2017-08-29 08:48 - Laura Moyers

- % Done changed from 6 to 50

- Status changed from Ready to Testing

#19 - 2017-08-30 20:33 - Laura Moyers

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

#22 - 2017-10-07 03:57 - Monica Ihli

Follow up notes from web testing:

Need to confirm the time zone of the server being used: System metadata is being stamped with +2:00 such as "2017-10-03T00:00:00.000+02:00".

Need to consult with MN for testing/implementation of updates.

+*Tier 1:*+

- *MNCore* - All tests pass except for *"getLogRecords - test event filtering"*. This is supposed to test ability to filter on event type. But https://labs.gbif.org:7005/mn/v1/log?event=read successfully does so. The actual url used per stack trace was https://labs.gbif.org:7005/mn/v1/log?toDate=2017-10-06T19:10:02.452%2B00:00&event=read&start=0&count=0, from which the test expected 0 results returned. The issue may be due to appearance of discrepancy between server time zones.
- *MNRead* - All tests pass except for *"getReplica - tests getReplica returns a valid object"* However comments indicate that if node is not yet registered, then this error can be ignored. (Node is not currently registered)
- *Authentication* - Complaining about *"Authentication - test with self-signed certificate"* failing to return a Not Authorized exception.
- *Content Integrity* - passes

+*Tier 2:*+
* MNAuthorization passes - Note: originally there were some issues with the online version but Rob was able to successfully run it from his local machine, which populated the test objects needed for it to pass now. This points to the need for some updating to what's on the server now.

+*Tier 3:*+

- *MNStorage API* - "update - tests that update works" fails. This specifically attempts to create an object and then update it. The test was unable to successfully perform an update and note of the update-related pids (such as mNodeTier3TestUpdate201727993310335) were found to have obsoletes/obsoleted properties. Furthermore, *update - tests with bad obsoletedBy info*, and *update - tests with bad obsoletes* failed to return invalid system metadata exceptions.

+*Tier 4:*+

- *MNReplication* - All tests pass.

Overall there seems to be some outstanding issues with the web tester version online which were confirmed to affect MN Authorization and MN read (different errors in stack trace), suggesting that what's on the server is outdated in some places.

#23 - 2017-12-05 16:37 - Monica Ihli

Last communication w/ Federico on 12/5:
GBIF is working on a T1/T2 implementation in small project called DataRepo (https://github.com/gbif/data-repo/) which is the back-end of their DataOne MN, and is almost ready to be released. Their plan is to have the DataOne MN working initially as Tier1 or Tier2, and then support tiers 3 and 4.

From their side they are pretty much ready to start testing operations of tiers 3 and 4 but prefer to do it incrementally.

#24 - 2017-12-15 14:48 - Monica Ihli

GBIF on 12/14: We have a testing public environment now for our member node, https://ws.gbif-uat.org:7000/mn/v1/, it took us a while to make it available because of firewall restrictions, this MN is backed by another service that we putting on production that is called the GBIF Data Repo which can be found here http://api.gbif-uat.org/v1/data_packages/ (staging/user acceptance test environment)

12/15:
* Asked Federico an estimation of how many data objects will be exposed via the MN. Response: amount of objects will be, initially, no more 20, we'd like to expose GBIF snasphots , each is about 4 GB.

- Checking the MN, I observed that a count of 20 is being returned when there is only 1 object being exposed by the MN. For example only one object was being returned from https://ws.gbif-uat.org:7000/mn/v1/object?start=0 on the morning of 12/15, but the list objects response is:

42592444-b1c7-4086-9625-66153b8c90b5
ba0fb20240817bab7ef68667c1915962
2017-12-14T13:58:43.590Z
6821285

/ns2:objectList

Provided a link to type documentation at
http://jenkins-1.dataone.org/jenkins/job/API_Documentation_trunk/ws/api-documentation/build/html/apis/Types.html#Types.Slice . GBIF is excited about their progress and eager to get registered in Sandbox. This will probably take place after the new year as the listObjects() functionality is central to harvesting and should be responding predictably before we attempt to harvest content.

#25 - 2017-12-22 14:50 - Amy Forrester

(Tim Robertson) add Federico Mendez (fmendez@gbif.org) as well to the list. He and I together will be able to answer questions, and cover during vacations and travel.

#26 - 2018-01-29 18:48 - Monica Ihli

Approaching GBIF w/ invitation to consider using GMN with OAI-PMH as a temporary stand-in to get GBIF content online until they get further with custom development.

1/29/2018: {Amy} email tim & Frederico with the idea.

* Tim out of office until 2/26

Continue --> Task #5836

#27 - 2018-03-01 19:53 - Amy Forrester

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

- Information URL set to http://gbif.org/

- Base URL changed from http://gbif.org/ to https://labs.gbif.org:7005/mn

- Name set to Global Biodiversity Information Facility

#28 - 2018-03-01 19:55 - Amy Forrester

- Logo URL deleted (https://github.com/DataONEorg/member-node-info/blob/master/production/graphics/web/GBIF.png)

- Date Upcoming set to 2014-07-17

- Information URL deleted (http://gbif.org/)

- Base URL changed from https://labs.gbif.org:7005/mn to http://gbif.org/

- Name deleted (Global Biodiversity Information Facility)

#29 - 2018-03-01 19:56 - Amy Forrester

- Base URL changed from http://gbif.org/ to https://labs.gbif.org:7005/mn

- Name set to Global Biodiversity Information Facility

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

- Information URL set to http://gbif.org/

#30 - 2018-04-06 20:18 - Monica Ihli

Contacted by Federico to report that they are testing MN on web tester, but some now deleted identifiers seem to have been cached.

#31 - 2018-09-26 19:45 - Rob Nahf

Notes on the MN API: I am testing their node as a v1 Tier 1. They chose NOT to implement the optional pidFilter parameter to getLogRecords(), and formatId parameter to listObjects().

Their packages are zip acrhives of 6-month snapshots of their collection, and are composed of 1 ORE, 1 EML metadata doc, and 1 zip archive. The zip archive is multiple GB, so may be difficult to get without adjusting client timeouts and possible memory .