Sessions 3/4: Member Node Breakouts

John Cobb
Matt Jones
Laura Moyers
7 July 2013
DataONE Users Group
Schedule

1:00-2:20 and 2:40-4:00 – Member Node Breakouts

- Member Node Overview and Process Overview
- Documentation and Communications
- Data driven Science, Data archives (MNs) and DataONE Integration opportunities
- Summary / Questions & Comments

http://epad.dataone.org/dug2103-BK1-MNs
Member Nodes

Authoritative members of the Federation
Curate data holdings
  Provide unique identifiers for each object
  Ensure availability, quality, and reliability
Log and report accesses to objects
Control access to data and metadata
Replicate holdings for other MNs
Deploy a DataONE-compatible software system
DataONE Design goals

Embrace heterogeneity via Data Packaging
Preservation
High availability
Reproducibility, Immutability and Versioning
Logging

All within a distributed, heterogeneous, autonomous federation of Member Nodes
Persistent Identifiers

Goal: Uniquely identify data or metadata objects to support reproducible analysis via data citation

• Every object in DataONE gets a Persistent Identifier
• Not-reusable
• Indirect reference to immutable content
• 800 Unicode characters or less
• Whitespace and non-printing characters illegal

http://mule1.dataone.org/ArchitectureDocs-current/design/PIDs.html
Some identifiers

- doi:10.5063/AA/duc_merp.126.4
- ark:/13030/m5zp459k/1/cadwsap-s5800837-001.xml
- urn:uuid:e26ef510-cfcd-11e1-ae97-f4e395c5a4c
- urn:lsid:ubio.org:namebank:11815
- duc_merp.126.4
- ข้อมูลที่เป็นประโยชน์
DataONE Packages

Goal: Aggregate heterogeneous data and metadata objects, linking among components

- Flexibly describe complex data structures
- Supports arbitrary file formats
- Virtual aggregations by reference
- Objects can be “in” multiple packages
- Extensible model for relationships
- Linked-data compatible

http://mule1.dataone.org/ArchitectureDocs-current/design/DataPackage.html
Package Model

Each object:
- Has unique identifier
- Content does not change

Data

System Metadata

Any data object

Resource Map

OAI-ORE RDF

Science Metadata

System Metadata

XML documents: ISO19115, EML, FGDC, …
Discover Content: Metadata Formats

Extract and index common fields from metadata standards

- Ecological Metadata Language (EML)
- FGDC Biological Data Profile (BDP)
- ISO 19115 Geospatial Metadata
- Dublin Core
- Darwin Core
- METS

Extensible to include many more

- DIF, NexML, WaterML, CF, NcML, ESML, DDI, MIENS, ...
Road Map to a Member Node

Plan
- Determine feasibility
- Join the DataONE federation
- Select the tier
- Plan the implementation

Develop
- Use existing MN software?
  - Yes
    - Develop iteration
    - Develop test
    - Establish a test system
  - No
    - Develop iteration

Deploy
- Deploy in staging environment
- Test in staging environment
- Deploy in production
- Mutual acceptance

Operate
- Announcement
- Ongoing operations
- Participate in MN Forums

Yes

No
Plan: Member Node Tiers

Plan

- Determine feasibility
- Join the DataONE federation
- Select the tier
- Plan the implementation

Tier 1: Public content
Tier 2: Access control
Tier 3: Write services
Tier 4: Act as a replication target
Wrap your own
- Implement the Tier 1 MN APIs however you like
- Implement Tiers 2-4 as you need

Use existing repository software
- Metacat (Tier 4)
- Generic Member Node (Tier 4)
- Dryad (Tier 1)
- Mercury (Tier 1)
- Merritt (Tier 1)
- ... more coming...
Deploy: The Bottom Line

- Hardware
  - CPUs
  - Storage
- Infrastructure
  - Power
  - Internet
  - Facilities
- Administration
  - MN operation
  - Data curation
- Implementation
  - Design
  - Development
- Long-term
  - Maintenance
  - Migration
  - Shutdown

Deploy

- Deploy in staging environment
  - Test in staging environment
  - Deploy in production
  - Mutual acceptance
Operate: Moving to Production

To flip the switch:
• Must pass all tests required for tier
• Security updates, patches applied
• Administrative procedures in place
• Agreements in place

Once running:
• Announcement
• Maintain member node integrity
• Respond to administrative requests
• Community participation
Typical Barriers

During deployment
Establishing secure SSL environment
Providing package ORE maps

Design Issues
Immutability & Versioning
Logging
Questions and Discussion

Member Node Checklist
http://mule1.dataone.org/OperationDocs/member_node_deployment/mn_checklist.html

http://epad.dataone.org/dug2103-BK1-MNs
MN Documentation and Communication

Current communication channels
• Are they working?
• What can make them better?

Documentation is a form of communication
• Current status
• What can make it better?
DataONE communications channels with MNs

- Member Node Forum (MNF)
  - Meets biweekly
  - Desire for MNF to be:
    - One stop communications start location
    - Especially for issues common to MNs (which most are)
    - A place for MN leverage (MNs helping MNs directly)
- IRC (Internet Relay Chat)
- Redmine – can follow status of MN tickets
- DataONE developers mailing list – great for technical POCs

http://epad.dataone.org/dug2013-BK1-MNs
MN documentation status

Current documentation

• Overview (what is DataONE, how to become a member node, etc.)
  • DataONE website https://www.dataone.org/
  • https://ask.dataone.org/
• Technical documentation
  • Architecture documentation
    http://mule1.dataone.org/ArchitectureDocs-current/
  • Operations documentation
    http://mule1.dataone.org/OperationDocs/

http://epad.dataone.org/dug2013-BK1-MNs
MN documentation status

What’s missing?

• A link between the high-level overview information and the technical details

• It needs to be
  • accessible
  • user-friendly
  • comprehensive

http://epad.dataone.org/dug2103-BK1-MNs
Road Map to a Member Node

Plan

Determine feasibility
Join the DataONE federation
Select the tier
Plan the implementation

Develop

Use existing MN software?
Yes
No

Developmen t iteration
Developmen t test
Establish a test system

Deploy

Deploy in staging environment
Test in staging environment
Deploy in production
Mutual acceptance

Operations

Announcement
Ongoing operations
Participate in MN Forums
Questions: Documentation

1. What should be seen on the DataONE website regarding Member Nodes?

http://epad.dataone.org/dug2103-BK1-MNs
Questions: Documentation

1. What should be seen on the DataONE website regarding Member Nodes?
2. What new questions and answers are needed on ask.dataone.org re: Member Nodes (and science users)?

http://epad.dataone.org/dug2103-BK1-MNs
Questions: Documentation

1. What should be seen on the DataONE website regarding Member Nodes?
2. What new questions and answers are needed on ask.dataone.org re: Member Nodes (and science users)?
3. What 2-3 things do you need to know, or some questions that you have had trouble finding the answer to, that would improve documentation?

http://epad.dataone.org/dug2103-BK1-MNs
Question: MN communications

When and how do you use

- IRC?
- Online documentation?
- Developers mailing list?
- CCIT calls
- Redmine?

How can we maximize the utility of the MNF and other communication means?

http://epad.dataone.org/dug2103-BK1-MNs
For science researchers, how can DataONE help amplify the science impact of the data archives represented by the Member nodes?

- **Known:**
  - Resilient access
  - Persistence
  - A single interface (may or may not be preferred)
  - Unified discovery across archives
- **Pilot demonstrations**
  - Scaling
  - Enabling new investigations (may be scaling related)
- **Not known:**
  - Routinely access multiple archives simultaneously (incl. synthesis)

- **Known not possible**
  - Build multi-archive interoperable workflows
  - Building useful/powerful derived data products
  - Silver bullets
  - Everyone gets a pony
  - Free lunches
What can MNs and DataONE do to augment synergistic opportunities?

- What are the general areas of opportunity?
- How to identify the ripe, low hanging fruit?
- What are the principal challenges?
- Can one DUG function as a self-sustaining reservoir for this discussion?

Questions:

http://epad.dataone.org/dug2103-BK1-MNs
Questions:

What can MNs and DataONE do to augment synergistic opportunities?

How to find and assist science users who want to use multiple archives (multiple MN’s)?

• What is the value to science users beyond direct access to individual archives?
• How to enable this value within DataONE?
• What are the important differences between using different MNs in DataONE versus using multiple access methods (for example, multiple datanets)

http://epad.dataone.org/dug2103-BK1-MNs
Questions:

What can MNs and DataONE do to augment synergistic opportunities?
How to find and assist science users who want to use multiple archives (multiple MN’s)?

Other Questions?

http://epad.dataone.org/dug2103-BK1-MNs
Contact Points

John Cobb  cobbjw@ornl.gov  865.576.5439
Matt Jones  jones@nceas.ucsb.edu
Laura Moyers  lmoyers1@utk.edu
Member Node Forum (contact us to be added)
http://www.dataone.org/contact
Metacat (Tier 4)

Flexible storage system for metadata and data
- Stores, search, and document data
- Java webapp runs on Linux, Windows, MacOS
- Deployed worldwide, maintained over 10 years
- Web-based search interface
- Customizable user interface
- Web metadata entry tool
- Tier 4 Replication capabilities
- Postgres or Oracle backend
- DOI Support
- OAI-PMH harvester
- GPL open source license
Generic Member Node (GMN) (Tier 4)

Lightweight system to wrap backend data stores

- Tier 4 compliant
- Simple to install
- Configurable to wrap different back end systems
- Python Django implementation
- Linux, MacOS X, Windows
- Apache2 open source license

- No user interface

- http://repository.dataone.org/software/cicore/trunk/mn/d1_mn_generic
Dryad (Tier 1)

Data from peer-reviewed articles in bioscience

- Metadata schemas focused on METS
- Submissions integrated with journal workflows
- Customized DSpace repository
- DOI Support
- OAI-PMH harvester
- Data usage displays
- Dublin Core Application Profile
- Data file format determined by depositor and journal policy
- Some curation and migration of file formats
Mercury (Tier 1)

Flexible data and metadata repository

- Supports common metadata formats (FGDC, Dublin-Core, EML, ISO-19115)
- Open source, Java-based webapp jointly developed by USGS, DOE, NASA and NSF
- SOLR-based member node implementation for metadata
- File-system store for data
- Uses existing Clearinghouse harvesting architecture
- DOI support
- Deployed across many environmental data projects
- Web-based search interface
- Customizable user interface
- Web metadata entry tool
Merritt (Tier 1)

Flexible data repository from California Digital Library

- Implementation uses Metacat for DataONE services
- Micro-Services approach to digital curation
- Easy-to-use interface for deposit and update
- Integration with EZID/DOI support
- Tools for long-term management