Member Node Topics:
Synchronization

DataONE Member Node Forum
2016-09-01

DataONE is a collaboration among many partner organizations, and is funded by the US National Science Foundation (NSF) under a Cooperative Agreement. This material is based upon work supported by the National Science Foundation under Grant Numbers 0830944 and 1430508. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.
Three Components of DataONE

- ITK
- MNs
- CNs
Three Components Integration

DataONE Service Specifications

ITK

HTTPS REST XML Messages

MNs

CNs
Shown in a High Level Design

Investigator Toolkit
- Data Discovery
- Analysis, Visualization
- Data Management
- Java Library
- Python Library
- CLI Tools
- REST URLs

Service Specifications

Member Nodes
- Service Interfaces
  - Read
  - Write
  - Translate to MN Internals
- Data Repository

Coordinating Nodes
- Service Interfaces
  - Search
  - Register
  - Replicate

  Coordination Layer
  - Identify
  - Preserve
  - Catalog
    - Monitor
  - Object Store
  - Index
Objects in DataONE
Objects and Packaging

CiTO
The Citation Typing Ontology

System Metadata
Identifier
Checksum
Object Type
Size
Access Control
Replication Policy

Science Metadata

documents
isDocumentedBy

aggregatedBy

Data

Sys MD

Resource Map

OAI-ORE
Open Archives Initiative Object Reuse and Exchange

Sys MD
Objects in DataONE
Everything on Member Nodes

Science Metadata

Data

Resource Map

Sys MD

Sys MD

Sys MD

MNs
Objects in DataONE

Synchronization to Coordinating Nodes

Sys MD → Science Metadata → Resource Map → Data → Sys MD

Sys MD → Resource Map → CNs → MNs → Harvest → MNs
Objects in DataONE
Synchronization to Coordinating Nodes
Objects in DataONE

Synchronization to Coordinating Nodes

1. Science Metadata to Data
2. Data to Resource Map
3. Resource Map to CNs
4. CNs to MNs
5. MNs to Data
6. Data to Sys MD
7. Sys MD to Science Metadata

Get Content
Coordinating Node Processing
Indexing for Discovery
DataONE MN-CN Synchronization (generalized)

MN

CN

CN.queue

CN.store

CN.index_queue

CN.index

listObjects(from_LHD, to_now)

LHD = Last Harvest Date

PID

addToQueue(PIDs)

getSysMeta(PID)

PID.Sys.Meta

Internal verification of system metadata

save(systemMetadata)

loop [for each PID in Hz queue]

next(PID)

alt [if Science Metadata or Resource Map]

getObject(PID)

PID.Object

Only Science Metadata and Resource Maps are retrieved to the CNs.

setLHD(most_recent_PID)

PID_Object

add(PID)

Science Metadata and Resource Maps are checked for schema compliance.

loop [for each PID in index_queue]

index(PID_SystemMetadata)

index(PID_Object)

index(Resource_Map)
DataONE MN-CN Synchronization (generalized)

1. listObjects(from_LHD, to_now)

2. getSysMeta(PID)
   - PID_SYS_Meta
   - addToQueue(PIDs)
   - next(PID)
   - internal verification of system metadata
   - save(systemMetadata)

3. getObject(PID)
   - PID_Object
   - setLHD(most_recent_PID)
   - add(PID)
   - Only Science Metadata and Resource Maps are retrieved to the CNs.
   - Science Metadata and Resource Maps are checked for schema compliance.

4. loop
   - [for each PID in index_queue]
   - index(PID_SYS_Meta)
   - index(PID_Object)
   - index(Resource_Map)
1. **List Objects** Failure.
   PIDs never retrieved from Member Node

- Member Node not registered or not approved
- Member Node document not listing required services
- Network failure
- Authentication or authorization failure
- Member Node response timed out
- Member Node not handling the requested time slice as expected
- Response not in expected format
2. Get System Metadata Failure.
Content does not appear in \texttt{CN.getSystemMetadata, CN.get}

- Network failure
- Member Node response timed out
- Authentication or authorization failure
- Invalid System Metadata
- Duplicate identifier (CN will compare checksums)

Check for \texttt{SynchronizationFailed} messages on Member Node
Try the v2 method \texttt{CN.synchronize()} to resubmit a PID.
   Content does not appear in `CN.getSystemMetadata`, `CN.get`
   - Network failure
   - Member Node response timed out
   - Authentication or authorization failure
   - Unexpected content type

4. Content Validation Failure
   - Bad content transfer (unlikely)
   - Schema not registered

Check for `SynchronizationFailed` messages on Member Node
Try the v2 method `CN.synchronize()` to resubmit a PID.
5. Indexing Failure.
Apparent when content not discoverable in search index.

- Content not replicated between Coordinating Nodes
- No indexer rules available
- Indexer rule failure
- Incomplete package contents
- Unexpected content formatting for parser rules

CN administrators check CN replication for content.
Technical Resources

Education
- purl.dataone.org/architecture
- ...architecture/design/UseCases/06_uc.html
- dataone.org/previous-webinars

Communications
- mnforum@dataone.org
- developers@dataone.org
- irc.ecoinformatics.org #dataone

Sources
- repository.dataone.org/software/cicore
- github.com/DataONEorg

Tools
- search.dataone.org... search-dev... search-sandbox... search-stage...
- cn.dataone.org/cn/v2/diag/subject /sysmeta /object
- Browser, curl, Python, Java libraries
- examples.dataone.org