

# DataONE Member Node Forum

---

## Special Topic: Archival of Content

Thursday, 1 October 2015



# DataONE MNs: Archival of Content

---

- What is archival?
- Why archive?
- How do you archive content?





# Baseline MN Implementation Requirements

---

All Member Nodes, regardless of Tier, share the same baseline requirement of managing their object collection such that:

- 1. All objects are identified by a globally unique identifier (aka PID).**
- 2. The original bytes of submitted objects are always available using the original identifier.**
- 3. Content access policies are respected.**
- 4. Updates are treated as a new objects.**
- 5. Voluntary deletes are handled as “archive” actions.**
- 6. Compulsory (or administrative) deletes, are coordinated and executed via the DataONE Coordinating Nodes.**



# Baseline MN Implementation Requirements

---

## **6. Compulsory (or administrative) deletes, are coordinated and executed via the DataONE Coordinating Nodes.**

These are true deletes where the content is removed from all nodes, and is done ONLY in cases when illegal or inappropriate content is discovered and needs to be removed. Because content may be replicated to other nodes, a Member Node performing a compulsory delete must notify DataONE of the action to ensure any replicas are also removed. Otherwise the content may remain discoverable and accessible at other nodes.



# Baseline MN Implementation Requirements

---

## **5. Voluntary deletes are handled as archive actions.**

Objects published to DataONE are not deleted, but archived. This helps ensure DataONE's preservation requirements are upheld. An archived object has its *archived* field set to *true* in its system metadata, and the object remains accessible by its identifier. This applies to all objects ever exposed through the DataONE API regardless of how the content is added to a Member Node. For example, if a Member Node manages their collection with external tools, the system should behave such that deleted content is treated as archived, rather than actually deleted.



# DataONE MNs: What is archival?

---

Archival is the storage of documents, data, etc. in an archive or other repository.

Archival in DataONE refers to the action of ensuring data is no longer discoverable via a general search (such as ONEMercury) but is still retrievable by a user who has the PID for that data object/package.

Archival supports DataONE's principle of ***providing persistent access to content.***



# DataONE MNs: Why archive?

---

The content owner requests that their content be removed from your repository (and thus, from DataONE).

You may have discovered a flaw in the data (from data collection, analysis, etc.).

Your repository is closing up shop.

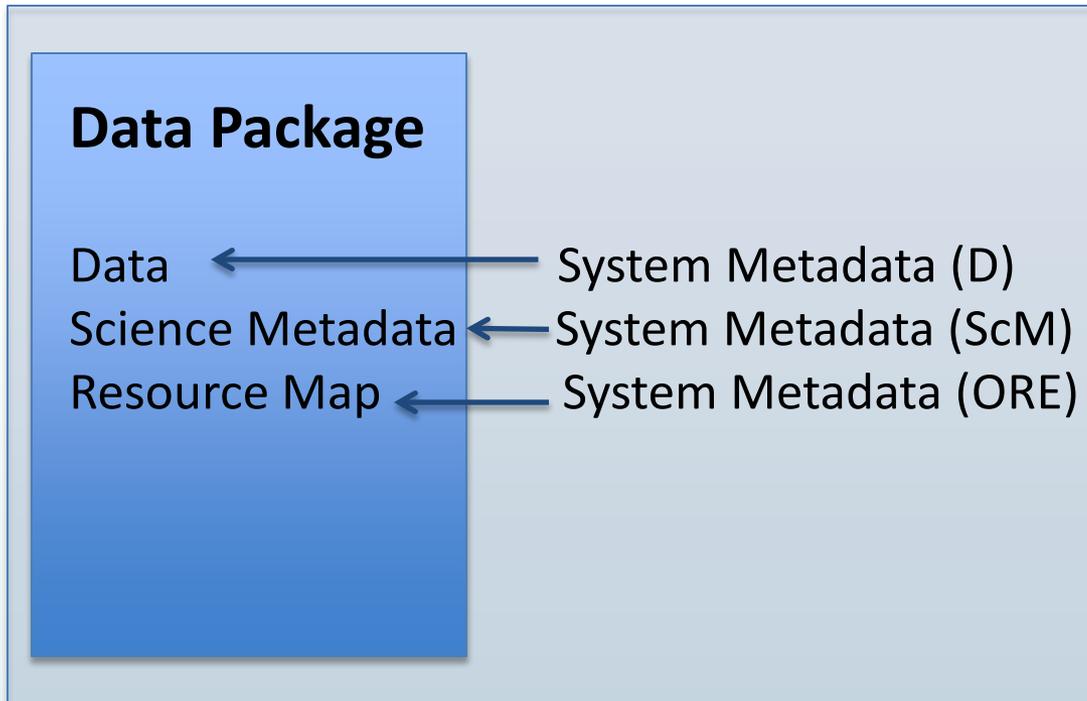
Other reasons???



# DataONE MNs: How to archive?

---

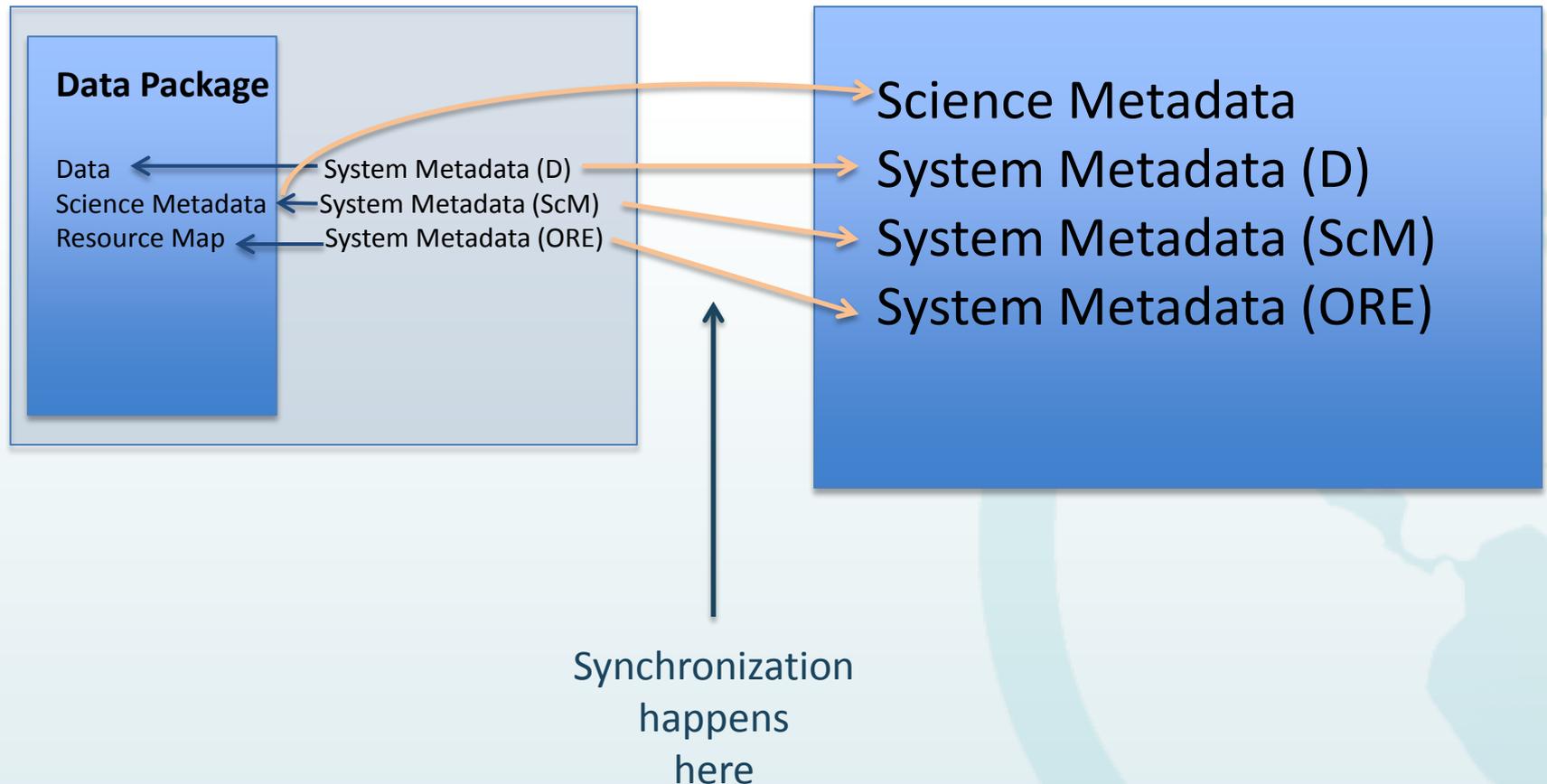
On the Member Node:



# DataONE MNs: How to archive?

On the Member Node:

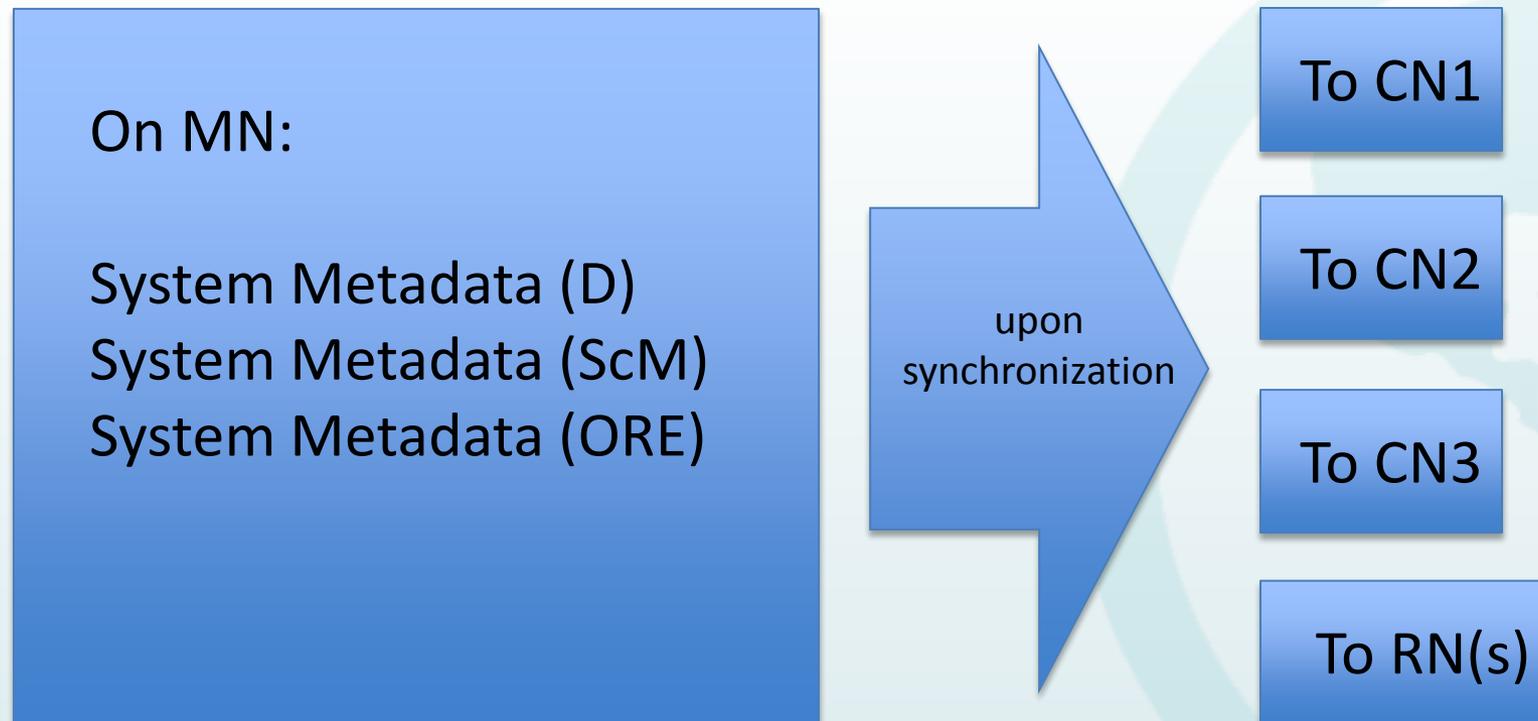
On the Coordinating Node:



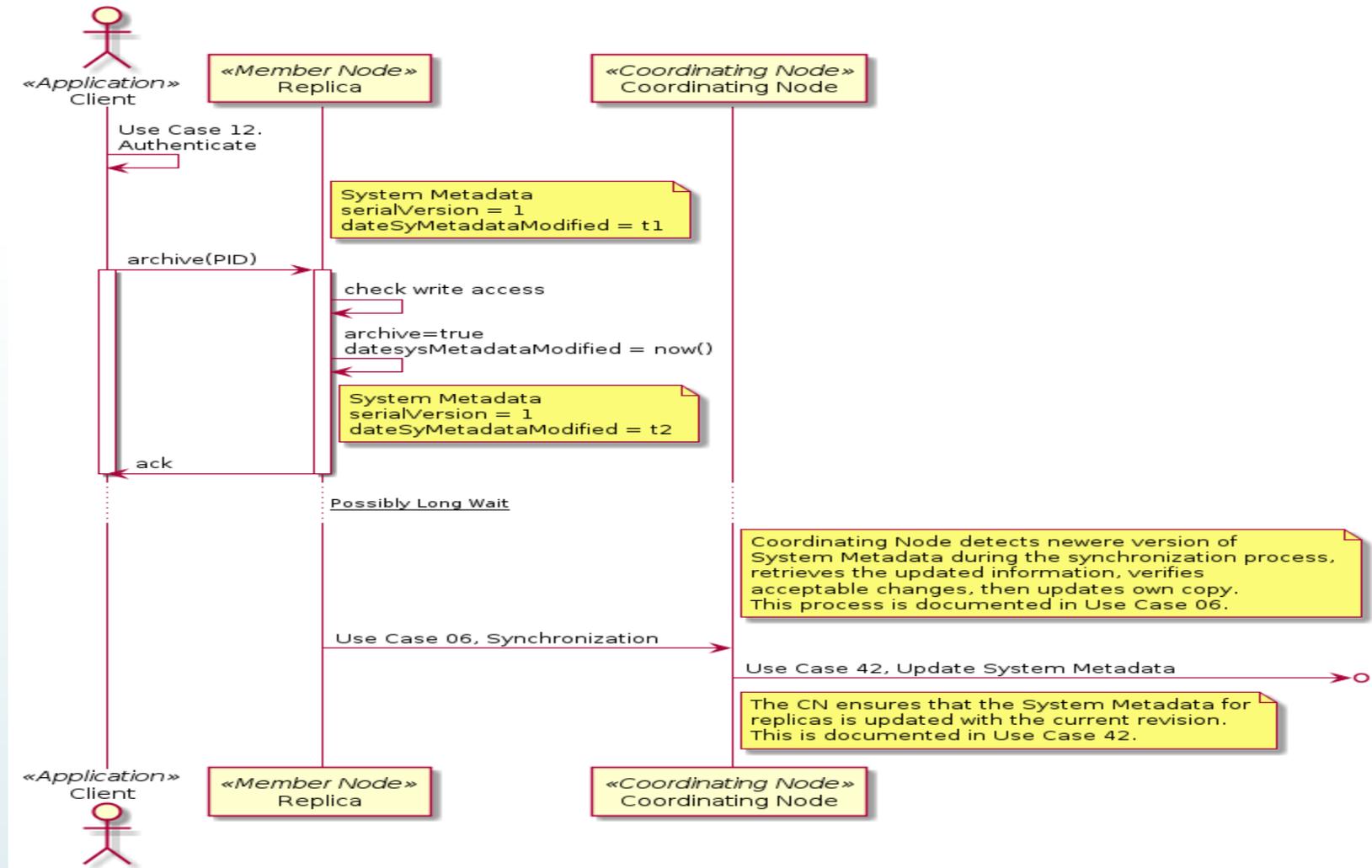
# DataONE MNs: How to archive?

The Member Node Operator (or data owner) will change archived=true in the system metadata (for all three: data, science metadata, resource map) on the Member Node:

See documentation here: [https://releases.dataone.org/online/api-documentation-v1.2.0/apis/MN\\_APIs.html#MNStorage.archive](https://releases.dataone.org/online/api-documentation-v1.2.0/apis/MN_APIs.html#MNStorage.archive)



# DataONE MNs: See Use Case 41



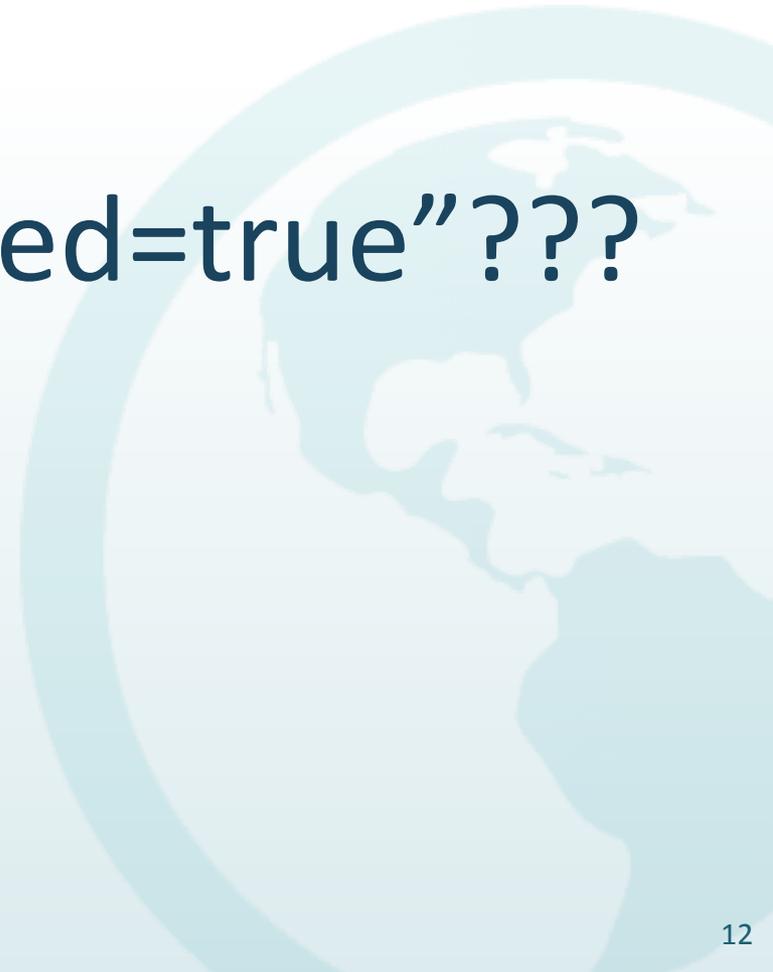


# DataONE MNs: How to archive?

---

The real question is....

How do I set “archived=true”???



# An example using curl

---

Using curl to archive an object. ANODE can be any Member or Coordinating Node that has a copy of the object:

```
# The member node base URL ANODE="https://my.mn.org/base/url"
# A client certificate with a subject that has write permission on the object
CERT="my_client_certificate.pem"
# The identifier of the object to be archived. Note that the PID will need # to
# be URL path encoded in this example. PID="object.identifier"
# Invoke the archive operation on the identifier
curl -E ${CERT} -X PUT "$NODE/v1/archive/${PID}"
```



# Discussion

---

Managing content discoverability/access:

- Delete (very rare; content no longer discoverable or retrievable, and PID cannot be reused)
- Archive (not discoverable, but retrievable; PID cannot be reused)
- Change access policies (controls discoverability and access)

Versioning of content – when might it be appropriate to archive prior versions of content?



# Discussion

---

What happens when a repository (which has exposed its content via DataONE) must end its operations?

Ideally, the closing Member Node's content should be replicated at another DataONE Member Node. That target MN should become the authoritative MN for the former MN's content. Then, depending on policy decisions at the original MN, the content may be archived from the target MN.

This ensures preservation of the original MN's content while allowing the original MN and its base repository to manage their operations appropriately.

# DataONE Member Node Forum

---

- The Member Node Forum meets bi-weekly on Thursdays at 2pm US Eastern Time
  - <https://www1.gotomeeting.com/join/726137377>\*
- The intent is to provide an opportunity for current and developing Member Nodes to connect, discover and share solutions to problems, and build community.
- Points of Contact:
  - Laura Moyers, University of Tennessee ([lmoyers1@utk.edu](mailto:lmoyers1@utk.edu))
  - Mark Servilla, University of New Mexico ([mark.servilla@gmail.com](mailto:mark.servilla@gmail.com))

\* GoToMeeting information current as of 10/1/2015