Provenance Use Cases

Use Case 41 - Scientists can provide tracking information about derived products

Revisions

2014-08-25-01 Draft
An editable version of this document can be found at:
https://docs.google.com/a/nceas.ucsb.edu/document/d/1I0RFtyDo_Pa2tYHPP210Rcxuf62M7M3CCteozXRs6Xs/edit

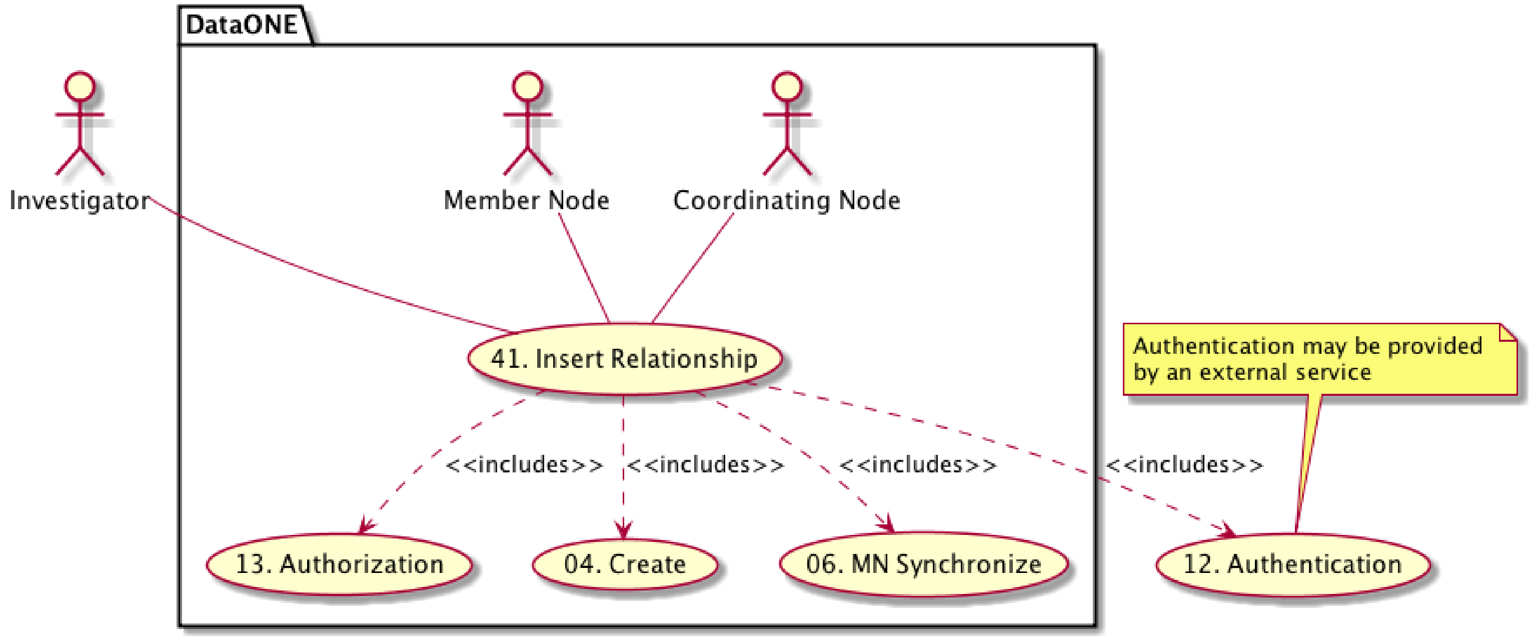
Goal

In DataONE-enabled client software, investigators can easily provide tracking information as they create new products from existing data files.

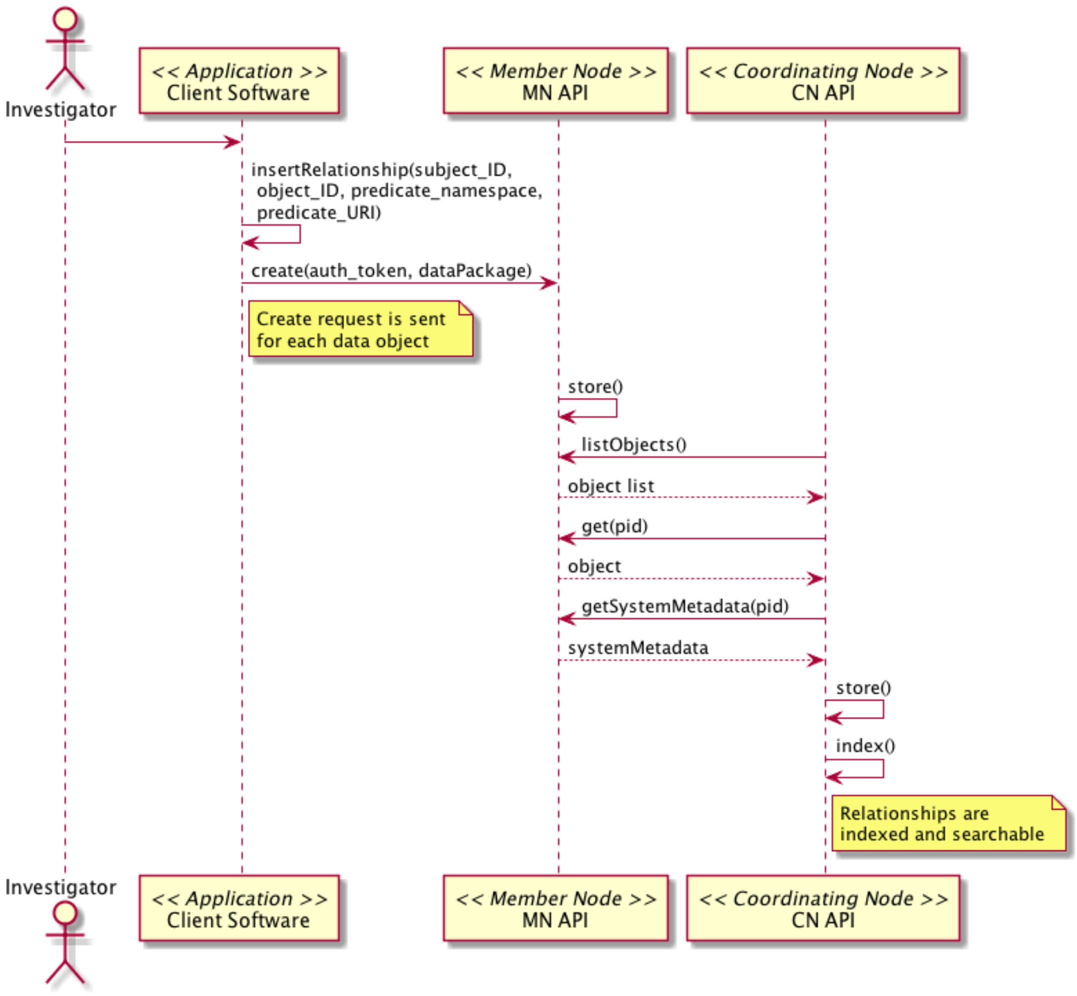
Summary

Investigators can upload derived datasets to a Member Node and provide traceable links to the primary resources used to create them.

Use case diagram



Sequence diagram



Actors

- Investigator
- Client software
- Member Node
- Coordinating Node

Preconditions

- The primary resource dataset needs to be registered on the Member Node.
- The Investigator needs write access to a Member Node.
- The client software must be DataONE-enabled and provenance-aware.

Triggers

Postconditions

- The derived datasets are stored on the Member Node
- The data package includes formal links between the primary and derived datasets

Notes

Use Case 42 - Scientists can examine original dataset(s) that were used to create a synthetic dataset found in DataONE.

Revisions

2014-08-25 Draft

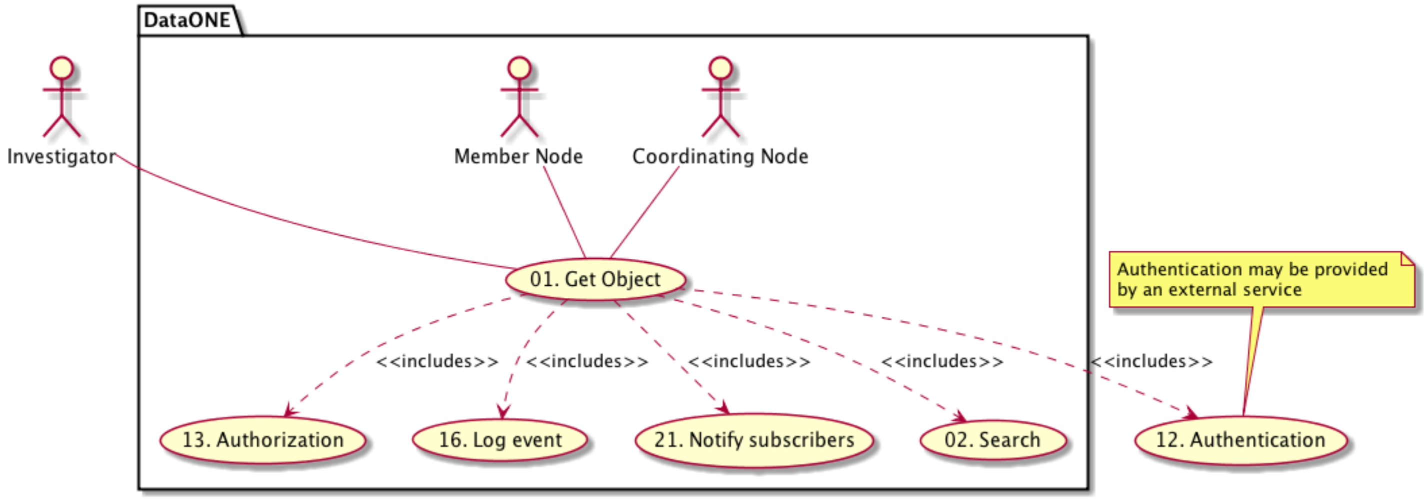
Goal

Scientists examining a synthetic dataset in DataONE are able to determine which dataset from a DataONE Member Node was used in the synthesis and can examine that dataset.

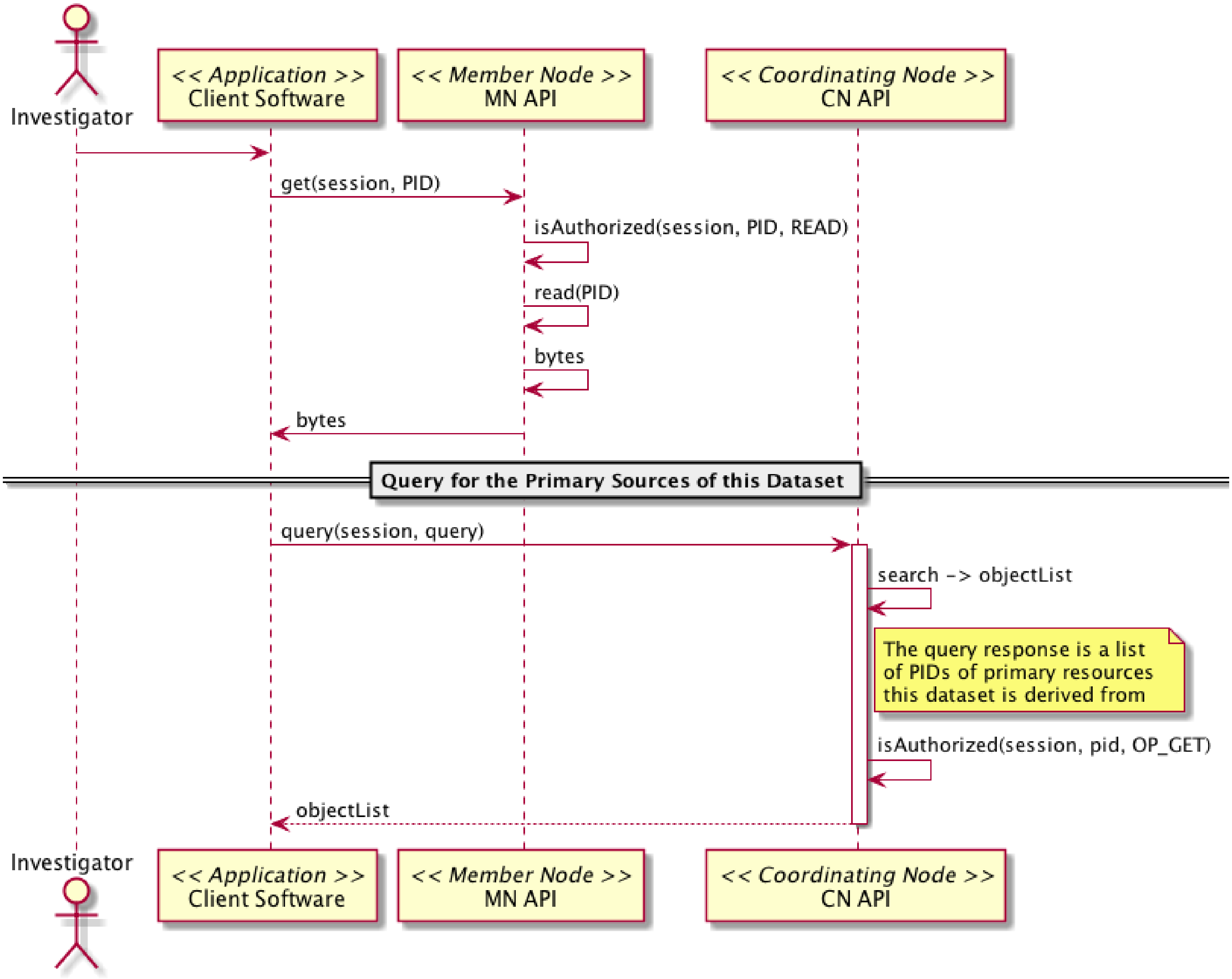
Summary

A scientist that has searched for data relevant to their studies has found a synthetic dataset in DataONE. They are able to view the relationships between the original and derived datasets, and can download the originals for examination.

Use case diagram



Sequence diagram



Actors

- Investigator
- Client software
- Member Node
- Coordinating Node

Preconditions

- The scientist who uploaded the synthetic dataset to a Member Node provided provenance information.
- The client software and user interface must be DataONE-enabled and provenance-aware.
- The synthetic and primary dataset(s) have been indexed by the DataONE Coordinating Nodes
- The synthetic and primary dataset(s) need to be accessible to that authenticated user or the public.

Triggers

Postconditions

- The scientist can understand the provenance links and can download the primary and derived datasets.

Notes

Use Case 43 - Scientists can discover synthetic research that uses their dataset within DataONE.

Revisions

2014-08-25 Draft

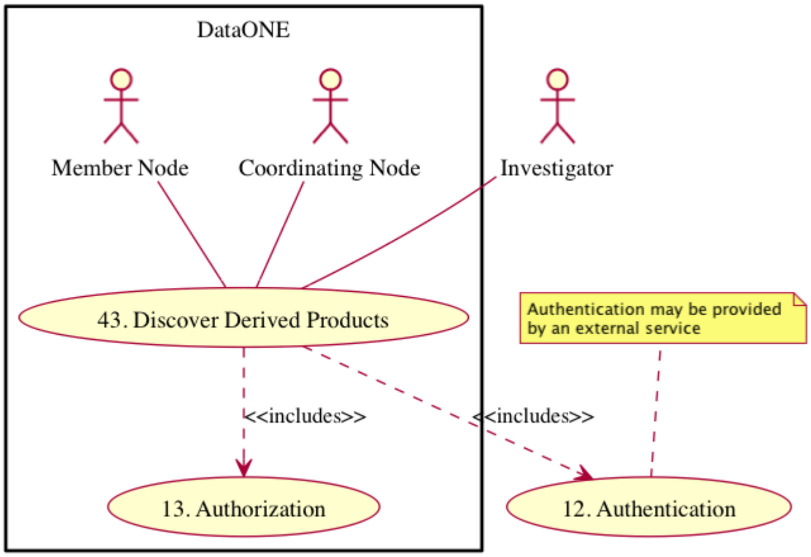
Goal

To provide a traceable link to derived works for each dataset.

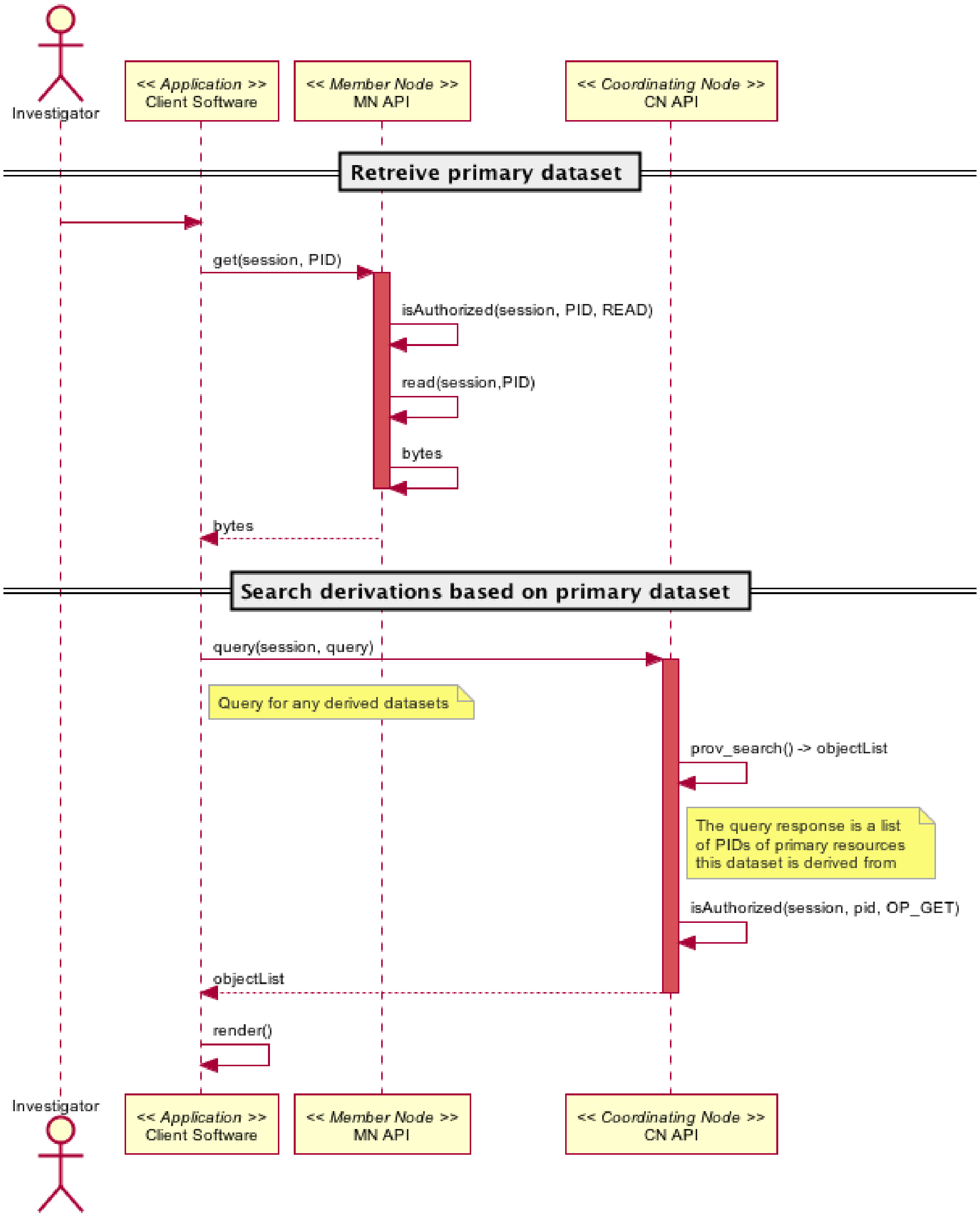
Summary

A scientist that has uploaded their dataset to DataONE and has allowed derived works in their intellectual rights statement can view and understand which derived works have used their dataset.

Use case diagram



Sequence diagram



Actors

- Investigator
- Client software
- Member Node
- Coordinating Node

Preconditions

- The client software and user interface must be DataONE-enabled and provenance-aware.
- The synthetic and primary dataset(s) have been indexed by the DataONE Coordinating Nodes.
- The scientist who uploaded the synthetic dataset to a Member Node provided provenance information.
- The intellectual rights of the primary dataset allows for derived works.

Triggers

Postconditions

- DataONE users can examine a list of derived works of each dataset.

Notes

Use Case 44 - Scientists reviewing data tables or figures generated by a script can examine and rerun the script using the same input data.

Revisions

2014-08-25 Draft

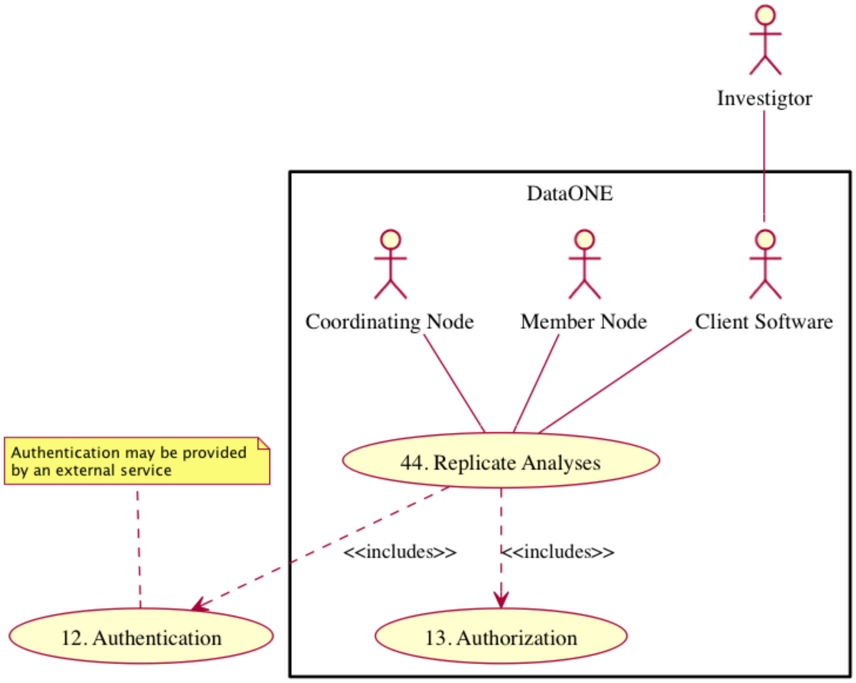
Goal

To assist reproducible science by providing a link between script or models, the input data used, and the generated output.

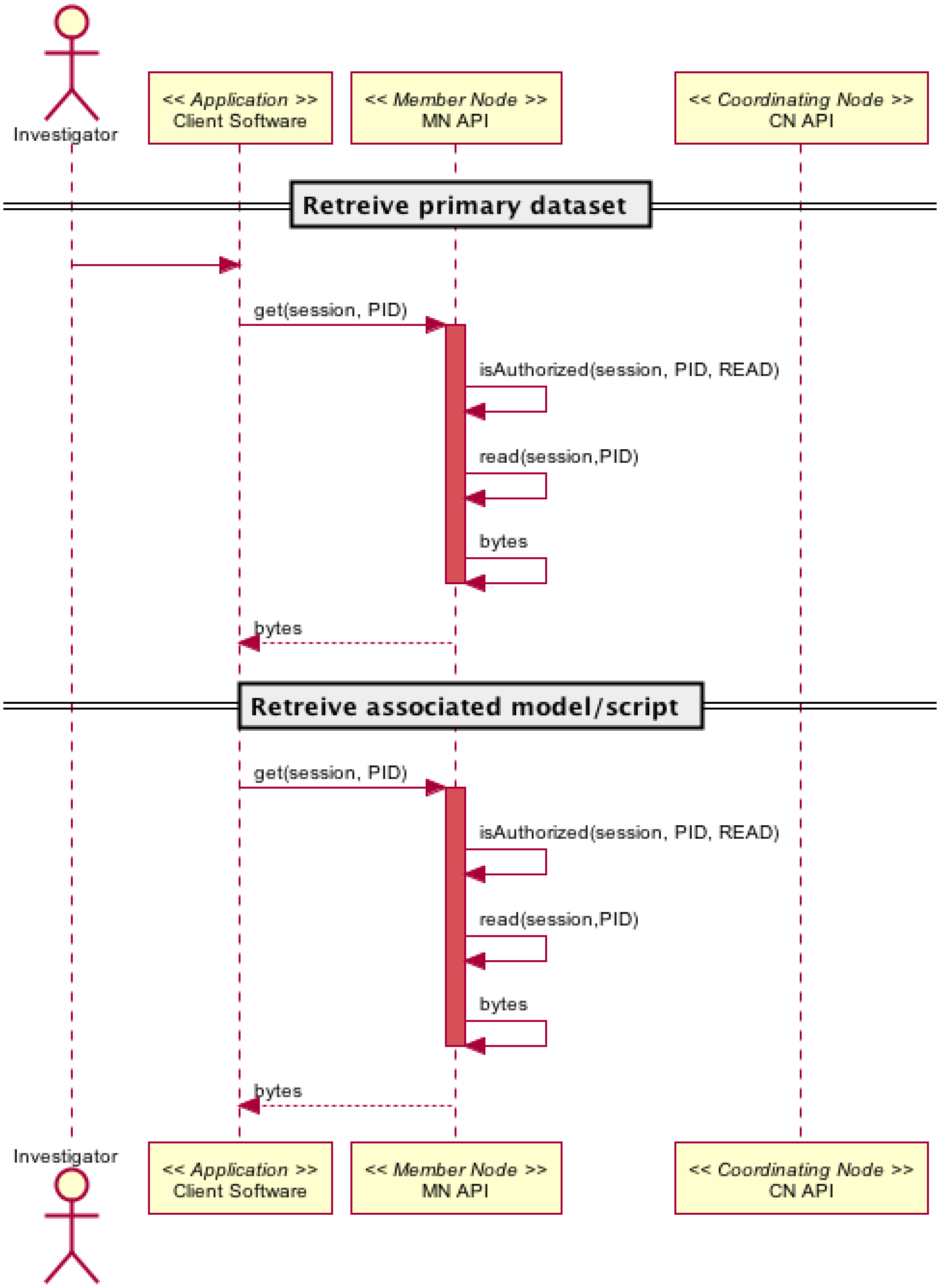
Summary

A scientist reviewing a data table or figure in DataONE can discover the script or model that was used to generate it. The scientist can subsequently download and rerun the script to reproduce the same results as the original run.

Use case diagram



Sequence diagram



Actors

- Investigator
- Client software
- Member Node
- Coordinating Node

Preconditions

- The client software and user interface must be DataONE-enabled and provenance-aware.
- The dataset has been indexed by the DataONE Coordinating Nodes.
- The data package includes a model or script and its product.
- The data package has a provenance link to the input data.
- The scientist must have access to the same version of scientific analysis software that was used to generate the table or figures.
- The input parameters to the model or script need to be provided in order to produce the same output.

Triggers

Postconditions

- A scientist can review a script or model for quality.
- A scientist can rerun the script to reproduce the same results.

Notes

Use Case Implementation Examples

- **An R Client example of Use Case 41 (Scientists can provide tracking information about derived products):**

```
(… create DataONE data objects and a DataONE data package…)

insertRelationship(data.package, id.result, c(id.script),
"http://www.w3.org/ns/prov", "http://www.w3.org/ns/prov#wasGeneratedBy")

insertRelationship(data.package, id.script, c(id.data), "http://www.w3.org/ns/prov",
"http://www.w3.org/ns/prov#used")

insertRelationship(data.package, id.data, c(id.data2, id.data3),
"http://www.w3.org/ns/prov", "http://www.w3.org/ns/prov#wasDerivedFrom")
```

```
createDataPackage(d1client, data.package)
```

A full test script is available here:
https://github.com/DataONEorg/rdataone/blob/master/tests/Test-Insert-Relationship-With-rdataone.R

- **User interface mockups of Use Cases 42, 43, and 44 are in PDF format here:**

  Science Metadata view
  https://github.com/DataONEorg/sem-prov-design/blob/master/docs/use-cases/images/metadata_view_with_use_cases.pdf?raw=true

  Data search view
  https://github.com/DataONEorg/sem-prov-design/blob/master/docs/use-cases/images/data_search_with_use_case.pdf?raw=true