# Infrastructure - Story #3262

## Object provenance needs to be supported across components

2012-09-25 16:59 - Chris Jones

| | | | |
|---|---|---|---|
| **Status:** | Closed | **Start date:** | |
| **Priority:** | Normal | **Due date:** | |
| **Assignee:** | Chris Jones | **% Done:** | 100% |
| **Category:** | | **Estimated time:** | 0.00 hour |
| **Target version:** | | | |
| **Story Points:** | | | |

**Description**

One of the goals within DataONE is to be able to support reproducible science. The "Provenance Working Group": http://www.dataone.org/working_groups/scientific-workflows-and-provenance-working-group (led by Bertram Ludascher and Paolo Missier) has been working toward this end by developing a provenance model that traces derivations of objects through scientific workflow systems (VisTrails, Kepler, etc.). The model they are working on is now called D-PROV (used to be D-OPM), and it strives to be compatible (by extension) with the more generic W3C PROV model. The working group has a use case where user Alice has a dataset D1 that is processed through a workflow WF1 to produce a derivative dataset D2 with a provenance trace Pr1. These products are then uploaded to the DataONE system along with their system and science metadata (SM1, MD1), linked together as a collection using a resource map RM1. A second user, Bob, wants to find datasets that were derived by the WF1 workflow. Once finding D2 in the DataONE system, Bob then wants to use the D2 dataset as an input to a new workflow WF2 to produce a new derived dataset, D3. The data (D3), science metadata (MD3), system metadata (SM3), and provenance (Pr2) artifacts produced during WF2 are then uploaded to the DataONE system as a new collection using a resource map (RM2).

A serialization of the provenance traces produced by the workflow engine (in the use case, VisTrails) is needed. This trace will (for now) be an RDF/XML document compatible with the W3C PROV model, using D-PROV extensions that are specific to provenance concepts relating to environmental science (as opposed to, say, provenance associated with corporate mergers, etc.). Two of the VisTrails developers (David Koop and Fernando Seabra Chirigati) will produce the provenance serialization, and will need to be able to easily insert the trace into the resource maps (RM1, RM2) for the two collections. This will require the use of new predicates in the resource map that allow for triples that "provide provenance information for" the dataset in the data package. The python and java libclient tools will need methods that allow for the insertion of these provenance triples.

To enable enhanced search based on provenance information, we need to initially create indexed attributes in the CN Solr index that are parsed out of the provenance serializations. These attributes will be similar to "derivedBy" and "derives", or others in the provenance model we think are most useful for search.

The ONEMercury interface needs to be modified to enable search based on provenance attributes. This will likely be accomplished in a development version of the the code, and so a development environment with at least a single CN and an MN accessible by the ProvWG members needs to be deployed so they can reliably interact with the DataONE API.

### History

**#1 - 2012-10-05 14:21 - Chris Jones**

*- Milestone changed from None to CCI-1.3*

*- Due date set to 2012-11-10*

*- Target version set to Sprint-2012.44-Block.6.2*

*- translation missing: en.field_remaining_hours set to 0.0*

**#2 - 2012-12-12 18:41 - Chris Jones**

*- Target version changed from Sprint-2012.44-Block.6.2 to 2013.2-Block.1.1*

*- Due date changed from 2012-11-10 to 2013-01-19*

**#3 - 2013-03-01 18:56 - Chris Jones**

*- Due date changed from 2013-01-19 to 2013-03-30*

*- Target version changed from 2013.2-Block.1.1 to 2013.12-Block.2.2*

**#4 - 2013-05-31 16:02 - Chris Jones**

*- Due date changed from 2013-03-30 to 2013-06-22*

*- Target version changed from 2013.12-Block.2.2 to 2013.24-Block.3.4*

**#5 - 2013-08-02 15:23 - Dave Vieglais**

*- Due date changed from 2013-06-22 to 2013-08-24*

*- Target version changed from 2013.24-Block.3.4 to 2013.33-Block.4.4*

**#6 - 2014-01-06 17:51 - Chris Jones**

*- Due date changed from 2013-08-24 to 2014-02-15*

*- Target version changed from 2013.33-Block.4.4 to 2014.6-Block.1.3*

**#7 - 2014-03-14 18:28 - Chris Jones**

*- Due date deleted (2014-02-15)*

*- Target version deleted (2014.6-Block.1.3)*

*- Start date deleted (2012-09-25)*

**#8 - 2018-01-17 20:27 - Dave Vieglais**

*- % Done changed from 0 to 100*

*- Status changed from New to Closed*