Working Group on Provenance Solutions

Research Data Provenance IG

Last updated Tue, Jan 31, 2017 12:15:06 AM

# Charter

## Overview

Tracking provenance for research data is vital to science and scholarship, providing answers to common questions researchers and institutions pose when sharing and exchanging data.

The tasks for this Working Group focus on finding, detailing and recommending best practices for provenance representation and management.

This group will conduct its work in the manner of a business analysis task: identifying business needs and determining solutions to business problems. Since RDA WGs are not themselves research groups (rather groups of researchers and research agencies), this group will look for existing practice and re-present that for use rather than generate new practice.

The six activity areas of the Working Group will be:

1. Common provenance Use Cases
2. Provenance patterns
3. Sharing provenance
4. Strategies for enterprise provenance management
5. Tools for provenance
6. Provenance data collections

## Deliverables

The deliverables for this Working Group are separated into three time-based cohorts as below. Short-term goals are mostly about seeking existing practice. Medium-term about determining possible output forms for the activity areas. Long-term about delivering those outputs and After-term about ensuring continuation of output custodianship, where required.

### Medium-term (M12)

1. A provenance use case recording system.
2. An initial collection of provenance use cases, elicited from other interest groups and working groups.
3. First documented provenance patterns generalised from use cases.
4. A report on investigation of provenance sharing implementations.
5. A review of existing enterprise provenance management implementations.
6. A listings of provenance tools compiled from interviews with RDA members and the provenance research community.
7. A directory of open and non-open provenance data collections.

### Long-term (M18)

1. A taxonomy for provenance use cases.
2. Recommendations for aligning new use cases with provenance patterns.
3. Lessons for provenance sharing and enterprise management implementations.
4. A synthesis and critical comparison of community recommendations for provenance tool custodianship.
5. A summary of best practice principles for provenance data collection stewardship.

### After-term (M18+)

1. A sustainability plan for ongoing tool and data collection custodianship.

# Value Proposition

Effective provenance management is sought by many members of the RDA and wider science data community. We propose a working group to help those members adopt existing provenance management practice. This help will be in the form of documenting provenance use cases: centralising a list of them and generalising them to reveal common ones; documenting existing technical and business processes for provenance management, assisting organisations with sharing provenance and listing existing sources of real provenance information.

We propose a working group on *provenance solution patterns.*

* The patterns should relate to core RDA interests, perhaps data/data and data/people relationships.
* Provenance vocabularies offer a level of generality/specificity that address what we perceive to be implementation gaps.
* Our goal: constructive engagement with and response to published RDA recommendations.

### WG activity points

1. **Provenance Use Cases**
   * Use Cases for provenance data or systems are often articulated in terms understood by a particular community however in our group's experience, many provenance Use Cases are differently worded instances of general Cases.
   * The establishment of a published set of UCs would allow people to compare their UCs with known UCs for which recommended implementations and other patterns may already be known.
2. **Provenance patterns**
   * Some ways of doing things in provenance are better than others. This activity is to generate provenance design patterns (for any provenance task such as representation, transmission, use etc.) perhaps in response to a series of provenance use cases that we would generate.
   * The patterns should relate to core RDA interests, perhaps data/data and data/people relationships.
3. **Sharing provenance**
   * This may only be a single class of provenance use cases but it is one that is less maturely answered by the provenance research community than, say, provenance representation. This activity might be to generate requirements for the research community to answer or perhaps find that no more research is needed for sensible recommendations for provenance sharing.
4. **Strategies for implementation**
   * Some provenance use cases apply to whole organisations (or consortia) and some organisations (or consortia) may already have experience in implementing solutions to them. This activity will list such use cases and seek descriptions of implemented or proposed solutions from members.
5. **Tools for provenance**
   * In addition to several well-known provenance conceptual models, there are tools to assist with the management of provenance. We will list those tools with comparisons in relation to RDA interests (perhaps taken from IG and other WG members).

# Engagement

In addition to serving the RDA community directly, this Working Group aims to serve the immediate interests of existing RDA groups. Provenance is foundational to many other RDA groups' activity and thus maximal impact on the RDA community can be achieved by aligning and assisting work in existing groups. Therefore this working group will engage heavily with other groups and source its primary requirements and exemplars from other groups. Examples of intersections we believe will be productive include the following:

Publishing Data Workflows WG:

Interest in workflow persistence and quality control, data deposit and citation, reference models and implementation.

Dynamic Data Citation WG:

Interest in a conceptual model for citation fidelity despite changes over time.

PID Information Types:

The Use Case "A.10 Provenance tracing."

Reproducibility IG:

The role of provenance models in support of replication.

PID IG:

Requirements for PIDs to maintain provenance content.

Archives and Records Professionals for Research Data IG:

Need for semantic understanding of archived material.

Data Discovery IG:

Upper ontology elements relevant to data discovery.

Preservation e-Infrastructure IG:

Semantic content of preserved data holdings.

# Work Plan

### Timeline

* **Dec 16**: Identify initial set of focus areas and discuss.
* **Feb 17**: Draft case statements distributed.
* **Apr 17**: Discuss draft case statements and formation of WGs at Plenary 9.
* **May 17**: Finalize and circulate case statements for WGs.
* **May - Oct 17**: Short-term goals.
* **Sep 17**: Meeting WGs and summary of activity at Plenary 10.
* **Nov 17 - Apr 18**: Medium-term goals.
* **Apr 18**: Group health check Plenary 11.
* **May 18 - Oct 18**: Long-term goals.
* **Sep 17**: Final group Plenary 12.
* **October 18+**: After-term goals.

# Initial Membership

Membership will be sought from the Provenance IG and supplemented with a call to both other RDA groups and known non-RDA provenance communities, such as the provenance research community.

Since this group's work is likely to be highly relevant to, or even directed by, other WGs, it may be sensible to have other WG members attend this group's meetings either in a liaison role or as members in their own right.