Working Group on Provenance Patterns

Research Data Provenance IG

Last updated Fri 03/03/2017 13:44:28.27

# 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 design 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 design 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 design 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 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. **Common 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. It will also allow people to consider provenance UCs posed by others that may be of future iterest to them.
2. **Provenance design 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 enterprise provenance management**
   * 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).
   * We will also seek to establish a mechanism to keep these tool lists up-to-date beyond the life of the WG.
6. **Provenance data collections**
   * The provenance research community knows that provenance ontologies and tools are used due to communication with them and research papers but are only anecdotally aware of many current provenance datasets (i.e. whole datasets of provenance information) and have not yet counted datasets linking to standardised provenance information. In order to know the state of operational system's adoption of provenance models and in order to provide access to public provenance data for both education and actual use, we will list as many current provenance datasets as we can find owned by RDA members and others and catalogues of datasets linking to standardised provenance information.
   * We will also seek to establish a mechanism to keep this listing up-to-date beyond the life of the WG.

# 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.

# Adoption

## Other RDA groups

This WG proposal is engagement-driven, primarily with other RDA groups, thus it is in them that we expect to see initial adoption.

Where another RDA group presents us with a provenance use case, we hope to either:

* associate that use case with a generic use case and a thus a pre-made generic resolution
* provide a direct provenance pattern-based resolution directly

In either case, we hope to promote a pattern that the RDA group will adopt and promote to its members.

## Prov WG member institutions

Most current RDA Provenance IG members are likely to become Prov WG members for their interest in provenance is adoption and that too is the WG's goal. It is likely that outputs from this group, having been generated by its members in response to their direct needs and similar needs of other RDA groups, will therefore be fed back into their home institutions for adoption there.

## Non-RDA groups

The international provenance research community is in contact with many potential consumers of provenance patterns due to their profile as experts on provenance. The potential consumers don't always receive the advice they are seeking due to differences in their aims and that of the research community's. The research community needs to push the provenance envelope forward and not dwell on previous work, even when that work may contain patterns perfectly suited to the potential consumers' needs.

The Provenance IG has and the WG will have, if membership proceeds as expected, good contacts with the international provenance research community with several IG members having made substantial contributions to provenance research initiatives such as the [Open Provenance Model](http://openprovenance.org/), [ProvONE](http://vcvcomputing.com/provone/provone.html), the [PROV W3C standard](https://www.w3.org/TR/prov-overview/) and having presented at many recent provenance conferences such as [IPAW 2014](http://www.ipaw.info/ipaw14) & [2016](http://www.ipaw.info/ipaw16), [TaPP 2015](http://workshops.inf.ed.ac.uk/tapp2015/), [TaPP 2017](http://batesa.web.engr.illinois.edu/tapp17/) *(coming)*, [Classification Soc 2017](https://tcs.wildapricot.org/Meeting), [ISWC 2016](http://iswc2016.semanticweb.org/) and [IDCC 2017](http://www.dcc.ac.uk/events/idcc17).

The full listing of the IG's involvement in provenance conferences is available on the RDA Prov IG wiki:

* <https://www.rd-alliance.org/group/research-data-provenance/wiki/provenance-conferences>

## Continued adoption

Some of the outputs from this WG are targeted at continued adoption over time. This proposal includes a deliverable for a "sustainability plan for ongoing tool and data collection custodianship" after having initially established a provenance use case database, listings of provenance tools and provenance datasets. Such a plan is currently missing from the international provenance community despite widespread recognition that it would be useful. This was recognised at IPAW 2016 independently of any RDA involvement.

It is expected that at the conclusion of this WG, the current provenance IG will have some role in the custodianship of its outputs.