Creating provenance super graphs using pingbacks

N.J. Car [a] and S. Woodman [b]
[a] *CSIRO Land & Water, Dutton Park QLD, Australia*
[b] *CSIRO Digital Productivity, Clayton, VIC, Australia*
Email: nicholas.car@csiro.au

Abstract: In computer science, provenance information is about a digital object's generation and, like 'provenance' in the art world, is recorded to answer such questions as "who produced what, when?" or "what data was used to make this product?". It is critical for scientific data's transparency of production.

If provenance information is recorded for work done to a dataset, questions such as "which datasets were derived from this dataset?" can be answered. This illustrates the concept of 'forward provenance'. Organisations that supply datasets can benefit from receiving forward provenance information from external agents that use their datasets in many ways. One great benefit in the science community is to determine impact from the publication of a dataset – a key issue for government funded research institutions. Another benefit could be that if errors in a dataset are discovered, a dataset publisher could notify users of that dataset.

A key enabler of effective provenance information sharing is the W3C's PROV Data Model recommendation (http://www.w3.org/TR/prov-dm/) which specifies structure for provenance information representation. PROV is generic, high-level and derived from many precursor provenance information systems such as Proof Markup Language (PML) (https://en.wikipedia.org/wiki/Provenance_Markup_Language) and the Open Provenance Model (OPM) (http://openprovenance.org/). Adherence to PROV ensures provenance information can be understood, at least at some level, by heterogeneous producers and consumers of it. The PROV Ontology (PROV-O) (http://www.w3.org/TR/prov-o/) provides a formalisation of the PROV data model that can be used to record provenance information according to the Web Ontology Language (OWL) (https://en.wikipedia.org/wiki/Web_Ontology_Language) which is thus compatible with the Semantic Web. Provenance information structured according to PROV-O takes the form of RDF graphs which can be stored in graph or other databases or serialised as RDF documents.

Provenance data model standards such as PROV do not directly address mechanisms for provenance discovery and access. PROV-AQ ("Access & Query") is a W3C technical note (http://www.w3.org/TR/provaq/) that specifies how to use standard Internet protocols to obtain information about the provenance of resources on the Web. One part of PROV-AQ specifies mechanisms for provenance 'pingbacks' which can be used to inform parties about use of their data thus communicating forward provenance with data originators. PROV-AQ provides a high-level specification for the process of sending pingbacks and also pingback message content guidelines. It does not describe how or when a pingback generator or receiver would generate, process or use pingback information.

In this presentation, we detail our extensions to the PROMS Server (https://wiki.csiro.au/display/proms) that enable provenance pingbacks within and between organisations. We demonstrate adherence to both PROVAQ, the well-known data description vocabulary DCAT (http://www.w3.org/TR/vocab-dcat/) and formalisms specified in the Data Provider Node Ontology (DPN-O) (http://purl.org/dpn) which provides a generalised architecture for data and data service description.

Specifically, we describe how our extensions to the PROMS Server allow it to:

1. Determine when and for which things pingbacks should be sent;
2. Implement a user-selectable set of strategies that use different approaches to determine reception endpoints for pingback messages;
3. Implement several user-selectable pingback message content options.

Finally, we will discuss how a network of PROMS Servers with these pingback extensions, by allowing the creation of 'super graphs' which are linked provenance graphs within and across organisations, may benefit Australian scientific agencies. We will present our first steps in establishing such a network, including PROMS Server installations using pingbacks across 3 agencies and initiatives.

Keywords: *Provenance, pingback, PROMS Server, PROV*