Provenance Central: More Mileage from Provenance Metadata

Bertram Ludäscher  
UC Davis, USA  
ludaesch@ucdavis.edu

Paolo Missier  
Newcastle University  
UK  
paolo.missier@ncl.ac.uk

Members  
DataONE Provenance Working Group

Metadata Provenance has long been recognized as an important form of metadata that helps to "get more mileage" from data. Provenance describes the lineage and processing history of data, by means of dependency relationships. These have recently been specified in the PROV data model for provenance, a W3C recommendation since June, 2013. The PROV model includes produce/consume relationships, for instance "data product D was generated by activity A", or symmetrically, "A used D", as well as attribution relationships, i.e., "D was attributed to Agent Ag", or "A was associated with Ag". Here an Agent can be a human, a software system, or other entity to which responsibility for data and activities can be attributed.

These specific relationships set provenance apart from other forms of metadata, notably descriptive or preservation metadata, not only semantically but also structurally (a provenance trace is a digraph).

There is a broad expectation within the Information Management community that provenance can be used to better understand and interpret data products, to assess and improve their quality and fitness for use, for debugging and reproducibility of results, etc.

Important sources of provenance generation are scientific workflow systems and other controlled environments and cyberinfrastructure that can be instrumented to capture provenance, including scripting languages like R, popular amongst e-scientists. Examples of scientific workflows include VisTrails [1], Taverna [2], e-Science Central [3], Pegasus [4], Kepler [5], Swift [6], SciCumulus [7], and more [8]. For these systems, a provenance trace that conforms to the standard model can be automatically constructed from a collection of observations of relevant system events. These may be data flowing in/out of a service interface that is part of a workflow, as well as lower-level OS, network, and file system events (the PASS project [9]), or modular provenance generated by multiple components of a distributed system [10].

We are working to develop a provenance management system, which we call PBase. As a starting point, PBase is a graph database where storage and query services are optimized for the specific graph structure of provenance traces. Furthermore, however, we argue that in order to get the most value out of a corpus of traces, PBase must be able to: (i) maintain additional relationships to elements of the systems under observation, (ii) integrate traces both across multiple systems...
and over time, and (iii) support analytical services on top of the query services.

Regarding (i), we are developing D-PROV, an extension to PROV that adds a model of the workflow, or the R program whose execution generates the trace, and connect its structural elements to the runtime elements in the trace itself. This enables more expressive queries to span across program structure and trace.

Trace integration (ii) takes two forms. Firstly, simply having all traces in a (logically) centralized repository allows for queries that cut across multiple experiments ("list all traces in which service S was active, or dataset D was used") and along the timeline ("what has been the most popular service amongst independent workflows over the past three months?"). Secondly, in some cases it may be possible to "logically concatenate" traces that share data products. These traces effectively come to represent a new, virtual experiment [11,12].

Finally, when attribution relationships are also available (establishing responsibility for the traces, their activities, and data products), the queries in (ii) can be extended into analytical, mining services (iii) that elicit implicit social connections amongst the owners of the data. For example, it may be possible to cluster independent researchers based on their common use of the same data products, or of the same sets of services over time. Thus, PBase is a new way of making data and social connections explicit, thus increasing data (re)usability in unprecedented ways.

PBase is being developed by a Provenance Working Group which operates as part of the DataOne data preservation infrastructure for Earth Science Observation (www.dataone.org).

Acknowledgement

DataOne/the U.S. National Science Foundation (Grant #OCI-0830944).

Keywords

Provenance; metadata; data integration; social network analysis.

References


