

ProvStore: A Public Provenance Repository*

Trung Dong Huynh and Luc Moreau

Electronics and Computer Science, University of Southampton
Southampton, SO17 1BJ, United Kingdom
{tdh,l.moreau}@ecs.soton.ac.uk

Abstract. ProvStore is the first online public provenance repository supporting the new PROV standards by W3C. It allows users and applications to store and (optionally) publish the provenance of their data on the Web. Provenance documents can be transformed, visualized, and shared in various serializations, with all the functionality also available to automated applications via a RESTful API (OAuth supported).

1 Provenance Repository

ProvStore (available online at provenance.ecs.soton.ac.uk/store) is the first public repository of provenance supporting the PROV standards for provenance on the Web by the World Wide Web Consortium [MM13]. It provides free accounts to registered users, allowing them to upload and share provenance documents either privately or publicly in various representations (see Figure 1 for an example). Specifically, it supports the Provenance Notation (PROV-N), RDF encoded using the PROV Ontology (PROV-O) in Turtle or TriG formats, PROV-XML, and PROV-JSON [HJK⁺13].

By default, documents submitted to ProvStore are private and can only be accessed by their owners. Document owners, however, can choose to share their documents with others in two ways: making a document *public*, i.e. available to any visitor to ProvStore, or sharing it with specific ProvStore's users. The former is useful for users who want to expose the provenance of their resources (e.g. papers, reports, data sets) to the public; the link to a document on ProvStore can be attached as the provenance URI along with the corresponding resource.¹ In the latter, different access roles can be set to authorized users for fine-grain access control: administrator, editor, contributor, or reader. Except reader, all other roles and the owner can append new provenance bundles to a document after it has been created. It is suitable for sharing provenance between a collaborating team of humans and/or applications (see Section 3 for more information about ProvStore's application programming interface).

* ProvStore was funded by the Engineering and Physical Sciences Research Council (EPSRC) as part of project 'Orchid', grant EP/I011587/1.

¹ See www.w3.org/TR/prov-aq for more information on provenance access and query. Document links on ProvStore support HTTP content negotiation. For example, if the HTTP request specify a header `Accept: application/json`, the PROV-JSON representation of the provenance document will be returned.

The screenshot shows the ProvStore interface for a document titled "primer tutorial". The document is displayed in a code editor with the following PROV-N data:

```

document
  prefix foaf <http://xmlns.com/foaf/0.1/>
  prefix dcterms <http://url.org/dc/terms/>
  prefix ex <http://example/>

  entity(ex:article, [dcterms:title="Crime rises in cities"])
  entity(ex:dataSet1)
  entity(ex:dataSet2)
  entity(ex:composition)
  entity(ex:regionList)
  entity(ex:articleV1)
  entity(ex:articleV2)
  entity(ex:blogEntry)
  entity(ex:chart2)
  entity(ex:chart1)
  activity(ex:illustrate)
  activity(ex:compose)
  activity(ex:compile)
  activity(ex:correct, 2012-03-31T09:21:00, 2012-04-01T15:21:00)
  activity(ex:compile2)
  agent(ex:chartgen, [prov:type="prov:Organization", foaf:name="Chart Generators Inc"])
  agent(ex:derek, [foaf:inbox="mailto:derek@example.org", prov:type="prov:Person", foaf:givenName="Derek"])
  wasGeneratedBy(ex:chart1, ex:compile, 2012-03-02T10:30:00)
  wasGeneratedBy(ex:composition, ex:compose, -)
  wasGeneratedBy(ex:dataSet2, ex:correct, -)
  wasGeneratedBy(ex:chart1, ex:illustrate, -)
  used(ex:illustrate, ex:composition, -)
  used(ex:correct, ex:dataSet1, -)
  used(ex:compose, ex:dataSet1, -, [prov:role="ex:dataToCompose"])
  used(ex:compose, ex:regionList, -, [prov:role="ex:regionsToAggregateBy"])

```

On the right side of the interface, there is a "Metrics" table:

Metrics	
Nodes	16
Edges	20
Components	1
Diameter	7
MFD	
entity → entity	4
entity → activity	3
entity → agent	3
activity → entity	3
activity → activity	2
activity → agent	2

Fig. 1. The screenshot of a ProvStore document.

On each document (Figure 1), users can see its provenance descriptions in PROV-N, along with some statistics about the numbers of assertions. ProvStore also provides a number of provenance network metrics [EHM⁺12] calculated on the graph representation of the document. As mentioned above, access links to various provenance representations are included, in addition to a numbers of provenance transformations and visualizations (see Section 2). The provenance validity of the document can be checked directly from inside the document page (provided by the external ProvValidator service²).

2 Provenance Transformation and Visualization

A provenance document can contain bundles, which are a PROV construct to support bundling a set of provenance descriptions (so allowing provenance of provenance to be expressed) [MM13]. To support relating provenance statements within a document across its bundles, ProvStore can produce a *flattened* representation of the document in which all of its provenance statements are merged into a flat document. In this representation, the provenance of entities distributed in multiple bundles can be “connected” for further examination.

In addition to the flattened representation, ProvStore provides a number of provenance views: Data Flow (concerned with the flow of information or the

² provenance.ecs.soton.ac.uk/validator

transformations of things), Process Flow (concerned with the processes that took place), and Responsibility (assigning responsibility for what happened) [MG13, Ch. 3]. These views are simplified versions of the original document produced by selecting only the relevant provenance descriptions from it. They can facilitate the examination of provenance information by allowing users to focus on a single aspect of it rather than the full descriptions. Each of the views can be applied either on the original document or on its flattened version.

All versions (original or flattened, optionally simplified in a provenance view) of a ProvStore document can be visualized in a (static) graphical representation (in the SVG format). In addition, ProvStore provides interactive visualization tools for users to explore a provenance graph through a Hive plot (highlighting input, output, and intermediary nodes), a Wheel plot (showing the density of connections to/from nodes), and a Gantt chart (presenting entities, activities, and agents on a time line). The Hive and Wheel plots also allow filtering on provenance assertion types to simplify the visualizations.

3 RESTful Application Programming Interface (API)

All of the functionality described in the previous sections (with the exception of interactive features like validation and visualizations) can be accessed programmatically via a RESTful API³ over the Hypertext Transfer Protocol. ProvStore, hence, can serve as a provenance storage-and-publish service on the cloud, providing applications a means to make the provenance of their data available online as soon as it is generated/recorded. Authorized applications must authenticate with ProvStore's API either by using their secret API keys or by following the OAuth (version 1) protocol. With the latter, ProvStore enables users of any third-party applications registered with it to seamlessly access the users' provenance data directly from inside such applications.

References

- EHM⁺12. Mark Ebdon, Trung Dong Huynh, Luc Moreau, Sarvapali Ramchurn, and Stephen Roberts. Network analysis on provenance graphs from a crowdsourcing application. In Paul Groth and James Frew, editors, *Provenance and Annotation of Data and Processes*, volume 7525 of *Lecture Notes in Computer Science*, pages 168–182. Springer Berlin Heidelberg, 2012.
- HJK⁺13. Trung Dong Huynh, Michael O. Jewell, Amir Sezavar Keshavarz, Danius T. Michaelides, Huanjia Yang, and Luc Moreau. The PROV-JSON serialization. Technical report, World Wide Web Consortium, April 2013.
- MG13. Luc Moreau and Paul Groth. *Provenance: An Introduction to PROV*. Morgan & Claypool, September 2013.
- MM13. Luc Moreau and Paolo Missier. PROV-DM: The PROV Data Model. Technical report, World Wide Web Consortium, 2013. W3C Recommendation.

³ See provenance.ecs.soton.ac.uk/store/help/api for the full specification of the API and example codes.