So You Want to Track Provenance
Concepts and Considerations

Anna Krohn : anna.krohn@tufts.edu
RDA/US Scholars Program Internship and Research Data Provenance Interest Group

I. Data Provenance

Data provenance, the following and recording of data's origins, transformations, and movement, is an essential piece of metadata for establishing the reconstructibility, reproducibility, quality, and trustworthiness of data. Many groups within the sciences and humanities have realized the value of provenance and now wish to add it to their metadata. Adding provenance tracking to a dataset must first begin with an examination of the dataset's context. A literature review [1] shows three distinct context "traditions" each with their own subtypes and capturing/recording methods, as elaborated below. Figure 1 also shows how the main types can make use of the others' subtype concepts.

Provenance Types			
	Database	Workflow	Web
	"a relation between versions of a database describing how each part of the output was derived from data in earlier versions or external sources." [2]	"The record of the history of the derivation of the final output" of a workflow, a process of "computation steps and human-machine interaction steps" [5]	A process that preserves not only data creation/origin but also data access information. [4]
Subtypes	**Why -** what "pieces of input data validate the existence of an output value, for a given query"[3] **Where -** "pieces of input data contributing to the identified output variable" [3] **How -** tracing how pieces of input data were "involved in the calculation" of the output [4]	**Actor -** "recording processes information and the time of the execution" [6] **Input -** "tracking the set of input data used to infer a data product" [6] **Interaction -** "recording inter actions between components and the data passed between them" [6]	**Access -** includes both actions of publication and consumption of data [4]
Methods	**Annotation -** data provenance information collection that changes a query to produce not only output identical to the original query but additionally produces the desired provenance information. [5] **Query -** data provenance information collection where a query is run and then the input, output, and the query itself are examined to extract the desired provenance information. [5]	**Annotation -** "metadata comprising of the derivation history of a data product is collected as annotations and descriptions about the source data and processes." [7] **Inversion -** "uses the property by which some derivations can be inverted to find the input data supplied to them to find the output data."[7]	**Recordable -** "information on executions that are performed by the system itself or that can sufficiently be monitored by the system." [4] **Metadata -** "can not be recorded automatically but requires the evaluation of metadata that is published on the Web. Metadata-reliant provenance information comprises information about executions inaccessible to the system as well as information about actors and artifacts involved in these executions." [4]

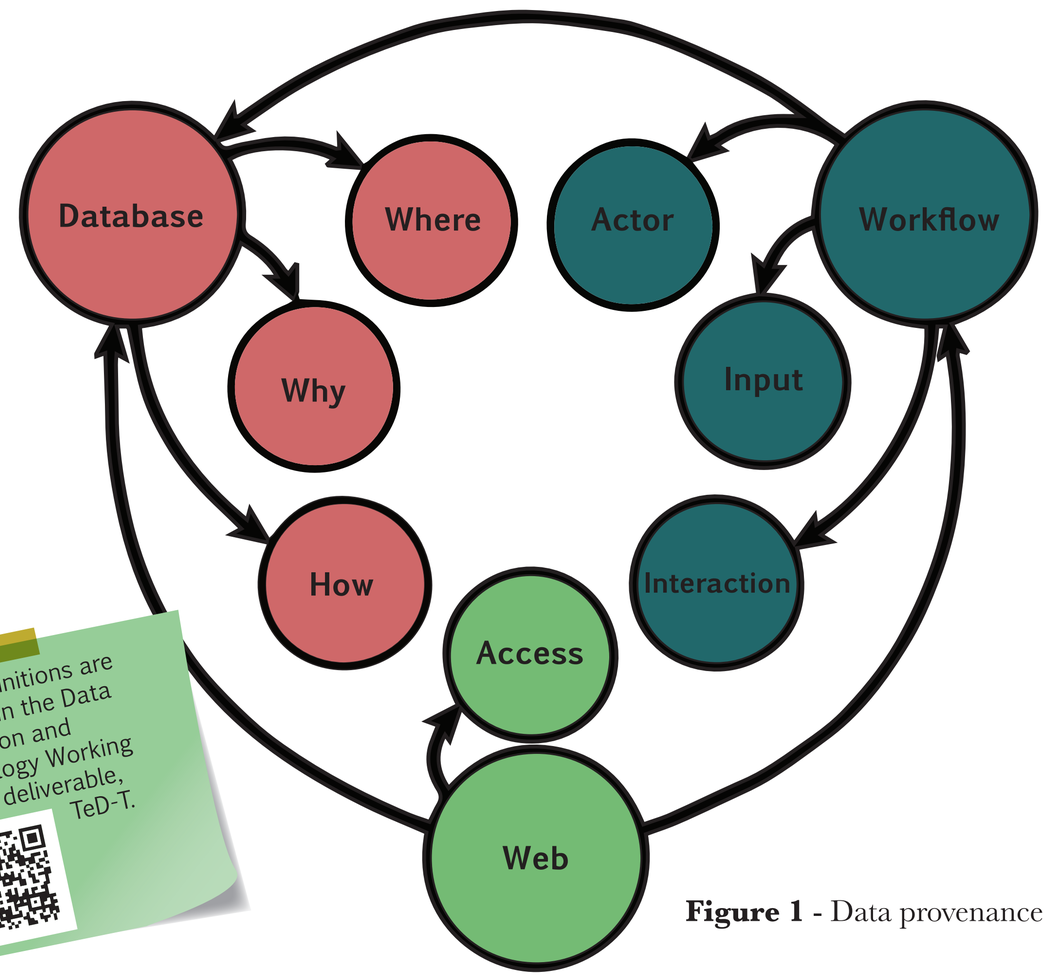

Figure 1 - Data provenance

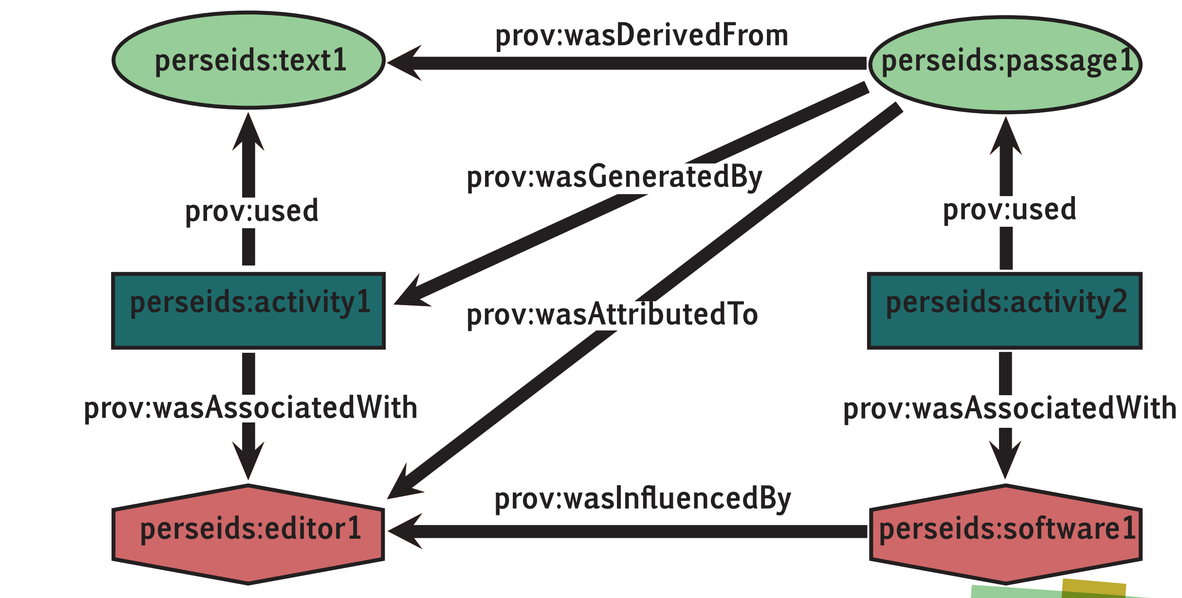

Figure 2 - In this example editor1 performs activity1 (excerption) to produce passage1 from text1 and then kicks off activity2 that consists of software1 acting on passage1. Another entity, passage2, would result from activity2 but has been omitted due to space considerations.

II. Standards and Tools

The nature of the dataset, what sort of provenance is applicable, and how it will be used should inform the choice of the standards and tools. Outside of the three provenance types, domain-specific standards might be encouraged such as the Open Geospatial Consortium Observations and Measurements, or the HL7 Data Provenance Project in development by the U.S. Department of Health and Human Services.

• Many ***Database*** metadata standards, utilized primarily in the library realm, provide some basic form of provenance tracking. Some of the best known examples are Dublin Core (DCMI) and Open Archives Initiative Object Reuse and Exchange (OAI-ORE).

• ***E-science workflows*** typically contain provenance tracking elements and are able to export stored provenance to other data standards. [6, 7]

• ***Web*** technologies have prompted the development of some of the more flexible standards, designed with the goal of being domain and technology-agnostic. The two largest of these are the Open Provenance Model (OPM) and PROV.

Except for OPM and HL7, these standards and others can be found in the Digital Curation Center List of Metadata Standards

III. Linguistic Annotation Example

To demonstrate the addition of provenance to a project, we chose and applied a standard to a use case, linguistic annotation as performed in the Perseids platform. Annotations can be extremely complex. They involve a series of steps that involve potentially multiple editors and pieces of software, editing can span a long period of time, and there is a need to track the intermediary states of the processes. We chose PROV because:

- there are clear database, workflow, and web elements
- a desire to disseminate the resulting work and provenance in XML and RDF formats

Figure 2 provides a PROV representation of a portion of an annotation workflow. Further work is required, but our initial analysis is as follows.

• ***Advantages*** - PROV defines simple, top-level concepts
- allows users to insert domain specific namespaces
- full vocabulary to express the full range of provenance types
- provenance documents can stand on their own, or be incorporated into other documents (in the case of PROV XML).

• ***Drawbacks*** - how best to store the PROV files
- potentially large files depending on the granularity
- unclear what the best practice is for representing some concepts
  - spans of time representing annotation completion
- how to link provenance documents from work done on the same texts

References

[1] Data Provenance Bibliography, https://www.zotero.org/groups/data_provenance.
[2] P. Buneman, S. Khanna, and W.-C. Tan, "Data Provenance: Some Basic Issues," in FST TCS 2000: Foundations of Software Technology and Theoretical Computer Science, vol. 1974, S. Kapoor and S. Prasad, Eds. Springer Berlin Heidelberg, 2000, pp. 87–93.
[3] P. Buneman, S. Khanna, and T. Wang-Chiew, "Why and Where: A Characterization of Data Provenance," in Database Theory — ICDT 2001, vol. 1973, J. Bussche and V. Vianu, Eds. Springer Berlin Heidelberg, 2001, pp. 316–330.
[4] O. Hartig, "Provenance information in the web of data," in Proceedings of the 2nd Workshop on Linked Data on the Web (LDOW2009), 2009.
[5] W.-C. Tan, Provenance in Databases: Past, Current, and Future. 2007.
[6] I. Altintas, O. Barney, and E. Jaeger-Frank, "Provenance Collection Support in the Kepler Scientific Workflow System," in Provenance and Annotation of Data, vol. 4145, L. Moreau and I. Foster, Eds. Springer Berlin Heidelberg, 2006, pp. 118–132.
[7] Y. L. Simmhan, B. Plale, and D. Gannon, "A Survey of Data Provenance in e-Science," SIGMOD Rec., vol. 34, no. 3, pp. 31–36, Sep. 2005.

Acknowledgements

I am grateful to the RDA and RDA/US Scholars Program for the opportunity to pursue this project, and specifically to Beth Plale and Inna Kouper for their organization and support of the internship. I especially wish to thank Bridget Almas and David Dubin for their guidance.

Scan for more information and the PDF of this poster.