Towards A Social Provenance Model for the Web

Andreas Harth Axel Polleres Stefan Decker

Digital Enterprise Research Institute
National University of Ireland, Galway

Abstract

In this position paper we firstly present the established notion of provenance on the Semantic Web (also referred to as named graphs or contexts), and secondly argue for the benefit of adding to the pure technical notion of provenance a social dimension to associate provenance with the originator (typically a person) of a given piece of information.

1 Basic Web Provenance Model

The notion of provenance is an essential concept when integrating data on the Web. Millions of people and organisations are contributing content to a massive, shared information space, which self-organises into a huge graph built from simple source-target links between documents and other resources addressable via Uniform Resource Identifiers (URIs). Part of the Semantic Web technology stack is the Resource Description Framework (RDF), a graph-structured data model widely used for knowledge representation which uses URIs to identify both nodes and arcs, arriving at subject-predicate-object triples.

The notion of provenance is intrinsic to the Web; the mechanism for accessing sources based on the URI addressing scheme via HTTP lends itself to a straightforward notion of provenance. The most natural interpretation of provenance on the Web denotes a URI to a data source which, when dereferenced via the Hypertext Transfer Protocol (HTTP), returns the data that the source provides at the time of the lookup. State of the art RDF repositories store subject-predicate-object-context quadruples, where the last element may denote the source of a given statement.

2 Named Graphs in SPARQL

The SPARQL Query Language for RDF[1] includes a notion of data provenance via "named graphs" since many RDF data stores hold multiple RDF graphs and record information about each graph. To reference data sources or more generally subgraphs from within queries, the query language includes the FROM, FROM NAMED, and GRAPH keywords for dealing with named graphs [1]. However, SPARQL does not define whether the name of the graph has to be retrievable or not, allowing to store and query RDF data that has not been published on the Web.

3 Application Scenarios

In the following, we present some application scenarios in which Web provenance tracking is useful.

- **Excluding Malicious Sources** For some applications, it is desirable to specify which sources to include or exclude during query evaluation or data processing. Having the ability to exclude certain sources (or even all sources within a given distance of certain sources) allows to reduce the amount of noise in query answers. For example, spam or accidentally created bogus statements can pollute datasets, but once identified could be easily disregarded from processing.

- **Prioritising Statements and Measuring Trust** Provenance can be utilised for ranking as well, e.g. attaching a measure for trustworthiness to the data source URI. When using data source URIs for ranking purposes, using the hostname rather than full data source URIs might increase the quality of the ranking procedure in the face of link spam generated by link farms. It is worthwhile

[1] `http://www.w3.org/TR/rdf-sparql-query/`

to note that in terms of data, what constitutes noise to one person might represent valuable information to another.

- **Open World/Closed World Assumption** Interesting in this context is the ability to add scoped negation, a variant of negation as failure, to a query or rules language [3] that allows to "close off" certain data partitions that are considered complete and as a result provides a mix between open and closed world assumption.

4 Associating People with Data Provenance

So far, we have presented the technical notion of provenance on the Web. Ultimately, people or groups of people are originators of data on the Web, so consequently a provenance model has to cater for that fact. At first glance, ownership of any URI, including data source URIs, can be traced to people and organisations via the Domain Name System (DNS), one of the cornerstones of Internet infrastructure. However, given that some sites pertain to entire organisations and others allow external users to publish information, a more fine-grained social provenance model is required, enriched with the ability to encode the policy of the publishing site regarding division of the URI space and responsibilities and liabilities for publishing data. Questions regarding to a DNS owner's policy for assigning URIs include for example creating subdomains and subdirectories.

The main benefit for a social context model is the ability to talk about provenance according to a higher abstraction level and granularity, namely the social context of sources. Given such a model, it is possible to scope queries according to people relative to a social network. Moreover, a such extended provenance model allows for posing queries that take the social network of people or the trustworthiness of people or groups of people into account.

A social model of provenance will allow to ask queries such as "return all movies that are highly rated by my acquaintances, but exclude person X since I dislike her movie taste". In summary, taking into account the social network into trust models provides a higher level abstraction closer to the real world in which people operate.

5 Future Directions

Areas for future work include the investigation of more general notions of context [2, 4] where provenance is only a special case. However, a trade-off decision has likely to be made between the efficiency of the system and the expressivity of the language.

One of the main prerequisites for the ability to abstract from data source URIs and talk in terms of persons and organisations as the originators of data items will be a language or formalism to describe data placement policies for URI spaces that go beyond the granularity of DNS domain names.

We believe adding a social component to provenance tracking on the Web will complement the recent Social Media phenomenon explicable by the popularity of wikis and blogs, which are inherently people-centric.

References

[1] J. J. Carroll, C. Bizer, P. Hayes, and P. Stickler. Semantic web foundations: Named graphs, provenance and trust. In *Proceedings of 14th international conference on World Wide Web*, May 2005.

[2] R. V. Guha. *Contexts: A Formalization and Some Applications.* PhD thesis, Stanford University, February 1995.

[3] A. Polleres, C. Feier, and A. Harth. Rules with contextually scoped negation. In *Proceedings of 3rd European Semantic Web Conference*, June 2006.

[4] M. Sintek and S. Decker. Triple – an rdf query, inference, and transformation language for the semantic web. In *Proceedings of 1st International Semantic Web Conference*, June 2002.