ProvQL: Understanding the Provenance of your Data

Giorgos Flouris, Argyro Avgoustaki, Irini Fundulaki, Dimitris Plexousakis
Institute of Computer Science (ICS),
Foundation for Research and Technology – Hellas (FORTH), Greece
{fgeo, argiro, fundul, dp}@ics.forth.gr

ProvQL is a high-level structured query language, suitable for seeking information related to data provenance. It is especially suitable for tracking the sources that contributed to data generation, and for helping the data expert assess the trustworthiness and reliability of data.

The proliferation of (open) data on the Web, in the form of Linked Open Data, has made possible to publish structured data so that it can be interlinked and become more useful through semantic queries. This way, web pages are not only useful for human readers, but are also able to include data and information that can be read automatically by computers (metadata).

In this setting, data is being recorded in the form of RDF quadruples, e.g., (Donald_Trump, president_of, USA, DBpedia), where the subject (Donald_Trump) is associated with the object (USA) with a property (president_of), and this information is recorded in a specific dataset (DBpedia). Note that all elements of the quadruple (e.g., Donald_Trump) identify a specific resolvable entity or relation, corresponding to an actual person, role, object, or concept (e.g., the person Donald Trump, the role of being a president etc.). This way, all references to "Donald_Trump", in any dataset, are automatically associated to each other.

This constitutes one of the main advantages of linking different datasets, namely the ability to combine data and information from different sources in order to produce new data or knowledge on a domain of interest. However, this advantage comes with a price. For example, if a certain quadruple is later identified to be erroneous or inaccurate, how can one track down its dependencies and identify the (derived) quadruples that should be retracted? And how can one assess the trustworthiness, credibility, authenticity, reliability or accuracy of his/her data, when part of it was the result of some form of reasoning over data coming from other, external, sources? And what about the problem of data accountability?

To address these questions, there is a great need for recording the **provenance** of published data, i.e., their history, their origins and the process through which data records (quadruples) were "copied" from one source to another, modified and transformed in the process and/or reasoned upon to produce new data records (see Figure 1).

In recent years, provenance has been widely studied in several different contexts, e.g., databases, workflows, distributed systems, Semantic Web, etc., and with respect to different aspects and applications. These studies have resulted to different theoretical provenance models, each with a different level of complexity and detail, such as the so-called why, where and how provenance models, as well as hybrid ones [R1]. Also, various representation models (such as CIDOC CRM_{dig} or W3C PROV) have been developed to support the representation of different types of provenance information (see Figure 2).

Although these works have addressed the problem of determining how provenance should be recorded and represented, the issue of **querying data provenance information** has not yet been adequately considered. Although answering provenance queries over existing models (such as W3C PROV) is possible using standard (adequate) query languages (such as SPARQL), this would require familiarity with the specific representation model, and would be suitable only for this specific representation model. Therefore, applications using this approach would break if the representation model changes.

The objective of our work is to define a **structured high-level query language for provenance**, called *ProvQL*, which will allow a user to express provenance queries. The language should be **implementation-agnostic** and **representation-agnostic**, i.e., unrelated to the specific representation model used to represent provenance. In addition, one should be able to filter the required data both on the basis of its provenance, and on the basis of the actual data itself. In particular, ProvQL should be able to answer queries like the following:

- Which data records or sources contributed in deriving this data record?
- Identify all data records whose provenance includes a specific data source (or data item).
- Identify all quadruples referring to "Donald_Trump" that originate from a specific source.

ProvQL is currently under development. We have already defined the syntax of the language, which follows an SQL-like format, as well as its semantics, based on the well-known approach of mappings [R2] that associate variables with data or provenance. The next step is to provide a suitable implementation of the language, efficient enough to tackle complex provenance queries over large datasets.

References:

[R1]: Cheney, J., Chiticariu, L. and Tan, W.C., 2009. Provenance in databases: Why, how, and where. Foundations and Trends in Databases, 1(4).
[R2]: Pérez, J., Arenas, M. and Gutierrez, C., 2009. Semantics and complexity of SPARQL. ACM TODS, 34(3), p.16.

Please contact:

Argyro Avgoustaki
Institute of Computer Science (ICS),
Foundation for Research and Technology, Hellas
argiro@ics.forth.gr
+30 2810 391683

Figures:

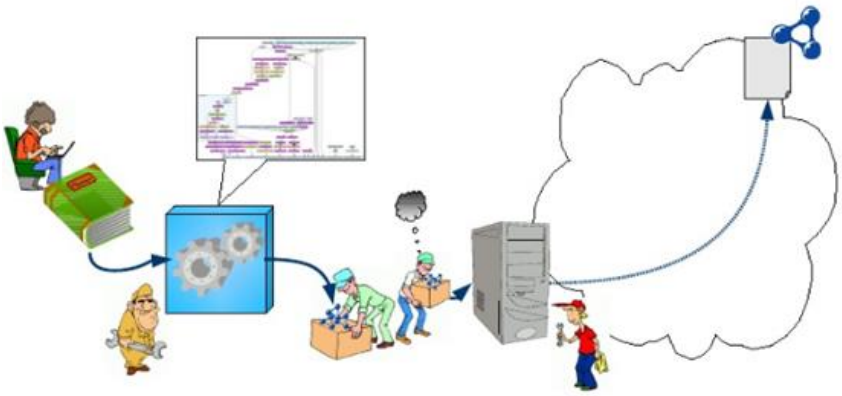

Figure 1. Data Provenance (taken from O. Hartig, J. Zhao, "Using Web Data Provenance for Quality Assessment", SWPM-09)

Figure 2. Excerpt from the W3C PROV model (taken from Provenance Analysis and RDF Query Processing Satya Sahoo, Praveen Rao)