Querying Data Provenance in the Internet of Things

Argyro Avgoustaki[1#], **Giorgos Flouris**[1*], **Irini Fundulaki**[1] and **Dimitris Plexousakis**[1]

[1] Institute of Computer Science – FORTH Hellas

# Presenting author: Argyro Avgoustaki, email: argiro@ics.forth.gr
* Corresponding author: Giorgos Flouris, email: fgeo@ics.forth.gr

ABSTRACT

The proliferation of data on the Web has made possible to publish structured data so that they can be interlinked and become useful through semantic queries. This way, web pages are not only useful for human readers, but also include data and information that can be read automatically by computers (metadata). In this setting, data are 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 relation (president_of), and this information is recorded in a specific dataset (DBpedia). Note that all elements of the quadruple identify a specific resolvable entity or relation, corresponding to an actual person, object, or concept (e.g, 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. 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, 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? 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, 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.

In recent years, provenance has been widely studied in several different contexts and with respect to different aspects and applications. In addition, various representation models have been developed to support the representation of different types of provenance information. 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.

The objective of our work is to define a new structured high-level query language for provenance, called ProvQL, which allows a user to express provenance queries. The language should be implementation-agnostic and representation-agnostic, i.e., unrelated to the provenance representation model. 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 is able to answer queries like "Which sources contributed in deriving this data record?", " Identify all data records whose provenance includes a specific data source (or data item)" and "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 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, to assess the language's expressiveness and evaluate its performance against large datasets.