Integrating Provenance into the Web of Data

Olaf Hartig[1] and Jun Zhao[2]

[1] Humboldt-Universität zu Berlin
hartig@informatik.hu-berlin.de
[2] University of Oxford
jun.zhao@zoo.ox.ac.uk

1 Introduction

During recent years an increasing number of data providers adopted a set of best practices for publishing and connecting structured data on the Web, leading to the creation of a globally distributed dataspace, the Web of Linked Data [1]. While this dataspace holds an enormous potential, using data from the Web poses questions of information quality and trustworthiness. Provenance information about who created and published the data and how, provides a promising means for quality assessment [2]. However, to enable successful and reliable provenance-based assessment methods we conceive it as an absolute necessity that provenance-related metadata becomes an integral part of the Web of Linked Data [3]. This is only possible if the publication of this metadata adheres to the same principles that are used for the data itself. Therefore, we present a vocabulary that allows providers of Linked Data to describe the provenance of their data with RDF. These descriptions have to be created and made available, ideally as Linked Data, again. To initiate such a practice with a minimal effort for data publishers we extended several Linked Data publishing tools with corresponding metadata components. In the remainder of this poster proposal we introduce our vocabulary in Section 2 and we describe our metadata extensions in Section 3.

2 The Provenance Vocabulary

In a recent study we reveal a lack of suitable vocabularies to describe provenance of Linked Data [4]. To fill this void we develop the Provenance Vocabulary[3]. We designed the vocabulary very closely to our model for Web data provenance [4]. As the provenance model comprises two dimensions, i.e., data creation and data access, our vocabulary accordingly consists of three categories of terms: general terms, terms for data creation, and terms for data access (cf. Figure 1). These terms allow to describe all common cases related to Linked Data publishing such as: manual data creation, Linked Data interfaces over native RDF stores, wrappers over relational databases or over Web APIs, file-based data creations, data changes, retrieval and provenance of source data, publication responsibilities, and service operators. To allow for a wide range of applications the vocabulary does not prescribe a specific granularity by which provenance has to be described. Hence, the classes are quite general. For instance, a DataItem could

[3] http://purl.org/net/provenance/

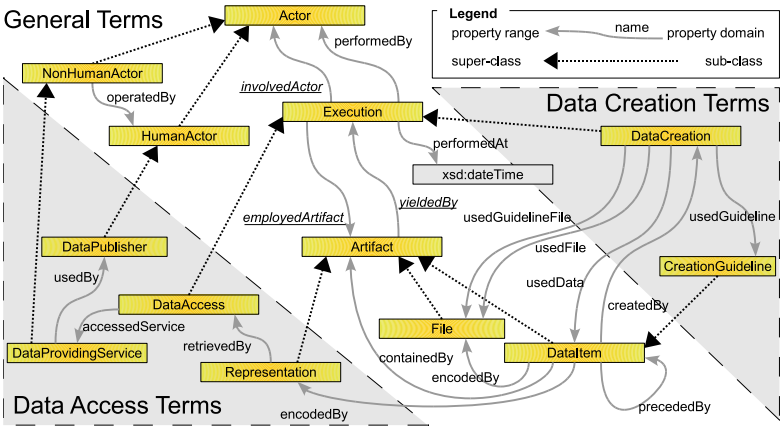

Fig. 1. Classes and properties defined by the Provenance Vocabulary core ontology.

be a whole linked dataset, an RDF graph as provided by a Linked Data server, or even a single RDF triple, depending on the granularity chosen. The vocabulary is defined as an OWL ontology and it is partitioned into a core ontology and the supplementary modules: Types and Integrity Verification.

3 Metadata Extensions for Linked Data Publishing Tools

To enable a large-scale augmentation of the Web of Linked Data with provenance metadata we extended several tools that are widely used for publishing Linked Data, including Triplify[4], Pubby[5] and D2R server[6]. Our extensions automatically generate and add provenance metadata to each RDF graph that is served following the Linked Data principles. Due to our extensions data publishers can easily enrich their data with provenance metadata by simply configuring a few parameters, such as the name and the URI identifying the publisher or the URI of the dataset. Hence, with data providers upgrading their Linked Data servers to the latest release of these tools the amount of provenance information added to the Web of Linked Data will increase significantly.

References

1. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data – The Story So Far. International Journal on Semantic Web and Information Systems **5**(3) (2009)
2. Hartig, O., Zhao, J.: Using web data provenance for quality assessment. In: Proc. of the Workshop on Semantic Web and Provenance Management at ISWC. (2009)
3. Hartig, O., Zhao, J.: Publishing and consuming provenance metadata on the web of linked data. In: Proc. of the 3rd Int. Provenance and Annotation Workshop. (2010)
4. Hartig, O.: Provenance information in the web of data. In: Proc. of the Linked Data on the Web Workshop at WWW. (2009)

[4] http://triplify.org/
[5] http://www4.wiwiss.fu-berlin.de/pubby/
[6] http://www4.wiwiss.fu-berlin.de/bizer/d2r-server/