Web Semantics: Science, Services and Agents on the World Wide Web 9 (2011) 149–164

Contents lists available at ScienceDirect

Web Semantics: Science, Services and Agents on the World Wide Web

journal homepage: http://www.elsevier.com/locate/websem

Modelling provenance of DBpedia resources using Wikipedia contributions ☆

Fabrizio Orlandi *, Alexandre Passant

Digital Enterprise Research Institute, National University of Ireland, Galway, Ireland

ARTICLE INFO

Article history:
Available online 21 April 2011

Keywords:
Provenance
Linked Data
DBpedia
Wikipedia

ABSTRACT

DBpedia is one of the largest datasets in the linked Open Data cloud. Its centrality and its cross-domain nature makes it one of the most important and most referred to knowledge bases on the Web of Data, generally used as a reference for data interlinking. Yet, in spite of its authoritative aspect, there is no work so far tackling the provenance aspect of DBpedia statements. By being extracted from Wikipedia, an open and collaborative encyclopedia, delivering provenance information about it would help to ensure trustworthiness of its data, a major need for people using DBpedia data for building applications. To overcome this problem, we propose an approach for modelling and managing provenance on DBpedia using Wikipedia edits, and making this information available on the Web of Data.

In this paper, we describe the framework that we implemented to do so, consisting in (1) a lightweight modelling solution to semantically represent provenance of both DBpedia resources and Wikipedia content, along with mappings to popular ontologies such as the W7 – *what*, *when*, *where*, *how*, *who*, *which*, and *why* – and OPM – open provenance model – models, (2) an information extraction process and a provenance-computation system combining Wikipedia articles' history with DBpedia information, (3) a set of scripts to make provenance information about DBpedia statements directly available when browsing this source, as well as being publicly exposed in RDF for letting software agents consume it.

1. Introduction

Collaborative websites such as Wikipedia[1] have recently shown the benefit of being able to create and manage very large public knowledge bases.[2] However, one of the most common concerns about these types of information sources is the trustworthiness of their content which can be arbitrarily edited by everyone. The DBpedia project,[3] which aims at converting Wikipedia content into structured knowledge, is then not exempt from this concern. Especially considering that one of the main objectives of DBpedia is to build a dataset such that semantic web technologies can be employed against it. Hence this allows not only to formulate sophisticated queries against Wikipedia, but also to link it to other datasets on the Web, or create new applications or mashups [3]. Thanks to its large dataset (around 1 billion RDF triples) and its cross-domain nature DBpedia has become one of the most important and interlinked datasets on the Web of Data [4]. Therefore ensuring provenance information of DBpedia data is crucial, especially for developers consuming or interlinking its content.

Research on Wikipedia, and on collaborative websites in general, shows that some information quality aspects (such as currency and formality of language) of Wikipedia are quite high [5]. However, as suggested in [8], the high quality level of certain aspects of Wikipedia articles does not imply that it is good on other dimensions as well. In fact, a substantial qualitative difference exists in Wikipedia between "featured" articles (high quality articles identified by the community) and normal articles [8]. For this reason it is important to identify quality measures for Wikipedia articles and estimate the trustworthiness of their content. Then, since the DBpedia content is directly extracted from Wikipedia, the same trust and quality values can be propagated to the DBpedia dataset. However, in order to obtain these values, it is essential to provide detailed provenance information about the data published on the Web.

The benefits of using data provenance to develop trust on the Web, and the Semantic Web in particular, have been already widely described in the state of the art (see [6,7]). Provenance of data provides useful information such as timeliness and authorship of data. It can be used as a ground basis for various applications and use cases such as identifying trust values for pages or pages fragments [2], or measuring users' expertise by analysing their contributions [27] and then personalize trust metrics based on the user profile of a person on a particular topic

☆ This work is funded by the Science Foundation Ireland under Grant Number SFI/08/CE/I1380 (Líon 2), an IRCSET scholarship and a grant by Cisco Research Center.

* Corresponding author. Tel.: +353 91 494 035.

E-mail addresses: fabrizio.orlandi@deri.org (F. Orlandi), alexandre.passant@deri.org (A. Passant).

[1] http://www.wikipedia.org.

[2] Statistics about Wikipedia: http://stats.wikimedia.org/EN/Sitemap.htm

[3] http://dbpedia.org/.

1570-8268/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.websem.2011.03.002

[21]. Moreover, providing also provenance meta-data as RDF and making it available on the Web of Data [23], offers more interchange possibilities and transparency. This would let people link to provenance information from other sources. It provides them the opportunity to compare these sources and choose the most appropriate one or the one with higher quality. In our specific context of DBpedia for example, by indicating by whom and when a triple was created (or contributed by), it could let any application flag, reject or approve this statement based on particular criteria (see Section 5).

In this paper we propose a modelling solution to semantically represent information about provenance of data in DBpedia and an extraction framework capable of computing provenance for DBpedia statements using Wikipedia edits. In particular, in the next section we review some related work in the realm of provenance management on the Web of Data and in trust and quality evaluation techniques on wikis. Comparing these two research fields we highlight the limitations that we found in both of them: the former lacks concrete and well established procedures to support the integration and publication of provenance of non- or semi-structured data on the Web of Data; the latter does not take into account the importance of making the information generated analysing users' edits available as Linked Data and providing details of the steps involved in the analysis. Then, in Section 3, we give some background information regarding lightweight ontologies such as Semantically-Interlinked Online Communities (SIOC) and its extensions which will be used in our modelling solutions. We decided to use the SIOC vocabulary and its extensions because it aims at describing the structure of online communities such as in wikis, and its *Actions* module suits well our need of defining events and user activities in wikis. In Section 4, we detail the W7 model for provenance representation, as previously designed by Ram et al. [33], and our implementation of this model with a lightweight ontology built to express it in RDFS. In the same section we also provide an alignment of our model with the Open Provenance Model (OPM), the reference ontology chosen by the W3C provenance incubator group. Finally, in Section 5, we describe the architecture of our DBpedia provenance extraction framework. Then we detail how we model provenance information for DBpedia statements and expose it as Linked Open Data. Before concluding we also show a set of scripts to directly browse information about the statements on the DBpedia pages.

2. Related work

Extracting and representing provenance information about data is a research topic that is going on from many years. Many studies have been conducted for representing provenance of data so far. Among all in [10,36] the authors provide comprehensive surveys about data provenance methodologies. The first one provides one of the first surveys in the field applied to a scientific data processing context. The second one provides, in a more generic context related to e-science projects, a taxonomy to understand and compare provenance techniques. Moreover, a comprehensive survey about provenance on the Web has been recently published by Moreau [11]. Considering also existing semantic models for provenance of data, what is common between most of the modelling solutions is the presence of three concepts involved in the data life-cycle: actors, processes and artefacts. Indeed, a modelling approach can be "process-oriented", "data-oriented" (the two distinctions made in [36]), or "actor-oriented" (as proposed by Harth et al. [22]).

However these studies, and most of the studies about data provenance, are not focused on integrating provenance information into the Web of Data. In [23] the author explicitly addresses the characteristics of provenance of data from the Web, and proposes the "Provenance Vocabulary".[4] We agree with the author on the fact that providing this information as RDF would make provenance metadata more transparent and interlinked with other sources, and it would also offer new scenarios on evaluating trust and data quality on the top of it. In this regard a W3C Provenance Incubator Group[5] has been recently established. The mission of the group is to "provide a state-of-the art understanding and develop a roadmap in the area of provenance for semantic web technologies, development, and possible standardisation". Requirements for provenance on the Web,[6] along with several use cases and technical requirements have been provided by the working group so far. These activities and documents have been recently included in a final report of the activities of the incubator group.[7] We invite the reader to consult this document in order to have more detailed information about the requirements for provenance needed in this work. In particular the requirements belonging to the following "dimensions": object, attribution, process, versioning, publication and scale. The report contains also mappings between the most relevant provenance ontologies. Many ontologies representing provenance of data are taken into consideration (such as the Provenance Ontology, the Provenir ontology, the Open Provenance Model (OPM), etc.), as well as other lightweight ontologies (such as the Dublin Core[8] vocabulary) that can partially represent provenance aspects of web data. An alignment of these ontologies is provided in the aforementioned W3C document, and the model taken as reference for the mappings is the OPM (more details later in Section 4). Finally, a comprehensive analysis of approaches and methodologies for publishing and consuming provenance metadata on the Web is exposed in [24].

Another research topic relevant to our work is the evaluation of trust and data quality in wikis. Recent studies proposed several different algorithms for wikis that would automatically calculate users' contributions and evaluate their quantity and quality in order to study the authors' behaviour, produce trust measures of the articles and find experts. WikiTrust [2] is a project aimed at measuring the quality of author contributions on Wikipedia. They developed a tool that computes the origin and author of every word on a wiki page, as well as "a measure of text trust that indicates the extent with which text has been revised".[9] On the same topic other researchers tried to solve the problem of evaluating articles' quality, not only examining quantitatively the users' history [27], but also using social network analysis techniques [28]. Another relevant contribution is in [19], where the author details the implementation of a system for expert finding in Wikipedia.

From our perspective, there is a need of publishing provenance information as Linked Data from websites hosting a wide source of information (such as Wikipedia) and also from relevant datasets (such as DBpedia). Yet, most of the work on provenance of data is, either not focused on integrating provenance information on the Web of Data, or mainly based on provenance for resource descriptions or already structured data. On the other hand, the interesting work done so far on analysing trust and quality on wikis does not take into account the importance of making the analysed data available on the Web of Data.

Interesting and related research in our context is also presented in [14,15]. First, the work by Vrandečić et al. describes a collaborative web application that allows users to aggregate sources of information on entities of interest from the Web of Data. It takes Wikipedia as its starting point for its entities and it provides the source of every information added by its users. Then, the research

[4] http://purl.org/net/provenance/ns.

[5] Established in September 2009. http://www.w3.org/2005/Incubator/prov/.

[6] http://www.w3.org/2005/Incubator/prov/wiki/User_Requirements.

[7] http://www.w3.org/2005/Incubator/prov/XGR-prov-20101214/.

[8] http://dublincore.org/.

[9] http://wikitrust.soe.ucsc.edu/.