Provenance Requirements for the Next Version of RDF

Jun Zhao, Christian Bizer, Yolanda Gil, Paolo Missier, Satya Sahoo

Abstract

The provenance (i.e., origins) of information on the Web is crucial in many applications to allow information quality assessment, data integration, trust judgments, reproducibility, accountability, and many other important tasks. This document summarizes the positions of the W3C Provenance Incubator Group on the requirements to support provenance and their implications on the current implementation and future prospects of the W3C RDF standard.

Introduction

The provenance of information is crucial to making determinations about whether information is trusted, how to integrate diverse information sources, and how to give credit to originators when reusing information. Broadly construed, provenance encompasses the initial sources of information used as well as any entity and process involved in producing a result. In an open and inclusive environment such as the Web, users who face information that is often contradictory or questionable would benefit from explicit provenance meta-information when making trust judgments. In particular, with the arrival of massive amounts of Semantic Web data (e.g., via the Linked Open Data community), the provenance of that data becomes an important factor in developing new Semantic Web applications. While it is important to choose a provenance data model that is practical and easy-to-use, however, a crucial enabler of the Semantic Web deployment is the explicit representation of provenance that is accessible to machines, in addition to humans.

In this paper we base on the principle that the data model is sufficiently expressive to represent both data and some of its provenance metadata and we focus specifically on representing this information using the RDF model. This uniform representation of RDF is appealing for a number of reasons. For example, it could form the basis for extending the inference capabilities of current Semantic Web reasoners, which operate on RDF graphs that represent large knowledge bases, to take automatically account for provenance metadata as well. Appealing as it sounds, however, this uniform representation of data and metadata requires additional capabilities that the standard RDF model currently does not offer. We use a well-defined set of provenance user requirements to articulate some of these shortcomings, and suggest new requirements to RDF that would make it suitable for seamless representation of provenance.

As a starting point, we take our user requirements gathered and documented by the W3C Provenance Incubator Group [charter], which was formed in September 2009 as part of the W3C Semantic Web Activity. Its charter was to provide a state-of-the art understanding and develop a roadmap in the area of provenance for Semantic Web technologies, development, and possible standardization. At the time of writing this document, the group has been in existence for six months, which is a half of its expected activity. The group has produced a number of documents, including a report of key dimensions for provenance [dimensions], more than thirty use cases spanning many areas and contexts that illustrate these key dimensions [use-cases], and a broad set of user requirements and technical requirements derived from those use cases [requirements-listed].

The Provenance Group's first published report summarizes requirements for provenance [requirements-report]. This paper is structured as follows: We begin by summarizing the requirements in that report, we then give an overview about current technologies and research on provenance tracking in Semantic Web applications, and we conclude with a recommendation which topics in relation to provenance could be addressed by a future RDF 2.0 working group.

Overview of Provenance Requirements

This section provides an overview of requirements for provenance. A detailed account is provided in [requirements-report]. The report uses extensive examples in three diverse scenarios that illustrate the need for provenance: 1) a news aggregator site that assembles news items from a variety of sources (such as news sites, blogs, and tweets), where provenance records can help with verification, credit, and licensing; 2) a data integration and analysis activity for studying the spread of a disease, involving public policy and scientific research, where provenance records support combining data from very diverse sources, justification of claims and analytic results, and documentation of data analysis processes for validation and reuse; and 3) a case of a commercial company that requires provenance information about their software development and testing procedure in order to defend the validity of their contract execution.

We grouped the requirements for provenance into three major areas of concern: content, management, and use. We now address each in turn, highlighting here only the requirements that are directly relevant to the RDF topic of discussion. For other requirements, examples, and more details we refer the reader to [requirements-report].

Content

Content refers to the type of information that provenance records need to contain. This content may include entities and processes that contributed to its creation or its delivery to a user/consumer. It may also include argumentations, design choices, and justifications for decisions.

- **Requirement 1: Identity** -- A key challenge is to be able to refer to the artifact that we are describing the provenance for. Within the RDF context, the artifact could be a single RDF statement, a set of statements or an arbitrary set of Web resources.
- **Requirement 2: Evolution --** An important requirement is the ability to describe the provenance of a dynamic, evolving resource. Over time, there may be updates and even new versions that change some aspect of the resource. A challenge is to describe how the new incarnations of the resource relate to one another, and to determine whether provenance records should be self-contained and attached to each incarnation, or instead refer to prior ones for details. As resources may be republished, perhaps repackaging, summarizing, or mixing their contents, their provenance records need to reflect such processes and their implications on the contents.
- **Requirement 3: Entailment --** Another important requirement is to distinguish what is directly asserted by the entities and processes that produce the resource from other information that may be inferred from those assertions or perhaps derived or hypothesized by a third party.

Management

Management refers to the mechanisms that make provenance information available and accessible in an open system like the Web. This includes the representation language for provenance records, the methods for publication and dissemination, and the methods for accessibility and querying of provenance records.

- **Requirement 4: Publication --** A publisher of provenance information needs to use some provenance representation language and link the provenance assertions to the actual resource information. The publisher may choose to publish only a subset of the provenance records, and should be able to identify themselves possibly with a signature that is verifiable by others.
- **Requirement 5: Querying** – Provenance information may be made accessible in some manner, and there must be mechanisms to find the provenance for a given resource. Query formulation and execution must be provided for provenance information. Ideally, there should be a convenient way to formulate queries that span primary and provenance information.

Use

Uses of provenance refer to the purposes and usage of provenance information. This includes presentation and visualization of provenance information, supporting abstraction and customization, integration of provenance from heterogeneous systems, allowing trust judgments based on provenance information, and handling imperfections in provenance records. None of the uses of provenance that the group analyzed seemed to lead to new requirements for RDF, so no requirements are added in this section.

State of the Art in RDF Provenance and Alternative Approaches

Many efforts have been put into supporting the provenance requirements described above, by proposing extensions to the existing RDF data model, alternative models of RDF and vocabularies/ontologies for describing patterns like evolutions, versioning, annotations etc. It is time to seek for a standardized means based on these existing approaches, preferable under the umbrella of a W3C working group. This section briefly outlines the current state of the art in representing provenance information in RDF and lists current approaches to extend RDF for the better representation of provenance information. The work aims in two general directions: 1) Approaches to represent provenance information and primary information together as an integrated model (Requirement 1); and 2) Development of schemata, ontologies, and vocabularies to represent and publish specific types of provenance information, such as attribution, versioning or entailment information (Requirements 2, 3, 4).

The official state of the art concerning the representation of RDF together with meta-information is RDF Reification [RDF-semantics]. The RDF reification vocabulary is propagated in the current RDF recommendation for making RDF statements identifiable. The RDF reification vocabulary consists of the four terms rdf:Statement, rdf:subject, rdf:predicate, and rdf:object. Together with additional vocabularies for representing attribution, versioning or entailment information, RDF reification can be used to represent provenance information. Nevertheless, RDF reification was never widely picked up by the community and there is hardly any reified RDF data published on the Web. Querying reified RDF statements with the SPARQL query language is cumbersome, thus reification only partly fulfills Requirement 5 about the convenient querying of provenance information. Over the years, various alternative approaches to RDF reification have been published. These include: N3 Formula [N3], N-Quads [N-QUADS], RDF Molecules [RDF Molecules], Temporal RDF [GHV07, HV06, TB09], Named

Graphs [Named Graphs], OWL Annotations [OWL2], the PaCE Model [Pace model], SPARQL Datasets [SPARQL] and etc. Two incompatible approaches have already found their way into W3C Recommendations: Named Graphs, which are part of the SPARQL Recommendation [SPARQL] and OWL Annotations which are part of the OWL 2 Recommendation [OWL2].

The second area that is important from the provenance angle is to ease the exchange of provenance information between systems and to publish provenance information together with primary information on the Web. Various groups have already developed models to represent specific types of provenance information, such as attribution, versioning or entailment information. The provenance incubator group has listed relevant vocabularies at [Relevant technologies]. These efforts include: The Open Provenance Model [opm], Dublin Core [dc], Open Archives Initiative - Object Reuse and Exchange (OAI-ORE) [oai-ore], Semantic Web Publishing Vocabulary [swpv, named-graphs], Inference Web - Open Proof Language [pml], the SWAN-SIOC alignment [swan-sioc], the data web versioning recommendation based on the Named Graphs [dataweb], the Changeset Vocabulary [changeset], POWDER [powder], RDF coloring [rdf-coloring] and etc.

Conclusion

In order to meet the provenance requirements described in this paper, we think that it is important to extend RDF within an upcoming RDF 2.0 Working Group with an efficient mechanism to provide provenance information about RDF data (Requirement 1). This could for example be achieved by providing for the identification of sets of RDF triples which represent primary information as well as sets that represent provenance information. Second, in order to ease the exchange of provenance records between systems, ease the exchange of versioning information (Requirement 2), and ease the exchange of entailment information (Requirement 3), it would be useful if the RDF 2.0 Working Group or a parallel Provenance Working Group would standardize vocabularies that cover the basic aspects of these areas together with best practices on how to publish provenance information on the Web (Requirement 4) and optionally reassuring such assertions with digital signatures (Requirement 4). For all three topics, various proposals have been make over the last years and it can thus be concluded that the time is right for moving from experimentation to standardization.

Acknowledgements

This position paper is based on the work of the W3C Provenance Incubator Group. We especially thank the following members of the group for their valuable feedback on the text of the paper: Luc Moreau, Deborah McGuiness, James Myers, Irini Fundulaki, Paul Groth and other members from the Incubator Group.

References

[Charter] http://www.w3.org/2005/Incubator/prov/charter

[Use-cases] http://www.w3.org/2005/Incubator/prov/wiki/Use_Cases

[Dimensions] http://www.w3.org/2005/Incubator/prov/wiki/Provenance_Dimensions

[Requirements-list] http://www.w3.org/2005/Incubator/prov/wiki/Requirements

[Requirements-report] http://www.w3.org/2005/Incubator/prov/wiki/Requirements_Clean

[RDF-semantics] P. Hayes and B. McBride, “RDF Semantics”, http://www.w3.org/TR/rdf-mt/, W3C Recommendation 10 February 2004

[N3] T. Berners-Lee, “Notation 3, a readable language for data on the Web”, http://www.w3.org/DesignIssues/Notation3

[N- QUADS] R. Cyganiak, A. Harth, and A. Hogan, “N-Quads: Extending N-Triples with Context”, http://sw.deri.org/2008/07/n-quads/

[RDF Molecules] L. Ding, T. Finin, Y. Peng, P. P. da Silva, and D. L. McGuinness. “Tracking RDF Graph Provenance using RDF Molecules”. In Proc. of ISWC (Poster), 2005.

[GHV07] C. Gutierrez, C.A. Hurtado, and A.A. Vaisman, "Introducing Time into RDF," *IEEE Trans. Knowl. Data Eng.*, vol. 19, 2007, pp. 207-218.

[HV06] C.A. Hurtado and A.A. Vaisman, "Reasoning with Temporal Constraints in RDF," *PPSWR*, 2006, pp. 164-178.

[TB09] J. Tappolet and A. Bernstein, "Applied Temporal RDF: Efficient Temporal Querying of RDF Data with SPARQL," *ESWC*, 2009, pp. 308-322.

[Named Graphs] J. Carroll, C. Bizer, P. Hayes, and P, Stickler: “Named Graphs”. Journal of Web Semantics, Vol. 3, Issue 4, p. 247-267, 2005.

[OWL2 Spec] B. Motik, P. Patel-Schneider, and B. Parsia. “OWL 2 Web Ontology Language Structural Specification and Functional-Style Syntax”. http://www.w3.org/TR/owl2-syntax/. W3C Recommendation 27 October 2009

[Pace model] S. Sahoo, O. Bodenreider, P. Hiltzler, A. Sheth, and K. Thirunarayan, “Provenance Context Entity (PaCE): Scalable Provenance Tracking for Scientific RDF Data”, SSDBM, 2010, To appear

[SPARQL spec] E. Prud'hommeaux and A. Seaborne, “SPARQL Query Language for RDF”, http://www.w3.org/TR/rdf-sparql-query/, W3C Recommendation 15 January 2008

[Relevant technologies] http://www.w3.org/2005/Incubator/prov/wiki/Relevant_Technologies

[opm] L. Moreau, B. Clifford, J. Freire, Y. Gil, Y, P. Groth, J. Futrelle, N. Kwasnikowska, S. Miles, P. Missier, J. Myers, J, Y.Simmhan, E.Stephan, and J. Van den Bussche (2009) The Open Provenance Model --- Core Specification (v1.1). Future Generation Computer Systems . (Submitted)

[dc] The http://dublincore.org/documents/dcmi-terms/

[oai-ore] C. Lagoze, H. Van de Sompel, P. Johnston, M.L. Nelson, R. Sanderson, and S. Warner, Open Archives Initative Object Reuse and Exchange (OAI-ORE), http://www.openarchives.org/ore/

[swpv] The Semantic Web Publishing Vocabulary (SWPV). http://www4.wiwiss.fu-berlin.de/bizer/WIQA/index.htm#WebVocab.

[pml] P.P. Da Silva, D.L. McGuinness, D.L. and R. Fikes. “A proof markup language for semantic web services”, Journal of Information Systems. Vol. 31, Issue 4, p. 381—395, 2006

[swan-sioc] A. Passant, P. Ciccarese, J.G. Breslin, J.G. and T. Clark. “SWAN/SIOC: Aligning Scientific Discourse Representation and Social Semantics”. http://www.w3.org/TR/hcls-swansioc/ W3C Interest Group Note 20 October 2009

[data-web] J. Zhao, A. Miles, G. Klyne, and D. Shotton. “Linked data and provenance in biological data webs”. Briefings in Bioinformatics, (2): 139-152, 2009

[changeset] S. Tunnicliffe and I. Davis, “Changeset”, http://vocab.org/changeset/schema.html

[powder] K. Scheppe, “Protocol for Web Description Resources (POWDER): Primer”. http://www.w3.org/TR/powder-primer/. W3C Working Group Note 1 September 2009

[rdf-coloring] G. Flouris, I. Fundulaki, P. Pediaditis, Y. Theoharis, V. Christophides: “Coloring RDF Triples to Capture Provenance”. International Semantic Web Conference 2009: 196-212