Provenance in Linked Data Integration

Tope Omitola, Nicholas Gibbins, and Nigel Shadbolt

Intelligence, Agents, Multimedia (IAM) Group
School of Electronics and Computer Science
University of Southampton, UK
{ t.omitola, nmg, nrs }@ecs.soton.ac.uk

Abstract. The open world of the (Semantic) Web is a global information space offering diverse materials of disparate qualities, and the opportunity to re-use, aggregate, and integrate these materials in novel ways. The advent of Linked Data brings the potential to expose data on the Web, creating new challenges for data consumers who want to integrate these data. One challenge is the ability, for users, to elicit the reliability and/or the accuracy of the data they come across. In this paper, we describe a light-weight provenance extension for the voiD vocabulary that allows data publishers to add provenance metadata to their datasets. These provenance metadata can be queried by consumers and used as contextual information for integration and inter-operation of information resources on the Semantic Web.

Key words: Linked Data, Public Open Data, Data Publication, Data Consumption, Semantic Web, Provenance, Data Integration.

1 Introduction

Whom do you trust? In the human realm, you must earn the trust of others, not assume it. You cannot, for example, expect them to simply hand you their money based on your assurances that you are an honourable person (although they may do that through referral). Conversely, you will give someone else your money or your information only after you have established that they will handle it appropriately (responsibly). And, how do you generate that trust? You generate that trust not by declaring "I'm trustworthy", but by revealing as much information of you as possible.

The above are examples of integration and inter-operation of transactions enabled by notions of quality and trust. Similar examples can be found on the Web. A user, on the Web, may be confronted with a potentially large number of diverse data sources of variable maturity or quality, and selecting the high quality and trustworthy data that are pertinent for their uses and integrating these together may be difficult. The advent of Linked Data brings the potential to expose data on the Web, pointing towards a clear trend where users will be able to easily aggregate, consume, and republish data. It is therefore necessary for end-users to be able to decide the quality and trustworthiness of information at hand.

While in some cases the answer speaks for itself, in others the user will not be confident of the answer unless they know how and why the answer has been produced and where the data has come from. Users want to know about the reliability or the accuracy of the data they see. Thus, a data integration system, to gain the trust of a user must be able, if required, to provide an explanation or justification for an answer. Since the answer is usually the result of a reasoning process, the justification can be given as well as a derivation of the conclusion with the sources of information for the various steps.

Provenance, also known as lineage, describes how an object came to be in its present state, and thus, it describes the evolution of the object over time. Provision of the provenance of information resources on the Web can be used as a basis for the assessment of information quality, improving the contextual information behind the generation, transformation, and integration of information on the Web.

2 Provenance

There are two major research strands of provenance in the literature: data and workflow provenance. In the scientific enterprise, a workflow is typically used to perform complex data processing tasks. A workflow can be thought of as a set of procedure steps, computer and human, that one enacts to get from the starting state to the goal state. Workflow provenance refers to the record of the entire history of the derivation of the final output of the workflow. The details of the recording vary from one experiment to another. It may depend on the goals of the experiment, or the regulatory and compliance procedures, and a number of other things. It may involve the recording of the software programs, the hardware, and the instruments used in the experiment.

Data provenance, on the other hand, is more concerned about the derivation of a piece of data that is in the result of a transformation step. It refers to a description of the origins of a piece of data and the process by which it arrives in a database.

3 Past and Current Work on Provenance

There are many surveys of existing work on provenance from workflows [2] and database[8] research communities. There have been work on the quality assessment of data that have addressed the issues of provenance[5]. There is the Open Provenance Model[13] which allows the characterisation of the dependencies between "things", and it consists of a directed graph expressing such dependencies. It is not light-weight but can be used to describe part of the provenance relationships that is a concern of a dataset publisher.

Berners-Lee's "Oh yeah?" button [3] was meant to challenge the origins, i.e. provenance, of what is being asserted and request proofs, by directly or indirectly consulting the meta-information of what is being asserted. Named graphs [6] are models that allow entire groups of graphs be given a URI and

provenance information can be attached to those graphs. The Semantic Web Publishing Vocabulary (SWP)[4] is an RDF-Schema vocabulary for expressing information provision related meta-information and for assuring the origin of information with digital signatures. It can be used within the named graph framework to integrate information about provenance, assertional status, and digital signatures of graphs. An RDF graph is a set of RDF triples, therefore an RDF graph may contain a few triples or very many. The Named Graph framework does not give a good control on the granularity of the collection of data items to attach provenance to. In this work, we do use some elements of the SWP.

The Provenance Vocabulary[9] provides classes and properties enabling providers of Web data to publish provenance-related metadata about their data. The vocabulary provides classes, called Artifacts, Executions, and Actors, that can be used to specify provenance for data access and data creation, at the triple level. An Actor performs an Execution on an Artifact. In the Provenance Vocabulary, there are different types of actors that perform different types of executions over diverse types of artifacts. Although encoding at the triple level is fine-grained and lets provenance data be attached to a triple, a big dataset may contain a large number of triples, and encoding at triple level may lead to the provenance information be much more than the actual data.

In the linked data world, data are usually collected together and provided as datasets. The provision of the provenance information of datasets' elements is an interesting problem.

4 Provenance of Linked Data and Datasets: Challenges

There are two major challenges in the provision of provenance information of linked data, viz Provenance Representation and Provenance Storage.

4.1 Provenance Storage

Provenance information can sometimes be larger than the data it describes if the data items under provenance control is fine-grained and the information provided very rich. However, one can reduce storage needs by recording data collection that are important for the operational aspects of the dataset publisher's business.

Provenance can be tightly coupled to the data it describes and located in the same data storage system or even be embedded within the data file, as advocated in tSPARQL [10]. Such approaches can ease maintaining the integrity of provenance, but make it harder to publish and search just the provenance. It can also lead to a large amount of provenance information needing to be stored. Provenance can also be stored by itself [26] or with other metadata. Once you decide how to store the provenance data, provenance representation itself is another major challenge.

4.2 Provenance Representation

There are two major approaches to representing provenance information, and these alternate representations have implications on their cost of recording and the richness of their usages. These two approaches are:

- The Inversion method: This uses the relationships between the input data, working backwards, to derive the output data, giving the records of this trace. Examples include queries and user-defined functions in databases that can be inverted automatically or by explicit functions [14]. Here, information about the queries and the output data may be sufficient to identify the source data,
- The Annotation method: Metadata of the derivation history of a data are collected as annotation, as well as descriptions about source data and processes. Here, provenance is pre-computed and readily usable as metadata.

While the inversion method is more compact than the annotation approach, the information it provides is sparse and limited to the derivation history of the data. The annotation method, however, provides more information that includes more than the derivation history of the data and may include the parameters passed to the derivation processes, the post-conditions, etc.

We advocate the use of the annotation method as it gives richer information of the data and the data set we may be interested in. The voiD (Vocabulary of Interlinked Datasets) [1] vocabulary can be employed to describe the provenance information of the data we are interested in. voiD is an RDF based schema to describe datasets. With voiD, the discovery and usage of datasets can be performed both effectively and efficiently. Using voiD also have the added benefit of storing the provenance information with the other metadata of our datasets. There are two core classes at the heart of voiD:

1. A dataset (*void:Dataset*), i.e. a collection of data, which is:
   - published and maintained by a single provider,
   - available as RDF,
   - accessible, for example, through dereferenceable HTTP URIs or a SPARQL[1] endpoint
2. The interlinking modelled by a linkset (*void:Linkset*). A linkset in voiD is a subclass of a dataset, used for describing the interlinking relationship between datasets. In each interlinking triple, the subject is a resource hosted in one dataset and the object is a resource hosted in another dataset. This modelling enables a flexible and powerful way to state the interlinking between two datasets, such as how many links there exist, the kind of links, and who made these statements.

5 voidp: Provenance Extension to voiD

In [7], a linked data(set) publisher was advised to reuse terms from well-known vocabularies wherever possible, and one should only define new terms one cannot

[1] http://www.w3.org/TR/rdf-sparql-query/

find in existing vocabularies. Reusing existing vocabularies takes advantage of the ease of bringing together diverse domains within RDF, and it makes data more reusable. By reusing vocabularies, the data is no longer isolated nor locked within a single context designed for a single use. We adhered to this advice and have made use of the following ontologies:

- Provenance Vocabulary[11],
- The Time Ontology in OWL[12],
- The Semantic Web Publishing Vocabulary[5],

In addition, the namespace for voidp is:
`@prefix voidp: <http://purl.org/void/provenance/ns/>.`

The classes are:

- `Actor`: Here, we reuse the Actor class in the Provenance vocabulary to specify an entity or an object that performs an action on a particular data item (or a data source or data set),
- `Provenance`: this class is a container class for the list of DataItem(s) we are putting under provenance control,
- `DataItem`: this class models the item of data we put under provenance control.

The properties are:

1. `activity:` this property specifies that a particular dataset has some items under provenance control,
2. `item`: specifies the item under provenance control,
3. `originatingSource`: the item's original source,
4. `originatingSourceURI`: the URI of the item's original source,
5. `originatingSourceLabel`: the label text used to describe the item's original source,
6. `certification`: if the dataset is signed, this property is used to contain the signature elements. This is an important element to prove the origin of a dataset as it is being sliced and diced during its evolution,
7. `swp:signature:` represents the signature of the dataset,
8. `swp:signatureMethod`: specifies the signature method,
9. `swp:authority`: defines the authority of the relationship between the item under provenance control and the dataset publisher,
10. `swp:valid-from` and `swp:valid-until`: these are the valid start and end dates of that (authority) relationship,
11. `processType`: specifies the type of transformation or conversion procedure carried out on the item's source, e.g. the transformation may be due to some scripts being run on the source data,
12. `prv:createdBy`: specifies the actor that executes an action on the item that is being recorded.
13. `prv:performedAt`: date when the transformation is done,

14. `prv:performedBy`: the URI of the actor that performs the recording of the provenance activity on the item.

These classes and properties are sufficient to be useful for specifying information for both workflow and data provenance.

6 Experiments and Results

Our group, the EnAKTing group[2], is dedicated to solving fundamental problems in achieving an effective web of linked data, and as part of our work, we make use of some of the United Kingdom's government data. As part of our group's work, we recently converted a set of government data files from comma-separated-values (csv) to RDF datasets.

6.1 Source Datasets

Some of these data files were:

- Mortality data:
  http://www.statistics.gov.uk/downloads/theme_population/ Table_3_Deaths_Area_Local_Authority.xls,
- Population data: *http://www.statistics.gov.uk/downloads/ theme_population/Mid-2003_Parl_Con_quinary_est.xls*,
- Energy:
  http://www.decc.gov.uk/assets/decc/statistics/regional/ road_transport/file45728.xls,
- CO_2 emission:
  http://www.decc.gov.uk/assets/ decc/ statistics/climate _change/ 1_20100122174542 _e_@@ _localregionalco2 emissionsest20057.xls,
- Crime:
  http://www.homeoffice.gov.uk /rds/pdfs09/ hosb1109chap7.xls.

We used voiD to describe these datasets. These datasets and their voiD descriptions were inserted into our RDF database, 4store[3]. The voiD descriptions used can be found at http://152.78.189.49/voidp/. The provenance elements can be seen in the voiD descriptions.

Example Scenario Query. We may be interested in an example query such as the following:

"Give the originating urls of the datasets for Robbery and female population for the County of Durham in the United Kingdom for 2004. Also give the CO_2 emission values and total energy consumption values for that same area. Only give datasets that are from the United Kingdom Home Office and from the United Kingdom's Department of Energy and Climate Change".

[2] http://enakting.org
[3] http://4store.org/

Running such a query, we are given the source urls that were stated in subsection 6.1 (Source Datasets).

7 Conclusions

The provenance of a data element can be used to elicit that datas quality and/or trustworthiness. Data quality can be used as contextual information to aid in data integration. This paper described **voidp**, a light-weight provenance extension for the voiD vocabulary that allows data publishers to add provenance metadata to the elements of their datasets, enumerating its classes and properties. These provenance metadata can be used by a data integration system or consumer for data aggregation and inter-operation.

8 Acknowledgements

This work was supported by the EnAKTing project, funded by EPSRC project number EP/G008493/1.

References

1. K. Alexander, R. Cyganiak, M. Hausenblas, and J. Zhao. Describing linked datasets. *LDOW2009*, 2009.
2. D. Artz and Y. Gil. A survey of trust in computer science and the semantic web. *Journal of Web Semantics*, 5(2), 2007.
3. T. Berners-Lee. Cleaning up the user interface. *http://www.w3.org/DesignIssues/UI.html (retrieved Nov. 2010)*, 2009.
4. C. Bizer. Semantic web publishing vocabulary (swp) user manual. *www4.wiwiss.fu-berlin.de/bizer/WIQA/swp/SWP-UserManual.pdf (retrieved Nov. 2010)*, 2006.
5. C. Bizer. *Quality-Driven Information Filtering- In the Context of Web-Based Information Systems*. VDM Verlag, 2007.
6. J. J. Carroll, C. Bizer, P. Hayes, and P. Stickler. Cleaning up the user interface. *WWW '05 Proceedings of the 14th international conference on World Wide Web*, 2005.
7. C.Bizer, R. Cyganiak, and T. Heath. How to publish linked data on the wel. *http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/ (retrieved Nov. 2010)*.
8. J. Cheney, L. Chiticariu, and W. C. Tan. Provenance in databases: Why, where and how. *Fundations and Trends in Databases*, 4(1), 2009.
9. O. Hartig. Provenance information in the web of data. *LDOW2009*, 2009.
10. O. Hartig. Querying trust in rdf data with tsparql. *Lecture Notes in Computer Science*, 5554, 2009.
11. O. Hartig and J. Zhao. Using web data provenance for quality assessment. *Proceedings of the 1st Int. Workshop on the Role of Semantic Web in Provenance Management, ISWC*, 2009.
12. J. R. Hobbs and F. Pan. An ontology of time for the semantic web. *ACM Transactions on Asian Language Processing (TALIP): Special issue on Temporal Information Processing*, 3(1), 2004.

13. L. Moreau, B. Clifford, J. Freire, J. Futrelle, Y. Gill, and e. a. P. Groth. The open provenance model core specification (v1.1). *Future Generation Computer Systems*, 2010.
14. J. Widom. Trio: A system for integrated management of data, accuracy, and lineage. *CIDR*, 2005.