Exploring Provenance in a Linked Data Ecosystem

David Corsar, Peter Edwards, Nagendra Velaga, John Nelson, and Jeff Z. Pan

dot.rural Digital Economy Research, University of Aberdeen
{dcorsar,p.edwards,n.r.velaga,j.d.nelson,jeff.z.pan}@abdn.ac.uk

Abstract. We describe our work exploring provenance within an open linked data ecosystem being developed in the travel/transport domain. We discuss techniques to infer provenance of sensor data, maintain provenance of third party data, and reference sources not available as linked data within a provenance record.

Keywords: provenance, linked data, intelligent information infrastructures, transport.

1 Introduction

In this paper, we discuss our work exploring provenance within an open linked data ecosystem being developed in the travel/transport domain[1]. Provenance is often cited as a key enabler for trusted information systems [1–3], particularly in dynamic open environments. The ecosystem we are developing integrates several open datasets published by government and online communities, with data generated by the crowd using a mobile phone app. Using these datasets, *GetThere*, a real time passenger information system, provides users with travel information such as timetables, estimated vehicle arrival times, vehicle locations, and details of network disruption. Given the diversity of datasets and providers within the ecosystem, issues such as information quality [4] and trust naturally arise. By providing a record of the agents, entities, and activities involved in producing a resource, provenance has a role to play in addressing these issues. We are exploring how provenance can be used to support the assessment of data within the ecosystem, for the purpose of ensuring passengers are provided with high quality, trustworthy information.

2 Provenance and Open Data

The diversity of data sources within the ecosystem has presented several challenges, a number of which related to provenance. Here we outline those challenges and our solutions to date.

[1] http://www.dotrural.ac.uk/irp

P. Groth and J. Frew (Eds.): IPAW 2012, LNCS 7525, pp. 226–228, 2012.

Making Implicit Provenance Explicit. Along with providing users with information, *GetThere* also asks users to act as sensors during their journeys on public transport. Observations generated by passengers are represented within the ecosystem using the Semantic Sensor Network (SSN) ontology[2]. Along with describing sensors and observations, the SSN ontology also captures implicit provenance information (i.e. provenance information not expressed using a provenance model), such as the sensor that generated the observation, the sensing method it used, and the inputs/outputs of that method.

To make this provenance information available explicitly (i.e. expressed using a provenance model), we have defined a series of axioms that map SSN concepts to PROV-O[3], the OWL encoding of the provenance interchange format being developed by the W3C Provenance Working Group[4]. These axioms define two equivalent class relationships between: *ssn:Process* (which represents sensing processes) and *prov:Activity*, and *ssn:Observation* and *prov:Entity*, along with three equivalent properties: *ssn:hasInput* and *prov:used*, *ssn:hasOutput* and *prov:generated*, and *ssn:sensingMethodUsed* and *prov:wasGeneratedBy*. These axioms allow an ontology reasoner to materialise PROV-O information for observations, inferring that an observation is an *Entity*, that the sensing process is an *Activity*, and the relevant *generated*/*wasGeneratedBy* links.

Associating Provenance with Remote Linked Data. Several of the datasets within the ecosystem are provided by third parties and accessed via remote SPARQL endpoints. Having access to the provenance of this data would provide agents with valuable additional information to use when evaluating such data. Unfortunately, in many cases provenance is not associated directly with the data. For example, the UK Government's public transport linked dataset does not include provenance; however, the web page linking to the endpoint states that the data "dates from March 2010"[5]. As other (non-linked data) versions of this dataset are updated regularly, associating provenance with the linked data would support, for example, automated assessment of timeliness. The main challenge here is determining an appropriate method for associating provenance with third party data only available via a remote SPARQL endpoint.

To address this we use the SPARQL 1.1 Service Design Ontology[6] and PROV-O to describe remote SPARQL endpoints, the data they make accessible, and the data provenance. Each endpoint description consist of individuals representing the *sd:Service* (the endpoint), the *sd:DataSet*s accessed by the endpoint, and the *sd:Graph*s within each dataset. By defining *sd:Graph* as equivalent to *prov:Entity*, we can build a provenance record for it, including a description of the *prov:Activity* that generated the graph and the data it contains.

[2] `http://purl.oclc.org/NET/ssnx/ssn`
[3] `http://www.w3.org/TR/prov-o/`
[4] `http://www.w3.org/2011/prov/wiki/Main_Page`
[5] `http://data.gov.uk/linked-data` Accessed May 2012.
[6] `http://www.w3.org/ns/sparql-service-description`

Including Non-linked Data Resources in a Provenance Record. Several datasets within the ecosystem have been derived from data originally available in formats such as CSV or HTML. For example, the ecosystem includes a bus timetable dataset created by manually scraping the timetable information from the operator's web site into a spreadsheet, which a program then converts into RDF. The main challenge here is recording this and including references to the original resources (for example, the web page), which change over time, so we cannot, for example, reference the web page URL.

We overcome this using the aforementioned technique for representing the provenance of data accessible through a SPARQL endpoint. Here, the provenance record includes details of the timetable scraping process and references to: the program that was used (in a source code repository); a copy of the file(s) used by that program; and a downloaded copy of the scraped web page. Although this record is largely created manually and necessitates storing a copy of all the files used, it does allow agents to verify the linked data by comparing it with the original source data.

3 Conclusion

In order to associate provenance with the data in the ecosystem, it has been necessary to develop various approaches that accommodate the diversity in datasets and providers. Maintaining a provenance record in this environment is challenging, particularly when dealing with data provided and maintained by third parties. The approaches we describe above illustrate how semantic web technologies can be used to provide a starting point for associating provenance with such data. We believe these approaches point the way towards a set of guidelines for provenance management in open, linked data environments.

Acknowledgements. The research described here is supported by the award made by the RCUK Digital Economy programme to the dot.rural Digital Economy Hub; award reference: EP/G066051/1.

References

1. Artz, D., Gil, Y.: A survey of trust in computer science and the semantic web. Web Semantics: Science, Services and Agents on the World Wide Web 5(2), 58–71 (2007)
2. Carroll, J.J., Bizer, C., Hayes, P., Stickler, P.: Named graphs, provenance and trust. In: Proc. of the 14th lnt. Conf. on World Wide Web, pp. 613–622. ACM, New York (2005)
3. Dividino, R., Sizov, S., Staab, S., Schueler, B.: Querying for provenance, trust, uncertainty and other meta knowledge in RDF. Web Semantics 7(3), 204–219 (2009)
4. Hartig, O., Zhao, J.: Using web data provenance for quality assessment. In: Freire, J., Missier, P., Sahoo, S.S. (eds.) Semantic Web in Provenance Management. CEUR Workshop Proceedings, vol. 526. CEUR-WS.org (2009)