Provenance Principles for Open Data

Edoardo Pignotti*, David Corsar*†, Peter Edwards*†
*Computing Science & †dot.rural Digital Economy Hub
University of Aberdeen
Aberdeen, AB24 5UA
{e.pignotti, d.corsar, p.edwards}@abdn.ac.uk

ABSTRACT

Provenance plays a vital role in enriching the context surrounding open data, and can help support assessment of attributes such as trustworthiness and quality. In this paper we introduce a set of provenance principles to provide a guideline for individuals and organisations to publish more transparent open data.

Categories and Subject Descriptors

H.1 [**Information Systems**]: Models and Principles

General Terms

Theory, Documentation, Management

Keywords

provenance, open data, linked data

1. INTRODUCTION

The emergence of Open Data sources on the Web provides applications and services with a wealth of data which they can use to deliver services, potentially providing socio-economic benefits for all. The concept of the Web of Linked Data [2] provides a means to expose, connect and share information on the Web identified by URIs using RDF[1] as a data model. Examples include the *data.gov.uk* initiative which aims to expose UK public data, and *bio2rdf.org* which provides an atlas of post-genomic data. However, the Web of Data still suffers from many of the same problems as the Web of documents in terms of information quality, trust, attribution, etc. which is essential for ensuring high-quality applications and services.

An illustration of this is reflected in the following quote from the chairman of the UK Audit Commission Michael O'Higgins on the day that government spending data was released in November 2010: "And that's where I think the critical issue is - that what is being released is not in fact information, it is data. And data needs context to become information, and it is provision of that context that will be important."

Provenance plays a vital role in enriching the context surrounding open data, and can provide additional evidence to support assessment of attributes such as trustworthiness and quality. Provenance (also referred to as lineage or heritage) aims to provide additional documentation about the processes that led to the creation of a resource [4]. Goble [3] expands on the Zachman Framework [7] by presenting the '7 W's of Provenance': *Who, What, Where, Why, When, Which, & (W)How.* Each of these provides a unique type of provenance information which can be used individually or in combination with others to support the assessment of trustworthiness and quality of open data.

The Provenance and Linked Open Data mini-theme[2] was an activity supported by the UK e-Science Institute[3] which was investigating provenance challenges in the context of Linked Open Data. As an outcome of a series of workshops organised under this activity we have identified and discussed a set of principles for publishing provenance of open data similar to the Linked Open Data rules discussed by Berners-Lee[1]. These provenance principles are "expectations of behaviour" and therefore breaking them does not destroy anything but misses an opportunity to make data more transparent. In the remainder of this paper we introduce the provenance principles and discuss how such principles can provide a guideline for individuals and organisations to publish the provenance of open data.

2. PROVENANCE PRINCIPLES

The provenance principles are summarised as follow:

☆	Publish the provenance (7 W's) of data on the web whatever format (e.g. plain text).
☆ ☆	Publish provenance as structured data (e.g. database, spreadsheet, XML)
☆ ☆ ☆	Use URIs to identify individual elements within the provenance record.
☆ ☆ ☆ ☆	Link provenance record to other provenance records using RDF.

To illustrate the use of the provenance principles introduced in this paper we use an example dataset from National Public Transport Access Nodes[4] (NaPTAN). NaPTAN it is one of the few 5-star data datasets (according to the Berners-Lee

[1] http://www.w3.org/RDF/

[2] http://wiki.esi.ac.uk/Provenance_and_Linked_Open_Data

[3] http://www.esi.ac.uk/

[4] http://www.dft.gov.uk/naptan/

Digital Engagement '11 ,
November 15 – 17, 2011, Newcastle, UK

1 Star: *"National Public Transport Access Nodes (NaPTAN), Posted by Department for Transport on 11/07/2011"*

2 Star: Looking up `http://data.gov.uk/api/catalogue/rest/package/naptan`
returns (JSON object): `{ title : "National Public Transport Access Nodes (NaPTAN)", published_by : "Department for Transport [11406]"}`

3 Star: Looking up `http://data.gov.uk/api/catalogue/rest/package/naptan/extras/published_by`
returns the string *"Department for Transport [11406]"*

4 Star: Looking up `http://data.gov.uk/api/catalogue/rest/package/naptan/extras/published_by`
returns the URI `http://reference.data.gov.uk/doc/department/dft`

Figure 1: Examples of 1, 2, 3, and 4 star provenance.

principles) published by data.gov.uk to-date. Figure 1 uses examples from this dataset to illustrate the principles.

The first principle is to publish the provenance of data on the web, in any format. This may be, for example, as plain text on a Web page or as part of an existing document. The main benefit of publishing one star provenance, is that it provides users with some information which can be used for gaining a better understanding of the context surrounding the data. However, due to the lack of structure it is difficult for applications to use this information.

The second principle is to publish the provenance of data using a structured data format such as CSV, JSON, XML, DB etc. Most of the datasets on data.gov.uk fall into this category, as they provide provenance information about release date, last update, and the publisher as JSON and HTML.

The third principle is to use URIs to identify elements within the provenance record and to provide provenance information when someone looks up that URI. Publishing provenance in this way facilitates both applications and users to access specific pieces of the provenance record by referring to their published URI. There are currently no examples of such provenance on data.gov.uk.

The fourth principle is to publish the provenance record using RDF and to link it to external provenance records and datasets (when available) thus extending the provenance record. Once again, this type of provenance does not exist for datasets published on data.gov.uk. The benefit of four star provenance data is that it provides a framework (with the support of appropriate ontologies) for applications to reason about the provenance record by following explicit provenance links. A number of existing RDF-based provenance models could be used to support such reasoning including OPM [5], The Provenance Vocabulary[5] and the Provenir ontology [6].

3. DISCUSSION

Provenance is important to the digital economy because it provides the context necessary for users and services to utilise open data available on the Web more effectively. Provenance can provide crucial evidence for supporting the assessments of quality, reliability and trustworthiness of information; all issues that are important in open systems.

While the provenance principles introduced here provide a guiding framework for people and institutions to publish provenance, many issues remain. The mini-theme identified some of these in relation to Open Data as follows: identity - provenance must be unambiguously associated with a resource not only now, but into the future; granularity of description - provenance may be associated with a complex resource or with individual components of that resource.

Adherence to the principles introduced in this paper would allow users of open data to judge for themselves if the data were suitable for their intended application. While the most desirable scenario would be to deliver a 5-star data set (according to the Berners-Lee principles) combined with 4-star provenance, we argue that even a 2 star data-set could be enhanced by the provision of 3 or 4 star provenance.

In the future it will be important to identify a de-facto standard for representing 4 star provenance in RDF. The W3C Provenance Working Group[6] is currently working towards this goal. Moreover, the general user or developer might benefit by a standard set of APIs to create, query and visualise 4 star provenance information.

4. ACKNOWLEDGEMENTS

The research described here was supported by the UK eScience Institute, by the UK Economic & Social Research Council (ESRC) under the Digital Social Research programme (RES-149-25-1075) and by the RCUK Digital Economy Programme award for the dot.rural Digital Economy Hub (EP/G066051/1).

5. REFERENCES

[1] T. Berners-Lee. Linked Data. *IJSWIS*, 4(2):1, 2006.

[2] C. Bizer, T. Heath, and T. Berners-Lee. Linked data - the story so far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 2009.

[3] C. Goble. Position statement: Musings on provenance, workflow and (semantic web) annotation for bioinformatics. Workshop on Data Derivation and Provenance, Chicago, 2002.

[4] P. Groth, S. Jiang, S. Miles, S. Munroe, V. Tan, S. Tsasakou, and L. Moreau. An architecture for provenance systems. Technical report, ECS, University of Southampton, 2006.

[5] L. Moreau, B. Clifford, J. Freire, J. Futrelle, Y. Gil, P. Groth, N. Kwasnikowska, S. Miles, P. Missier, J. Myers, B. Plale, Y. Simmhan, E. Stephan, and J. V. den Bussche. The open provenance model core specification (v1.1). *Future Generation Computer Systems*, 27(6):743 – 756, 2011.

[6] S. Sahoo and A. Sheth. Provenir ontology: Towards a framework for escience provenance management. *Microsoft eScience Workshop*, 2009.

[7] J. A. Zachman. A framework for information systems architecture. *IBM Systems Journal*, 26(3):276–292, 1987.

[5] http://trdf.sourceforge.net/provenance/ns.html

[6] http://www.w3.org/2011/prov/wiki/Main_Page