Provenance on the Web, Leaving the Walled Garden Behind . . .

Peter Edwards
Computing Science & dot.rural
Digital Economy Hub
University of Aberdeen
Aberdeen, UK
p.edwards@abdn.ac.uk

Edoardo Pignotti
Computing Science
University of Aberdeen
Aberdeen, UK
e.pignotti@abdn.ac.uk

David Corsar
Computing Science & dot.rural
Digital Economy Hub
University of Aberdeen
Aberdeen, UK
dcorsar@abdn.ac.uk

ABSTRACT
Provenance has been identified as essential for the development of a more trustworthy machine-processable web. We discuss issues associated with provenance on the Web by comparing two different systems, a closed e-science platform and a more open public transport information system.

Categories and Subject Descriptors
H.1.0 [**Information Systems**]: Models and Principles—*General*; H.m [**Information Systems**]: Miscellaneous

General Terms
Design, Management

Keywords
Provenance, Semantic Web

1. INTRODUCTION
In recent years, the Web has drastically altered the way in which information and services are exchanged between individuals and organisations. However, these exchanges are often predicated on little or no information as to the quality, reliability or authenticity of content or individuals [1]. Provenance has been identified by the database and workflow communities as an essential step in supporting reliability, discovery, and trust of online information [8].

While many of the existing provenance solutions have focused on specific technologies to support narrow scientific domains, some recent research has focused on interoperability of provenance information across different systems. The International Provenance and Annotation Workshop series[1] initiated a number of challenges with the aim of enhancing understanding of, and establishing interoperability between existing systems. The result was an agreement that a core provenance representation was needed, culminating in the development of the Open Provenance Model (OPM) [4].

More recently, the W3C Provenance Working Group[2] has been established to develop standards that support the widespread publication and use of provenance information of Web documents, data, and resources. The working group aims to provide recommendations that define a language for exchanging provenance information among applications. A parallel activity supported by the UK e-Science Institute[3] is the Provenance and Linked Open Data mini-theme[4] which is investigating the provenance challenges of linked open data.

In this paper we consider provenance on the Web and its contrasting role within two different systems: a closed e-science platform which is highly ordered and is able to assume a degree of self-policing by users (the walled garden) vs a more open public transport information system which is prone to incorrect, incomplete or even deliberately false data, and which must function in real-time.

2. THE WALLED GARDEN
Researchers are increasingly turning to the use of online tools to collaborate and share resources: one such system is the ourSpaces[5] virtual research environment. This system has been developed to facilitate collaboration and interaction between researchers via the linking of research artefacts, projects, people and their social networks. Provenance in ourSpaces is crucial in order to support transparency and accountability of the research process by documenting the derivation history of research artefacts. At the core of ourSpaces [6] is a provenance framework based on OPM, which provides a specification to express data provenance, process documentation and data derivation. It is based on three primary entities namely *Artifact*, *Process* and *Agent* and associated causal relationships: *used*, *wasGeneratedBy*, *wasTriggeredBy*, *wasDerivedFrom* and *wasControlledBy*. The description of a digital resource in ourSpaces is provided by a provenance ontology, which defines the primary entities of OPM and additional provenance ontologies which extend the concepts defined in the OPM ontology with domain specific classes. In ourSpaces, the link between social network activities and digital artefacts is established

[1] www.ipaw.info/

[2] http://www.w3.org/2011/prov/wiki/Main_Page
[3] http://www.esi.ac.uk/
[4] http://wiki.esi.ac.uk/Provenance_and_Linked_Open_Data
[5] http://www.ourspaces.net

formally in order to obtain a full and transparent provenance representation. This is achieved by the integration of the FOAF[6] social networking vocabulary and the SIOC[7] online communication vocabulary, with the provenance ontologies.

In this environment, the quality of the provenance record is guaranteed by a degree of self-policing deriving from norms within the research community. Nevertheless, there is still a need to manage users so they comply with certain policies imposed by projects, funding organizations or institutions. Within ourSpaces policies can be created by a user, by the administrator of a project, group or organisation or by a system developer. For example, a user may impose certain access constraints on digital artifacts that they own, e.g. certain information about the artifact may only be accessible to users who contributed towards the creation of that artifact. Such policies are defined declaratively in terms of obligations, prohibitions, and permissions.

One of the most significant challenges in ourSpaces is usability as there is need for users to be able to create provenance descriptions of processes and artifacts without knowledge of the underlying representation. Another issue relates to the integrity of the provenance data. How do we assess completeness of the record? How should broken/missing links be handled? Various strategies present themselves including use of social annotations to highlight missing information or links; or using policies about documentation or research methodology to reason about the provenance record and to alert users when there are completeness issues.

3. OPEN INFORMATION SYSTEM

In contrast to the closed environment described above, open distributed systems exhibit a different set of characteristics and so have a separate set of issues. Open systems typically consist of interacting participants that are free to join and leave at any time, are not controlled by a single authority, and are driven by different aims and objectives and so may be unreliable, [3, 5]. Issues such as trust, reputation, reliability [5], information quality [2], and privacy naturally become prevalent in such systems. Although closed environments also experience these issues, they are amplified in open systems, as they cannot rely on measures to help address these issues such as self-policing based on community norms, or multiple interactions with participants for validating or acquiring missing information.

We are investigating the role of provenance in addressing these issues within the context of an open passenger information system for rural public transport users[8]. Here, open data from government agencies is integrated using linked data principles with data from operators, frequently changing passenger social networks, and crowd-sourced transport and journey experience reports from small numbers of passengers. This data is represented using several ontologies, including FOAF, the W3C Semantic Sensor Network ontology[9], and relevant transport ontologies (e.g. NaPTAN[10]).

[6]http://www.foaf-project.org/
[7]http://sioc-project.org/
[8]http://www.dotrural.ac.uk/irp
[9]http://www.w3.org/2005/Incubator/ssn/
[10]http://www.data.gov.uk/dataset/naptan

In such an open system users may introduce imperfect data (e.g. incomplete, erroneous, or fraudulent reports), which could adversely affect system outputs, reducing user trust; provenance is thus critical in this domain to support information quality and trust evaluations. However, we cannot expect such information to be supplied directly by the user, in stark contrast with the e-science example above. We therefore have to rely upon indirect sources such as details of the device used to supply information, past user reports (including feedback from others), the user's role within social networks, etc. This in turn raises several issues, such as ensuring each user's privacy cannot be violated by reasoning with the provenance information [7], and that the system remains responsive to the dynamic environment.

4. DISCUSSION

Based on our experiences to date with the applications above, we have identified the following issues associated with implementing provenance solutions in more open systems: What new sources of provenance can be exploited? How should systems deal with imperfect provenance? How should provenance itself be validated? How can policies associated with the provenance record be created, used and enforced? We argue that finding answers to these questions is essential if provenance is to have a role in the future Web.

5. REFERENCES

[1] O. Hartig. Provenance information in the web of data, *Linked Data on the Web (LDOW2009)*, April 20, 2009, Madrid, Spain.

[2] O. Hartig and J. Zhao. Using web data provenance for quality assessment. In *Proc. of the Workshop on Semantic Web and Provenance Management at ISWC*, 2009.

[3] T. Huynh, N. R. Jennings, and N. Shadbolt. An integrated trust and reputation model for open multi-agent systems. *Journal of Autonomous Agents and Multi-Agent Systems*, 13(2):119–154, 2006.

[4] 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*, July 2010.

[5] S. Ramchurn, T. Huynh, and N. R. Jennings. Trust in multiagent systems. *The Knowledge Engineering Review*, 19(1):1–25, 2004.

[6] R. Reid, E. Pignotti, P. Edwards, and A. Laing. ourspaces: linking provenance and social data in a virtual research environment. In *Proceedings of the 19th International Conference on World Wide Web, WWW 2010*, pages 1285–1288, 2010.

[7] Y. Simmhan and K. Gomadam. Social web-scale provenance in the cloud. In *IPAW 2010, Revised Selected Papers*, volume 6378 of *Lecture Notes in Computer Science*, pages 298–300. Springer, 2010.

[8] Y. L. Simmhan, B. Plale, and D. Gannon. A survey of data provenance in e-science. *SIGMOD Rec.*, 34:31–36, September 2005.