AAAI Technical Report SS-12-06
Wisdom of the Crowd

The Crowd and the Web of Linked Data: A Provenance Perspective

Milan Markovic , Peter Edwards , David Corsar and **Jeff Z. Pan**
Computing Science & dot.rural Digital Economy Hub, University of Aberdeen, Aberdeen, AB24 5UA
{m.markovic, p.edwards, dcorsar, and jeff.z.pan}@abdn.ac.uk

Introduction

Linked Data (Berners-Lee 2006) provides a method for publishing structured, interlinked data in a machine-readable form that can be used to build intelligent applications and services. However, the usefulness of these applications/services is dependent on the availability and correctness of the data they reason with. The crowd potentially has an important role to play in performing the non-trivial tasks of creating, validating, and maintaining the linked data used by applications and services. Additional information, such as how the data were created, when, by whom, etc., can be used in these tasks and others, such as evaluating the performance of the crowd and its members. Such information can be captured in a provenance record.

In this paper we discuss the role of the crowd in creating and maintaining the web of linked data, how provenance can be used to record the crowd's actions, and the requirements this places on the provenance model.

Crowd Wisdom & the Web of Linked Data

In this section we discuss two broad roles for the crowd within the web of linked data, namely, data creation and data maintenance.

Data Creation

The crowd can be utilised to create individuals and links between datasets. Creating individuals involves the use of the crowd to create new instance data either within existing datasets or as part of new datasets. To illustrate, consider a scenario where a user can report traffic disruption (e.g. road closure) from his mobile device in real time. Figure 1 illustrates the process: an *observer* uploads a message about the traffic disruption, information is then processed by another member of the crowd (*classifier*) and linked to appropriate datasets (e.g. road information provided by government). A linked data representation of these reports is then generated to provide other members of the crowd with more complex query and reasoning abilities.

The crowd can also be used to define new links between datasets, either as alignments (i.e. defining equivalent concepts) or as new relationships between previously unlinked concepts. For example, Spotthelink (Thaler, Simperl, and Siorpaes 2011) uses the crowd to define alignments between two ontologies.

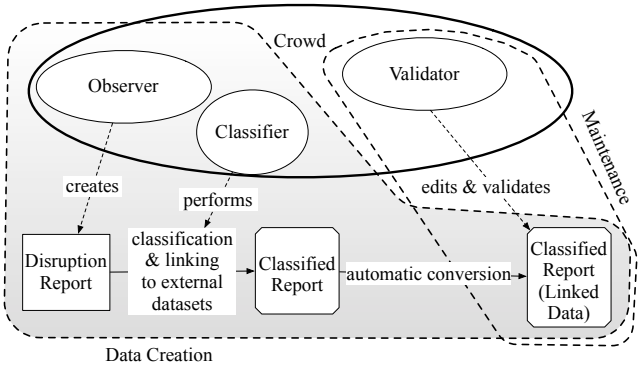

Figure 1: Linked data generation & validation activities associated with the traffic disruption scenario.

Data Maintenance

Two data maintenance tasks that can be performed by the crowd are validation and editing. Here validation involves members of the crowd (*validators*) evaluating data and annotating them according to some quality or correctness vocabulary. Editing is then the process of revising data that has been previously annotated as being of poor quality or incorrect. In the traffic disruption scenario mentioned earlier, in addition to data creation (and classification) the crowd can be used to validate and edit data generated within the system. This involves correcting/updating the data (e.g. if the title of the report does not represent its actual content) and also adding additional information. The crowd is thus capable of resolving situations with contradicting reports (e.g. by assigning different quality annotation to these reports).

Role of Provenance

We adopt the W3C Provenance Incubator Group[1] definition of provenance: "a record that describes entities and processes involved in producing and delivering or otherwise in-

[1] http://www.w3.org/2005/Incubator/prov/XGR-prov-20101214/

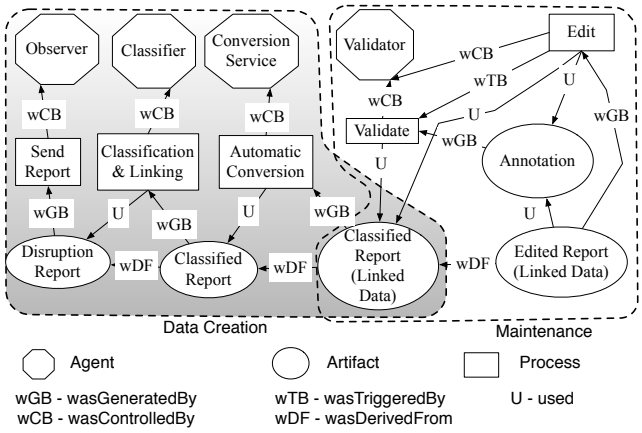

Figure 2: A provenance representation of the linked data generation & validation processes described in Figure 1.

fluencing a resource" (Gil et al. 2010). In the crowd context we interpret this as maintaining a record of the LD generated/maintained by the crowd and the process(es) involved. Figure 2 illustrates how the LD creation and validation processes described in Figure 1 can be characterised using the Open Provenance Model (Moreau et al. 2010). OPM is a generic model for representing provenance, in terms of processes (e.g. sending reports, linking reports), artifacts used and generated by those processes (e.g. a report), and the agents controlling these processes (e.g. an observer).

Previous research has identified provenance as essential for supporting information discovery and assessments such as reliability and quality (Simmhan, Plale, and Gannon 2005). Let us consider the example presented in Figure 2: the provenance record provides an audit trail that can support, for example the discovery of *classifiers* who generate reports that are frequently edited by *validators*, which in turn may form part of a reliability assessment of those crowd members, and assessment of the quality of their outputs. These types of analysis can aid processes such as selecting a workforce for future applications, or monitoring/evaluating crowd performance, particularly important when using small crowds where the negative effects of unreliable workers have a potentially greater impact than within larger crowds. The ability to achieve such assessments is directly influenced by the existence of a provenance record.

There are several issues associated with use of provenance in this way. We can expect the amount of information supplied to vary greatly (e.g. compare the description of the "Disruption Report" and "Annotation" artifacts in Figure 2). Generating the provenance graph (record) is also challenging, as it may require: ensuring links are correctly generated; referencing items not published as LD; referencing individual triples; and referencing triples deleted as part of an edit performed during maintenance.

Provenance Model Requirements

To support the use of provenance in the context of crowdsourcing and LD as discussed above, we have identified the following set of requirements that the chosen provenance model must meet:

- Ability to model objects (artifacts), their creators (agents), and the processes involved, as these form the key elements of the provenance graph.
- Support recording of the temporal context, to enable, for example, ordering of the provenance graph based on a time line and to support reasoning about the provenance graph as a frequently updating data stream.
- Ability to refer to objects that are not published as LD but that are involved in the provenance record, for example the "Disruption Report" artifact in Figure 2, in order to create the most complete provenance record possible.
- Incorporate a description of the agent's intent to provide insight into why the agent performed a process or created a particular artifact.
- The crowd is likely to generate large scale provenance records, and so the provenance model must be lightweight in order to enable real time reasoning using those records.

Discussion

We recognise that existing provenance models, such as OPM meet some of these requirements, e.g. modelling objects, agents and the processes involved. Other work (Pignotti et al. 2011) has begun to extend OPM with descriptions of intent. As part of our future work, we plan to investigate if OPM can be extended to meet our remaining requirements.

Research questions are thus: *Can the provenance record be used to identify reliable/unreliable members of the crowd?*; *What are the practical challenges of embedding provenance information in a real time crowd sourcing application?*; *What is the most appropriate provenance model for use in real time crowd sourcing applications?*

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

References

Berners-Lee, T. 2006. Linked data; accessed:10/01/2012; http://w3.org/designissues/linkeddata.html.

Gil, Y.; Cheney, J.; Groth, P.; Hartig, O.; Miles, S.; Moreau, L.; and Silva, P. P. D. 2010. Provenance xg final report. *Final Incubator Group Report*.

Moreau, L.; Clifford, B.; Freire, J.; Futrelle, J.; Gil, Y.; Groth, P.; Kwasnikowska, N.; Miles, S.; Missier, P.; Myers, J.; Plale, B.; Simmhan, Y.; Stephan, E.; and Van de Bussche, J. 2010. The open provenance model core specification (v1.1). *Future Generation Computer Systems*.

Pignotti, E.; Edwards, P.; Gotts, N.; and Polhill, G. 2011. Enhancing workflow with a semantic description of scientific intent. *Web Semantics: Science, Services and Agents on the World Wide Web* 9(2):222–244.

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

Thaler, S.; Simperl, E.; and Siorpaes, K. 2011. Spotthelink: Playful alignment of ontologies. In *6th Conference for Professional Knowledge Management*.