Quality Assessment, Provenance, and the Web of Linked Sensor Data*

Chris Baillie, Peter Edwards, and Edoardo Pignotti

Computing Science & dot.rural Digital Economy Research, University of Aberdeen
Aberdeen, UK
{c.baillie,p.edwards,e.pignotti}@abdn.ac.uk

Abstract. This paper presents a quality assessment framework for linked sensor data and discusses a role for provenance in quality assessment.

Keywords: provenance, linked data, quality assessment.

1 Introduction

In this paper we describe a framework for evaluating the quality of linked data and discuss how the provenance of such data could be introduced to the quality assessment process. The open nature of the Web enables anyone (or any 'thing') to publish any content that they choose which means that poor quality data can quickly propagate [1]. Therefore, a mechanism to assess quality is essential if agents (human or machine) are to identify reliable data to support tasks such as decision making and planning.

Data is generally regarded as high quality if it is 'fit for use' in that it meets a number of requirements [2]. These requirements place constraints on certain *quality dimensions* (e.g. *accuracy*, *timeliness*, *relevance*) and are described using *quality metrics* (e.g. *timely* data is no more than 10 minutes old). Quality assessments guided by such metrics often require additional metadata describing the context around data, something which can be provided by publishing information as Linked Data. We argue that this context should also include provenance information, a record of the entities and processes involved in data derivation. Provenance has been identified as an essential step in helping users to better understand, trust, reproduce, and validate data [3]. We argue that it should therefore also play an important role in evaluating data quality. Given the scope of the Web, we are investigating quality issues within the Web of Linked Sensor Data [4], a subset of the Web of Linked Data comprising semantic representations of sensors and their observations. In this paper, we provide a motivating example before describing the implementation of our quality assessment framework. We then discuss how provenance can be included in quality assessment and outline our future plans.

* 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.

P. Groth and J. Frew (Eds.): IPAW 2012, LNCS 7525, pp. 220–222, 2012.

2 Quality Assessment Framework

To help us better understand the requirements of quality assessment and the potential for provenance we have examined a number of scenarios. One of these uses data from the mobile phones of public transport users to provide details such as vehicle location, speed, etc. Examples of quality metrics in this scenario include *relevance*, examining the distance between the observation and the accepted route of travel, and *timeliness*, examining how old the observation is when it is used. Following an analysis of this scenario, and others, we have developed a number of requirements for a quality assessment framework: **(1)** data should be evaluated against a number of quality dimensions, **(2)** quality metrics are necessary to guide assessments, and **(3)** quality assessment results should be recorded to enable their re-use.

We represent sensor observations using the W3C Semantic Sensor Network Incubator Group's ontology[1]. This enables us to describe the context around sensor observations such as observed phenomena (e.g. location or speed in the transport scenario) and features of interest (e.g. a journey). Quality assessment is represented using the Data Quality Management[2] (DQM) ontology. *Quality metrics* are defined using SPIN[3] and attached to instances of *dqm:DataRequirement* (requirements 1 and 2); the results of assessment are captured using instances of *dqm:QualityScore* (requirement 3). We associate sensor observations (and their values) to quality scores via the *dqm:plainScore* property. (Fig. 1).

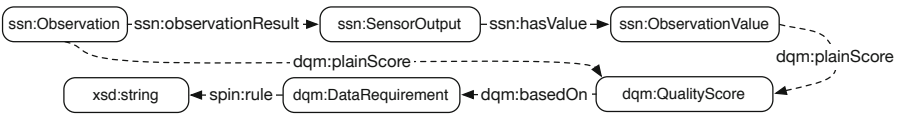

Fig. 1. Quality assessment characterised using the SSN and DQM ontologies

To evaluate our methodology we have developed a number of web services that enable the creation and manipulation of linked data representing sensor observations, quality annotations, and their provenance. The observation service generates RDF representations of sensor observations received from a smartphone app and stores them in a triple store. The quality assessment service takes the URI of an observation as a parameter and performs the evaluation using a SPIN reasoner. Guided by a number of data requirements, the reasoner produces quality scores for the observation and stores them in a dedicated quality annotation triple store, enabling the re-use of quality results. The observation and quality assessment services make use of a provenance service to document how

[1] `http://www.w3.org/2005/Incubator/ssn/ssnx/ssn`
[2] `http://semwebquality.org/dqm-vocabulary/v1/dqm`
[3] `http://www.spinrdf.org`

observations and quality annotations were created. This service uses the W3C Provenance Working Group's Prov-O model[4] to represent sensor observations as instances of `Entity` (physical, digital, or conceptual 'things' that one can provide provenance for) and sensing processes as an `Activity` (something that occurs over time and acts upon or with entities). Similarly, we represent data requirements and quality scores as `Entities`, and the quality assessment process as `Activities`. The next section outlines how this provenance information can be used in future quality assessments.

3 Provenance and Quality Assessment

At present, our framework evaluates quality using only the metadata associated with sensor observations. We have identified a number of ways in which provenance can impact upon data quality such as: the reputation of the agent responsible for creating the data, the type of device that created the data, and how the data has been transformed since it was created (e.g. rounded numeric values or type conversion). We therefore intend to implement new quality metrics that can examine observation provenance and make assessments of quality based on this information in addition to the wider contextual information. We also intend to investigate how the provenance of existing quality scores can be used to decide if existing quality assessment outcomes can be re-used instead of executing new ones. We have identified a number of use cases in which this could occur, such as agent A re-using a quality score that was generated after an assessment using agent B's data requirements because they are in the same social network and trust each other, or agent A re-using a quality score generated from a data requirement that matches one of its own requirements.

Finally, we acknowledge that quality assessment is highly subjective and therefore intend to allow users (or their agents) to define their own data requirements that will supersede the system-wide requirements that are in place at the moment.

References

1. Baillie, C., Edwards, P., Pignotti, E.: Assessing quality in the web of linked sensor data. In: 25th Conference on Artificial Intelligence (AAAI 2011), pp. 1750–1751. AAAI Press (August 2011)
2. Furber, C., Hepp, M.: Using semantic web resources for data quality management. In: 17th International Conference on Knowledge Engineering and Knowledge Management, pp. 211–225 (2010)
3. Miles, S., Groth, P., Munroe, S., Moreau, L.: Prime: A methodology for developing provenance-aware applications. ACM Transactions on Software Engineering and Methodology 20(3), 39–46 (2009)
4. Page, K.R., De Roure, D.C., Martinez, K., Sadler, J.D., Kit, O.Y.: Linked sensor data: Restfully serving RDF and GML. In: International Workshop on Semantic Sensor Networks 2009, vol. 522, pp. 49–63 (October 2009)

[4] `http://www.w3.org/TR/prov-o`