Assessing Quality in the Web of Linked Sensor Data*

Chris Baillie, Peter Edwards & Edoardo Pignotti

Computing Science & dot.rural Digital Economy Hub, University of Aberdeen, Aberdeen, AB24 5UA
{c.baillie, p.edwards, e.pignotti}@abdn.ac.uk

Abstract

Assessing the quality of sensor data available on the Web is essential in order to identify reliable information for decision-making. This paper discusses how provenance of sensor observations and previous quality ratings can influence quality assessment decisions.

Introduction

The Web has evolved from a collection of hyperlinked documents to a complex ecosystem of interconnected documents, services and devices. Due to the inherent open nature of the Web, data can be published by anyone or any ‘thing’. As a result of this, there is enormous variation in the quality of information[1]. An appropriate mechanism to assess the quality of Web content is essential if users (or their agents) are to identify reliable information for use in decision-making.

There are a number of different Information Quality (IQ) assessment frameworks; many of which depict IQ as a multi-dimensional construct. Bizer and Cygniak (2009) describe nine dimensions which include accuracy, completeness and timeliness. Lee et at (2002) limit their quality assessment approach to four quadrants: soundness, dependability, usefulness and usability but further decompose these into sub-criteria similar to those of Bizer.

The concept of the ‘Web of Data’ has recently emerged (Bizer, Heath, and Berners-Lee 2009). While this incarnation of the Web is still prone to issues of Information Quality, the associated rich metadata representations (which include links to other entities) should facilitate IQ assessment. Documenting data provenance can further enhance this representation by detailing the processes and entities that were involved in data derivation; both of which may have had an impact on data quality. Miles et al (2006) have identified the documentation of data provenance as an essential step to support users to better understand, trust, reproduce and validate the data available on the Web. With a suitably detailed representation of both the data and its provenance it should then be possible to reason about its quality (Hartig and Zhao 2009).

Given the vast scope of the Web, we have chosen to investigate the provenance and IQ issues associated with the Web of Linked Sensor Data (Page et al. 2009). We have identified the W3C Semantic Sensor Network Incubator Group (SSNXG) ontology[2] as an appropriate starting point for this work as it emerged after an extensive survey of existing sensor ontologies. We also require a generic model of provenance in order to support the diverse ecosystem of sensor platforms and data. We have investigated a number of existing models for representing provenance information but found many of these to be tailored to specific domains (e.g. workflows or databases); we have therefore selected the Open Provenance Model (Kwasnikowska, Moreau, and Van den Bussche 2010) as it provides a technology-agnostic model for describing the relationships between agents, processes and data

To better illustrate the issues described above we have developed the following scenario. A user is planning a vacation but is unsure whether or not to pack an umbrella. In order to make such a decision their agent accesses a weather sensor deployed at the location they will be travelling to. The sensor publishes data regarding temperature, atmospheric pressure, rainfall and wind speed. All of these factors could influence the decision. However, the user has never used this sensor before and is unsure of the quality of the sensor’s observations: How precise are the sensor’s observations? Does the sensor consistently produce accurate observations? Are the sensor observations up-to-date?

From this and a number of similar use cases we have identified a number of research questions relevant to the issues of IQ assessment introduced in this paper. Firstly, *is it possible to reason about quality in the Web of Linked Sensor Data?* Secondly, *can the outcomes of quality assessment be represented in the Web of Linked Sensor Data?* Thirdly, *can sensor data provenance be used to facilitate quality assessment?* Finally, *can the provenance of past quality assessment activities extend the provenance graph to facilitate future quality assessment decisions?*

*The research described here was supported by the UK Engineering & Physical Sciences Research Council award for the dot.rural Digital Economy Hub (EP/G066051/1)

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

[2] www.w3.org/2005/Incubator/ssn/

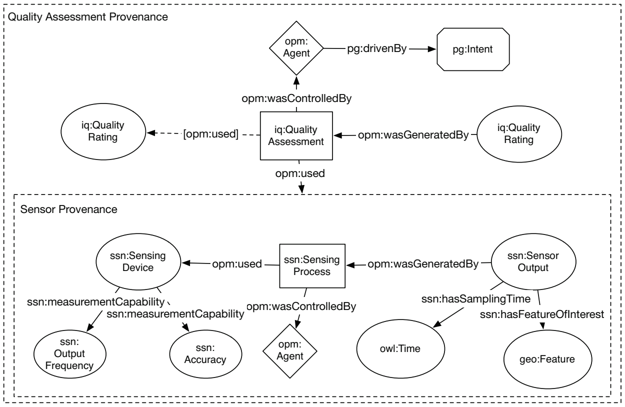

Figure 1: Our Model of Information Quality Assessment

Provenance and Linked Sensor Data

In order to characterise quality assessment in the Web of Linked Sensor data we need to extract information, such as measurement range, frequency, latency, precision, etc. from the sensor metadata. Moreover, we require a language to specify how quality should be assessed. SWRL[3] provides an initial mechanism to define such logic in the form of rules which output quality assessment metadata based on an OWL ontology which describes qualitative and quantitative IQ values. However, the measurement of quality is subjective as it relates to the ability of the information to satisfy the user's intent characterised as a combination of goals and constraints (Pignotti et al. 2010). Consider the user whose goal is to decide whether to pack an umbrella. An example of a constraint might be that they will only use sensors with precision rated at ±3°C and observations created within the last 24 hours.

While assessing sensor precision, we can discount sensing devices rated outside of the range defined by the user. Moreover, as the observation is published as linked data we can enhance our assessment by considering other linked entities. For example, if the sensor is deployed in Aberdeen (Scotland) we can compare the temperature reported by the sensor with the average temperature in Aberdeen published in another linked data source. A temperature observation of 25°C in February should be given a low quality score.

To determine another quality attribute, consistency, we must examine whether or not the sensor's observations are of a regularly occurring, dependable nature. To accomplish this, we require the sensor's provenance record which details a number of previous observations. If the observation can be considered an outlier relative to previous observations then this can be used as an indicator of quality.

We argue that completed quality assessments and their results should also be incorporated into the provenance graph to facilitate future assessments. For instance, if a previous quality assessment was performed by a trusted agent with intent similar to our own then the result of that assessment may be re-used. Figure 1 presents a conceptual view of our model of quality assessment. The *Sensor Provenance* (lower layer) characterises sensors, sensing processes and their properties while the *Quality Assessment Provenance* (upper layer) characterises the assessment process, assessment outcomes, controlling agent and their associated intent.

To explore the issues of IQ and provenance within the Web of Linked Sensor Data we have developed an infrastructure of sensing devices. These devices are based on the Arduino electronics prototyping platform and feature sensors that measure physical phenomena such as temperature, motion, light and vibration. The data transmitted by these devices are stored and made available as Linked Sensor Data[4]. Using this platform we are investigating our approach to IQ assessment through the development of an IQ reasoning apparatus driven by sensor metadata and provenance. We plan to evaluate our approach by identifying a number of case studies with which to assess the suitability of our framework. We also intend to investigate the role of policy languages such as OWL-POLAR(Sensoy et al. 2010) or AIR[5] in capturing the constraints associated with the quality assessment process.

References

Bizer, C., and Cygniak, R. 2009. Quality-driven information filtering using the wiqa policy framework. *Journal of Web Semantics* 7:1–10.

Bizer, C.; Heath, T.; and Berners-Lee, T. 2009. Linked data - the story so far. *International Journal on Semantic Web and Information Systems* 5(3):1–22.

Hartig, O., and Zhao, J. 2009. Using web data provenance for quality assessment. In Freire, J.; Missier, P.; and Sahoo, S. S., eds., *1st Int. Workshop on the Role of Semantic Web in Provenance Management*, volume 526.

Kwasnikowska, N.; Moreau, L.; and Van den Bussche, J. 2010. A formal account of the open provenance model. *(submitted)*.

Lee, Y. W.; Strong, D. M.; Kahn, B. K.; and Wang, R. Y. 2002. Aimq: a methodology for information quality assessmemt. *Information and Management* 40:133–146.

Miles, S.; Groth, P.; Munroe, S.; and Moreau, L. 2006. Prime: A methodology for developing provenance-aware applications. *ACM Transactions on Software Engineering and Methodology* 39–46.

Page, K. R.; Roure, D. C. D.; Martinez, K.; Sadler, J. D.; and Kit, O. Y. 2009. Linked sensor data: Restfully serving rdf and gml. In *International Workshop on Semantic Sensor Networks 2009*, volume 522, 49–63. CEUR.

Pignotti, E.; Edwards, P.; Gotts, N.; and Polhill, G. 2010. *Enhancing Workflow with a Semantic Description of Scientific Intent*. Journal of Web Semantics (to appear).

Sensoy, M.; Norman, T. J.; Vasconcelos, W. W.; and Sycara, K. 2010. Owl-polar: Semantic policies for agent reasoning. *International Semantic Web Conference 2010*.

[3] http://www.w3.org/Submission/SWRL/

[4] http://dtp-126.sncs.abdn.ac.uk:8081/snorql/

[5] http://dig.csail.mit.edu/TAMI/2007/amord/air-specs.html