Evaluating Quality using Sensor Metadata and Provenance*

Chris Baillie
Computing Science & dot.rural
Digital Economy Hub
University of Aberdeen
c.baillie@abdn.ac.uk

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

Edoardo Pignotti
Computing Science
University of Aberdeen
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 sensor metadata, provenance, and previous quality ratings can influence quality assessment decisions.

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

General Terms
Theory, Documentation, Management

1. INTRODUCTION
The Web has evolved from a collection of hyperlinked documents into a complex ecosystem of interconnected documents, services and devices. Vint Cerf recently stated that "*we don't know whether the information we find [on the Web] is accurate or not. So we have to teach people how to assess what they've found*". This highlights how the inherent open nature of the Web enables anyone or any 'thing' to publish any data they choose, resulting in enormous variation in quality[1]. An appropriate mechanism to assess the quality of Web content is therefore essential if users (or their agents) are to identify reliable information for use in decision-making.

There are currently a number of different Information Quality (IQ) assessment frameworks, many of which depict IQ as a multi-dimensional construct. Bizer and Cygniak [1] describe nine dimensions which include accuracy, completeness and timeliness. Lee et at [5] limit their quality assessment approach to four quadrants: soundness, dependability, usefulness and usability but further decompose these into subcriteria similar to those of Bizer.

The concept of the 'Web of Data' has recently emerged [2]. 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. Miles et al (2006) have identified the documentation of data provenance (the entities and processes associated with a data product [4]) 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 [3].

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 many of these are tailored to specific domains (e.g. workflows or databases); therefore we have selected the Open Provenance Model [4] as it provides a technology-agnostic model for describing the relationships between agents, processes and data

To better illustrate the issues described above we describe a scenario from dot.rural's Informed Rural Passenger project[3]. A user is planning a bus trip into town but is unsure when to expect the bus to arrive at their local bus-stop. To decide when to leave, their agent accesses a number of sensors deployed on the bus. These sensors publish data describing GPS location, observation accuracy, time, and direction of travel. All of these factors could influence the decision. However, the user has never queried these sensors before and is unsure about the quality of their observations: How precise are the sensors? Do they consistently produce accurate observations? Are the sensor observations up-to-date?

2. RESEARCH OBJECTIVES
To provide a focus for our research we have developed the following hypothesis: *publishing linked sensor data and its provenance provides additional context that enhances quality assessment.*

Moreover, in order to investigate this hypothesis we have identified a number of research questions: *Is it possible to reason about quality in the Web of Linked Sensor Data? Is it possible to reason about quality using provenance informa-*

*This research is supported by the award made by the RCUK Digital Economy programme to the dot.rural Digital Economy Hub; award reference: EP/G066051/1.
[1]http://www.w3.org/2005/Incubator/prov
[2]www.w3.org/2005/Incubator/ssn/
[3]http://www.dotrural.ac.uk/irp/

tion? Does the additional context provided by linked sensor data and its provenance enhance quality assessment? Can quality assessment be enhanced using the provenance of past quality assessments?

3. SENSOR NETWORK TESTBED

We have developed a sensor network testbed to provide a realistic platform for our research. The testbed features a number of sensor nodes that monitor physical phenomena such as temperature, vibration, motion and light. Each node is connected to a LAN using either ethernet or wireless connectivity and publishes its data, via a Java servlet, to a MySQL database. We have wrapped this database in a d2r server[4] to allow access to the sensor data as RDF.

We have also implemented a data visualisation Web service[5] giving access to our quality assessment mechanism. This allows a user to view data from a sensor (e.g. Temperature or Vibration) on a particular sensor node within a given time window. Graphs are produced displaying the observations generated by the selected sensor during the required time window. Clicking individual points on the graph causes the service to assess the quality of the selected observation. The assessment process uses a number of rules defined in terms of the concepts and relationships appearing in our quality ontology. This ontology was influenced by the earlier Qurator ontology [7] and describes how quality metrics (methods of assessment) can evaluate quality dimensions using certain quality indicators (values representing certain dimensions of data quality). Our first set of rules are used to identify quality indicators; for example, from the current time and the time the observation was created we produce the *data age* quality indicator. Our second set of rules describe quality metrics; for example, if the data age indicator shows the data as more than 12 hours old we can infer that the timeliness of the data is poor. We recognise, however, that these constraints are subjective and will vary depending on task and agent. The results of these quality assessments are attached to the observation's linked data graph as quality annotations so that they can later be used for either display purposes or perhaps further reasoning.

4. FUTURE WORK

In addition to exploring further quality metrics with our existing service, we intend to investigate how capturing the provenance of sensor data can be used to evaluate sensor data quality. For example, to determine another quality attribute, consistency, we must examine whether 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 could be used as an indicator of quality.

While our initial framework of quality assessment (Figure 1) will only provide default methods for evaluating data against quality indicators we acknowledge that this process is subjective. As such, we aim to develop a future version of the quality assessment model that will make use of user defined policies to tailor the assessment to individual preferences. Consider the user whose goal is to catch the bus into town. An example of a policy they might provide is that their agent will only use observations created within the past 10 minutes and with associated error margin less than ± 15 metres.

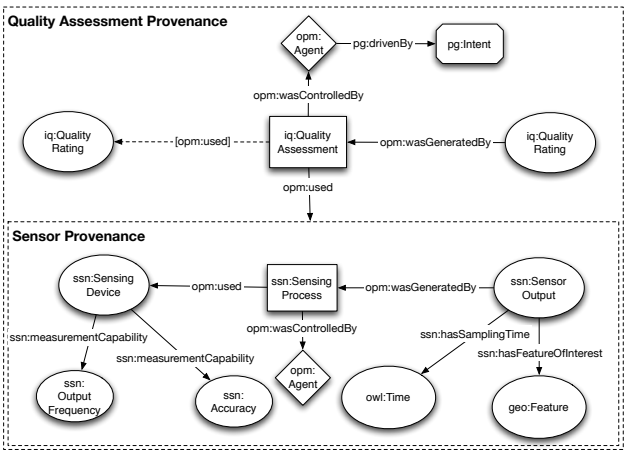

Figure 1: Our framework for quality assessment.

We are also investigating how the provenance of past quality assessment can be used to enhance the performance of future quality assessments. For example, if a trusted agent with similar intent (characterised as goals and constraints [6]) has previously assessed an observation then we could adopt the existing quality rating rather than compute a new one.

5. EVALUATION

We have developed a number of use cases that will be used to inform our methods of evaluation. These use cases will detail a number of scenarios in which the quality of sensor data must be assessed in order to complete certain tasks. From these scenarios we will develop a number of simulation models to enable us to explore our use cases and to assess the performance of our methods of quality assessment.

6. REFERENCES

[1] Christian Bizer and Richard Cygniak. Quality-driven information filtering using the wiqa policy framework. *Journal of Web Semantics*, 2009.

[2] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data - the story so far. *International Journal on Semantic Web and Information Systems*, 2009.

[3] Olaf Hartig. Provenance information in the web of data. In *Linked Data on the Web*, April 2009.

[4] Natalia Kwasnikowska, Luc Moreau, and Jan Van den Bussche. A formal account of the open provenance model. *(submitted)*, 2010.

[5] Yang W. Lee, Diane M. Strong, Beverly K. Kahn, and Richard Y. Wang. Aimq: a methodology for information quality assessmemt. In *Information and Management*, volume 40, 2002.

[6] Edoardo Pignotti, Peter Edwards, Nick Gotts, and Gary Polhill. *Enhancing Workflow with a Semantic Description of Scientific Intent*, volume 9, pages 222–244. Journal of Web Semantics, 2011.

[7] A. Preece, P. Missier, S. Embury, B. Jin, and M. Greenwood. *An ontology-based approach to handling information quality in e-science*, volume 20, pages 253–264. Wiley, 2008.

[4]http://www4.wiwiss.fu-berlin.de/bizer/d2r-server/

[5]http://dtp-126.sncs.abdn.ac.uk:8080/SensorNet/graph.jsp