Quality Reasoning in the Semantic Web

Chris Baillie, Peter Edwards, and Edoardo Pignotti

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

Abstract. Assessing the quality of data published on the Web has been identified as an essential step in selecting reliable information for use in tasks such as decision making. This paper discusses a quality assessment framework based on semantic web technologies and outlines a role for provenance in supporting and documenting such assessments.

Keywords: provenance, linked data, quality assessment.

1 Background

In recent years the World Wide Web has evolved from a collection of hyperlinked documents [3] to a vast ecosystem of interconnected documents, services and even people. Content on the Web suffers from a range of issues associated with data quality [5], as illustrated by this quote from one of the founders of the Internet:

> "The problem is - we don't know whether the information we find [on the Web] is accurate or not. We don't necessarily know what its provenance is. So we have to teach people how to assess what they've found. [...] there's so much juxtaposition of the good stuff and not-so-good stuff and flat-out-wrong stuff or deliberate misinformation or plain ignorance."
> *Vint Cerf, July 2010*

This highlights how the open nature of the Web enables anyone or any 'thing' to publish any content they choose. As a result, poor quality data can quickly propagate[1] and appropriate mechanisms to assess the quality of Web content are essential if agents (people or software) are to identify reliable information for use in tasks such as decision making and planning. Given the scope of the Web we have chosen to investigate these issues within the Web of Linked Sensor Data [11], a subset of the Web of Linked Data comprising semantic descriptions of sensors and their observations. Current examples of quality assessment frameworks such as Bizer and Cygniak's WIQA [2], and Lee et al's AIMQ [9] assess quality by examining data against a number of *quality dimensions* such as `accuracy`, `timeliness`, and `relevance` as defined by a number of *quality*

[1] `http://www.w3.org/2005/Incubator/prov/wiki/Use_Case_Report#Information_Quality_Assessment_for_Linked_Data`

P. Cudré-Mauroux et al. (Eds.): ISWC 2012, Part II, LNCS 7650, pp. 383–390, 2012.

metrics. Assessments such as these often require additional metadata describing the context surrounding data (e.g. the characteristics of the sensor or the phenomenon measured by the observation), something that can be provided by publishing linked data [3]. We argue here that this context should also include provenance information, a record of the entities and processes involved in data derivation, as this has been identified as an essential step to support users to better understand, trust, reproduce, and validate the data available on the Web [10]. Provenance should therefore play a key role in evaluating quality as it provides information about data sources, the method used in data creation, and how the data has transformed over time - including who had access to the data, who processed it, and how the data was previously assessed.

Patel-Schneider and Fensel [12] describe Berners-Lee's vision of a semantic web language stack comprising different layers, each providing an intermediate language standard. This stack uses XML as a base standard for representing metadata, each layer above this base then adds new capabilities for expressing semantics. At the top of this stack is a layer dedicated to trust, famously illustrated by Berners-Lee's "*Oh, yeah?*"[2] button, which asks the Web "*how do I know I can trust this information?*". Richardson et al [13] describe trust as "*belief in a statement [...] A high value means that the statement is accurate, credible and/or relevant*", dimensions which are similar to those identified as important in evaluating the quality of data. This suggests that quality assessment should play an important role in the semantic web stack, either as a layer on its own or as a sub-component of the trust layer.

In our work to date we have investigated a number of application scenarios that employ sensors such as transport telematics, physiological monitoring in healthcare, and environmental conservation. In the first of these scenarios a crowdsourcing system is used to generate data describing the locations of public transport vehicles. The system relies on passengers activating a smartphone app that monitors their location using the phone's built-in GPS receiver. Other users can then use this system to discover when the next bus will arrive at their local bus stop. There are a number of possible sources for low quality data in this scenario, including poor mobile phone network coverage, degradation of the GPS signal, and malicious users. Being able to evaluate the quality of data is essential if this service is to be reliable and trustworthy.

To provide a focus for our research we have developed the following hypothesis: *publishing semantic descriptions of data and their provenance provides additional context that enhances quality assessment.* There are two key elements here: *context* refers to metadata describing the situation in which the observation was created and its provenance, such as the observed phenomenon (e.g. temperature), the feature of interest (e.g. a city), or the agent that controlled the sensing process; *enhancements* refer to how quality assessment is improved or new forms of assessment are enabled.

We have identified three potential enhancements: a) being able to evaluate a wider range of quality dimensions; b) being able to include a wider range of

[2] `http://www.w3.org/DesignIssues/UI.html`

data properties while evaluating individual quality dimensions; and c) being able to reduce the time taken to evaluate quality by re-using results from previous quality assessments. These are described in greater detail in section 4.

2 Related Work

Recent years have witnessed growing interest in semantic sensor networks. For example, the Open Geospatial Consortium ran a Sensor Web Enablement initiative [4] which aimed to develop a number of standard encodings for sensor measures. Le-Phouc and Hauswirth [8] built upon this, illustrating how linked sensor data can be published by following the linked data principles. This enables links to other datasets that provide additional contextual information about the original data. For example, observations from a GPS device can link to data describing the transport route that a vehicle should be using. There are a number of existing ontologies describing sensors and their observations [1,7]. The W3C Semantic Sensor Networks Incubator Group developed its own ontology[3] after a survey of these existing sensor ontologies and represents a state-of-the-art model describing sensor networks. However, while these ontologies are suitable for describing sensors and their observations, they provide only minimal observation provenance in the form of a description of the sensing method used to produce the observation. We argue that this is insufficient as there is more to provenance than just the process that created the observation, including details of the agent that controlled the process and the entities that were used by the process (e.g. the sensing device).

Quality assessment is the process of determining how suitable a piece of information is for a particular use and is performed by evaluating data against a number of *quality dimensions* such as `accuracy`, `timeliness`, and `relevance`. Bizer and Cygniak's WIQA framework [3] is a collection of software components that perform quality assessment using a number of *quality metrics* to examine data content, its context, and any associated external ratings. To our knowledge, the WIQA framework does not enable users to author their own policies to guide the information filtering process. We argue that this is key to any quality assessment framework because quality is highly subjective and task dependent.

Hartig and Zhao [6] present an approach to using provenance information about the data on the Web to assess its quality and trustworthiness. Their solution identifies provenance elements and the relationships between them. These elements represent specific provenance information such as the data producer or the process of data creation. Once the provenance graph has been generated, this data can be used in order to assess information quality by assigning *impact values* to the nodes, representing how processes and agents may have influenced data quality. Again, this solution does not enable users to define their own quality metrics.

There is no consensus on how quality metrics should be defined. Furber and Hepp [5] describe the use of SPARQL rules to guide quality assessment. Their

[3] `http://www.w3.org/2005/Incubator/ssn/`

model of quality assessment (DQM) has provision for a limited number of quality dimensions (currently `timeliness`, `accuracy`, `completeness`, and `uniqueness`). Having analysed a number of real application scenarios we have requirements for dimensions that are not defined in DQM such as `availability` (the time between the observation being created and published on the server) and `relevance` (the extent to which the observation describes the phenomenon in which we are interested).

3 Work to Date

To provide a realistic platform for our research we have developed a basic sensor network framework that can receive input from Arduino[4] based sensors and also smartphones. Observations are transmitted as a JSON[5] string to the observation web service, which uses the W3C Semantic Sensor Network Incubator Group ontology to create a semantic representation, in RDF, of the observation. The example in Figure 1 illustrates a sensor observation described using this ontology. Sensors are characterised using instances of `ssn:Sensor` and their observations using a combination of `ssn:Observation`, `ssn:ObservationValue`, and `ssn:SensorOutput`. The SSN ontology provides a number of properties that enable us to describe certain aspects of the context in which the observation was created. For example, `ssn:observationSamplingTime` allows us to describe when the observation value was originally measured and `ssn:observationResultTime` can describe when the observation was made available. We can also describe the `ssn:Property` (the phenomenon measured by the observation, e.g. speed or temperature) and the `ssn:FeatureOfInterest` (the entity to which the `ssn:Property` applies, e.g. a vehicle or location). In implementing the passenger information scenario, described earlier, we have extended the SSN ontology to enable us to capture more contextual information. As observations are transmitted to a server from a mobile phone we create the `_:serverTime` property to describe when observations are received by the server. The GPS observations we are working with detail latitude and longitude and so we capture these using the W3C Semantic Web Interest Group's Basic Geo Vocabulary[6] `geo:lat` and `geo:long`. We can also represent the error associated with the observation using `_:accuracy`, along with the vehicle's `_:speed` and `_:heading`. This extra metadata enables our framework to perform a more comprehensive assessment of quality, as described later.

We characterise the quality assessment process using Furber and Hepp's Data Quality Management (DQM) ontology (Figure 2). This ontology enables the definition of `dqm:DataRequirements` that specify how quality assessment should be performed (i.e. *quality metrics*). A number of basic quality rules are built into the model (e.g. legal and illegal values, and unique values). However, these are not capable of describing application-specific data requirements such as calculating

[4] http://www.arduino.cc

[5] http://www.json.org

[6] http://www.w3.org/2003/01/geo/

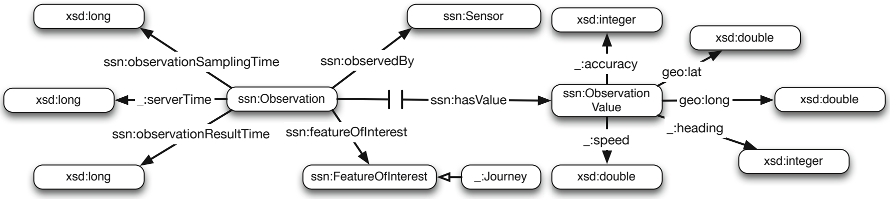

Fig. 1. An example sensor observation characterised using the SSN ontology

the distance between a GPS observation and a bus route. We have constructed a number of these requirements using the SPIN - SPARQL inferencing notation[7] which allow custom rules to be associated with `dqm:DataRequirement` instances (see example in Figure 2).

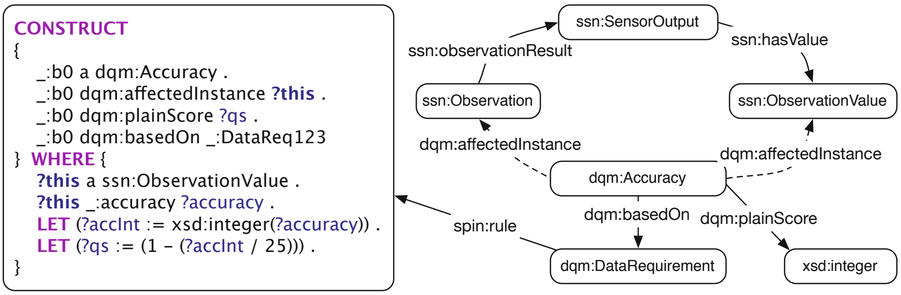
```
CONSTRUCT
{
   _:b0 a dqm:Accuracy .
   _:b0 dqm:affectedInstance ?this .
   _:b0 dqm:plainScore ?qs .
   _:b0 dqm:basedOn _:DataReq123
}  WHERE {
   ?this a ssn:ObservationValue .
   ?this _:accuracy ?accuracy .
   LET (?accInt := xsd:integer(?accuracy)) .
   LET (?qs := (1 - (?accInt / 25))) .
}
```

Fig. 2. Quality assessment characterised using the DQM ontology

We have also implemented a web based client that displays these sensor observations on a map based on the values of the `geo:lat` and `geo:long` properties. Clicking on these observations sends the observation's URI to the quality assessment service. This service employs a SPIN reasoner, guided by a number of rules, to evaluate the quality of selected observations which are returned to the web browser and displayed to the user. Figure 2 contains an example SPIN rule from the passenger information scenario that evaluates the accuracy of GPS observations, those with a low error are assigned a high quality score by the reasoner, which also annotates observations with the quality assessment results. Other examples include `timeliness` (observations older than 10 minutes are considered low quality), and `relevance` (observations farther than 250 metres from the expected route of travel are low quality). When assessment is complete instances of `dqm:QualityScore` are used to annotate the corresponding `ssn:Observation` or `ssn:ObservationValue` via the `dqm:affectedInstance` property.

[7] `http://spinrdf.org/`

4 Future Plans

At present our framework only examines the metadata describing the context surrounding sensor observations. We have already concluded that the SSN ontology is not sufficient to capture the provenance of sensor observations. We have identified the Prov-O[8] ontology as suitable for this task as it introduces a minimal set of concepts to represent provenance information in different application domains. Moreover, Prov-O conforms with the OWL 2 RL profile (scalable reasoning) which should facilitate the integration of provenance reasoning within our existing rule-based engine. This ontology is still being developed therefore we need to determine if Prov-O is capable of expressing the provenance of sensor observations. *Can the SSN and Prov-O ontologies be combined to represent the provenance of sensor observations?* For example, SSN can represent both the sensing process and the device that created an observation but with the inclusion of Prov-O we can also represent the agent that controlled the sensing process. *Can an SSN sensing process also be characterised as a Prov-O `Activity`?* We believe that this should be possible but need to investigate whether the semantics in both ontologies will permit this. The outcome of this investigation could be useful to the group developing Prov-O.

Another important question we need to address is: *Should the provenance of sensor observations be captured as they are created?* or *Should provenance be inferred only when a specific observation is requested?* Capturing the provenance of each observation could lead to the generation of large amounts of provenance data. Inferring provenance would avoid having to store much of this data but could increase the time taken to reason about its quality as the reasoner must perform two tasks (inferring provenance and performing quality assessment). We are also interested in answering the following question: *How can we use the provenance of existing quality scores to determine if these results can be reused?* This will involve either capturing or inferring the provenance of quality scores and authoring a number of new data requirements that can consider this provenance when performing new assessments. This raises the following issue: *Are DQM and Prov-O sufficient to characterise the provenance of quality scores?* For example, `dqm:DataRequirements` and `dqm:QualityScore`s could both be characterised as a `prov:Entity` and so a combination of the two ontologies could potentially provide a complete account of quality score provenance.

We also have a number of questions relating to how reasoning is performed within our quality assessment framework. *What kind of rules (based on the Prov-O / SSN / DQM combination) can be used to support quality assessment?* We have already identified a number of ways in which the provenance of sensor observations can be used to support quality assessment. For example, we can examine the reputation of the agent associated with the sensing process, the type of device that created the observation, and how the observation has been transformed since it was created (e.g. converting location observations between certain co-ordinate systems can reduce the accuracy of observations). We have

[8] `http://www.w3.org/2011/prov/`

also identified a number of scenarios in which agents could re-use quality scores, e.g. Agent A could re-use Agent B's quality result because they are in the same social network and trust each other, or because Agent B's data requirement matches one of Agent A's. We will continue to identify more scenarios that will, in turn, inform new data requirements.

Our hypothesis, in section 1, states that publishing semantic representations of data and their provenance provide additional context that enhances quality assessment. We will measure the extent to which the provision of additional contextual information is useful to quality assessment by documenting the number of quality dimensions that can be evaluated with and without this metadata. For example, a description of an observed value associated with a timestamp can only be evaluated for `timeliness`. However, adding a description of the observation's associated error enables assessments of `accuracy`, and a description of the *feature of interest* allows the assessment of `relevance`. Furthermore, increasing the amount of contextual information enables quality assessment to consider more metadata while evaluating each quality dimension. For example, as part of the earlier `accuracy` example we could also explore observation provenance to identify where `accuracy` may be reduced (such as a change in coordinate system). To evaluate this, we will analyse the number of RDF triples used in assessing each quality dimension. Capturing the provenance of past quality assessments should enable us to improve the performance of future quality assessments through the re-use of existing quality results. We will determine if this is true by analysing the time taken to perform a new quality assessment with or without the provenance of past assessments. The data required by these evaluations will be collected by deploying our solution as part of a larger software infrastructure to address issues in the passenger information scenario[9] outlined earlier. This will enable us to evaluate how our solution performs with real data and real users. We aim to show that the use of our quality assessment framework enables a service to better select data for presentation to its users based on a number of quality rules. For example, the service in the passenger information scenario can evaluate quality to ensure that the sensor observations produced by GPS devices on public transport vehicles are accurate, timely, and relevant to the user.

Our approach will be deemed to be successful if we can demonstrate that it is possible to assess the quality of sensor observations by examining metadata describing their characteristics and provenance. A further indicator of success will be if the deployment of our quality assessment framework within the passenger information scenario can be shown to provide tangible benefits to users.

Acknowledgement. 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.

[9] `http://www.dotrural.ac.uk/irp/`

References

1. Bermudez, L., Graybeal, J., Arko, R.: A marine platforms ontology: Experiences and lessons. In: Proceedings of the Semantic Sensor Networks Workshop at the 5th International Semantic Web Conference (November 2006)
2. Bizer, C., Cygniak, R.: Quality-driven information filtering using the wiqa policy framework. Journal of Web Semantics 7, 1–10 (2009)
3. Bizer, C., Heath, T., Berners-Lee, T.: Linked data - the story so far. International Journal on Semantic Web and Information Systems 5, 1–22 (2009)
4. Botts, M., Percivall, G., Reed, C., Davidson, J.: Ogc sensor web enablement: Overview and high level architecture. In: Nittel, S., Labrinidis, A., Stefanidis, A. (eds.) Proceedings of Geosensor Networks 2006, pp. 175–190 (2008)
5. Furber, C., Hepp, M.: Swiqa - a semantic web information quality assessment framework. In: European Conference on Information Systems, p. 76 (2011)
6. Hartig, O., Zhao, J.: Using web data provenance for quality assessment. In: 1st Int. Workshop on the Role of Semantic Web in Provenance Management, vol. 526 (2009)
7. Kim, J., Kwon, H., Kim, D., Kwak, H., Lee, S.: Building a Service-Orient Ontology for Wireless Sensor Networks. In: 7th IEEE/ACIS International Conference on Computer and Information Science, pp. 649–654 (2008)
8. Le-Phouc, D., Hauswirth, M.: Linked open data in sensor data mashups. In: International Workshop on Semantic Sensor Networks 2009, vol. 552, pp. 1–16. CEUR (2009)
9. Lee, Y.W., Strong, D.M., Kahn, B.K., Wang, R.Y.: Aimq: a methodology for information quality assessmemt. Information and Management 40, 133–146 (2002)
10. 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)
11. Page, K.R., Roure, D.C.D., 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)
12. Patel-Schneider, P.F., Fensel, D.: Layering the Semantic Web: Problems and Directions. In: Horrocks, I., Hendler, J. (eds.) ISWC 2002. LNCS, vol. 2342, pp. 16–29. Springer, Heidelberg (2002)
13. Richardson, M., Agrawal, R., Domingos, P.: Trust Management for the Semantic Web. In: Fensel, D., Sycara, K., Mylopoulos, J. (eds.) ISWC 2003. LNCS, vol. 2870, pp. 351–368. Springer, Heidelberg (2003)