Short Paper: Assessing the Quality of Semantic Sensor Data

Chris Baillie, Peter Edwards, Edoardo Pignotti, and David Corsar

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

Abstract. Sensors are increasingly publishing observations to the Web of Linked Data. However, assessing the quality of such data remains a major challenge for agents (human and machine). This paper describes how Qual-O, a vocabulary for describing quality assessment, can be used to perform quality assessment on semantic sensor data.

Keywords: Semantic Web, Linked Data, ontology, quality, provenance

1 Introduction

The number of sensors publishing data to the Web has increased dramatically, a trend that is set to accelerate further with the growth of the Internet of Things. However, in order to identify reliable datasets agents must first perform data quality assessment [1]. Data quality is defined in terms of 'fitness for use' and is assessed against one or more *dimensions of quality*, such as *timeliness* and *accuracy* using a set of *quality metrics* [8]. Such metrics typically produce values between 0 (indicating low quality) and 1 (high quality) by examining the context around data [2]. Some sensors now publish their data to the Web of Linked Sensor Data using vocabularies such as the Semantic Sensor Network (SSN) ontology [3] to describe the context around observations, such as the time observations were generated and the feature the measured phenomenon (e.g. temperature) affects (e.g. a geographical region). In this paper we will demonstrate how a generic quality assessment framework can use this context to facilitate data quality assessment. Moreover, we will argue that this context should be further enriched to include a description of sensor data provenance: a record of the agents, entities, and activities involved in data derivation. Capturing such metadata using a vocabulary such as the W3C PROV[1] recommendation enables users to better understand, trust, reproduce, and validate the data available on the Web [6]. Any quality assessment should thus examine data provenance as there are a number of provenance dimensions that can affect the quality of data. For example, derivation: *Was the observation derived from any other observations?* and attribution: *Who was associated with the generation of this observation?*.

[1] http://www.w3.org/TR/prov-overview/

2 Qual-O: An Ontology for Quality Assessment

In this section, we present the Qual-O ontology[2]. This model enables the specification of quality assessments that can examine the semantic representation of data (e.g. using SSN) and their provenance (e.g. using PROV). Moreover, this model can be used to describe the provenance of such assessments. For the remainder of this paper, we refer to the provenance used in assessment as *subject provenance* and the provenance of past assessments as *QA provenance*.

The concepts in Figure 1 preceded by the *qual* namespace characterise a minimal set of concepts for describing quality assessment, influenced by existing quality vocabularies such as the DQM ontology [5]. We adopt a different approach to these vocabularies insofar as we base our model on the PROV vocabulary to ensure it is capable of describing QA provenance. PROV is defined in terms of three main concepts: *Entity* (a physical, digital, conceptual, or other kind of thing with some fixed aspects), *Activity* (something that occurs over a period of time and acts upon or with entities), and *Agent* (something that bears some form of responsibility for an activity). However, *Entity* alone is insufficient to characterise metrics and subjects as they are simply *used* by an *Activity* to generate a further entity. Therefore, we use subclasses of *prov:Role* to describe the function of an entity with respect to an activity, e.g. *subject*, *metric*, and *result*. A *qual:Subject* is thus an *Entity* with a *subject* role, a *qual:Metric* is an *Entity* with a *metric* role, and a *qual:Result* is an *Entity* with *result* role. It then follows that a *qual:Assessment* is a kind of *Activity* that used one *qual:Metric*, one *qual:Subject*, and generated one *qual:Result*. The relations between each *qual* concept can also be defined in terms of PROV: *targets* and *guidedBy* are sub-properties of *prov:used* and describe the relationship between an Assessment and a Subject and Metric, respectively; *resultOf* is a sub-property of *wasGeneratedBy* and attributes a Result to its Assessment.

Defining a quality assessment model in this way has a number of advantages. Firstly, using OWL2 RL[3] enables the provenance of quality assessment to be inferred based on the concepts used to define the assessment. Secondly, the assessment activity is described including the time the assessment started and ended, and the kind of platform that performed QA (e.g. some computational service). Thirdly, we can attribute assessment results to the agent associated with QA to form descriptions of *why* the assessment was performed using agent intent [7] characterised as a set of goals and constraints (*int* namespace in Figure 1). For example, an agent can *only use sensor data with a score of 0.75* to achieve a goal, *decide whether to take a jacket with them on a walk.* Finally, we can associate QA results with specific subgraphs in the subject; for example, to describe the precise property and value in either the subject description or its provenance record. This could facilitate more complex QA re-use queries, e.g. '*select completed assessments with an accuracy score greater than 0.75 affecting location data generated by a GPS device and not mobile mast location*'.

[2] http://sensornet.abdn.ac.uk/onts/Qual-O.ttl

[3] http://www.w3.org/TR/owl2-profiles/

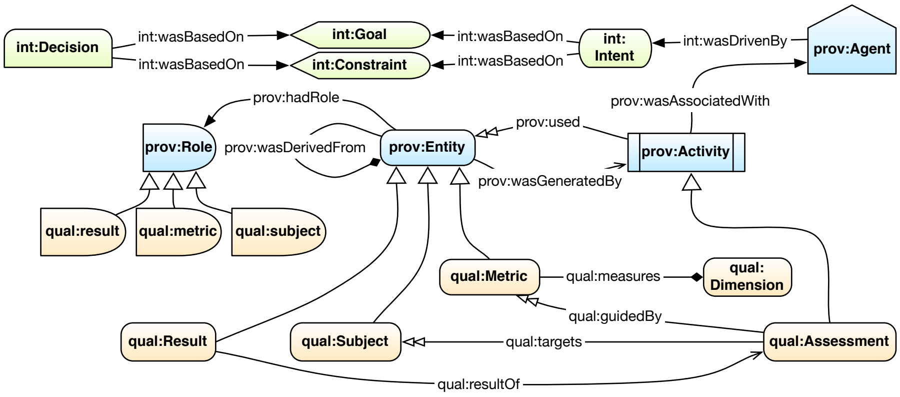


Fig. 1. An overview of the Qual-O ontology.

3 Assessing the Quality of Semantic Sensor Data

To investigate how our quality model performs it was first necessary to obtain some sensor data, described using a suitable ontology. To this end, we have developed a number of sensing devices based on the Arduino electronics prototyping platform. Each is equipped with sensors capable of describing, for example, temperature, humidity, location and acceleration (via GPS). Figure 2 provides an example of how we describe sensing devices and their observations using the SSN ontology. The Arduino is an instance of *Platform* with a number of attached *SensingDevice*s. These devices produce *ssn:Observation*s describing a specific *Property* (e.g. Temperature). The observed real-world phenomenon (e.g. Edinburgh, UK as described by DBPedia[4]) is captured using *FeatureOfInterest* and *ObservationValue* is used to describe the measured value.

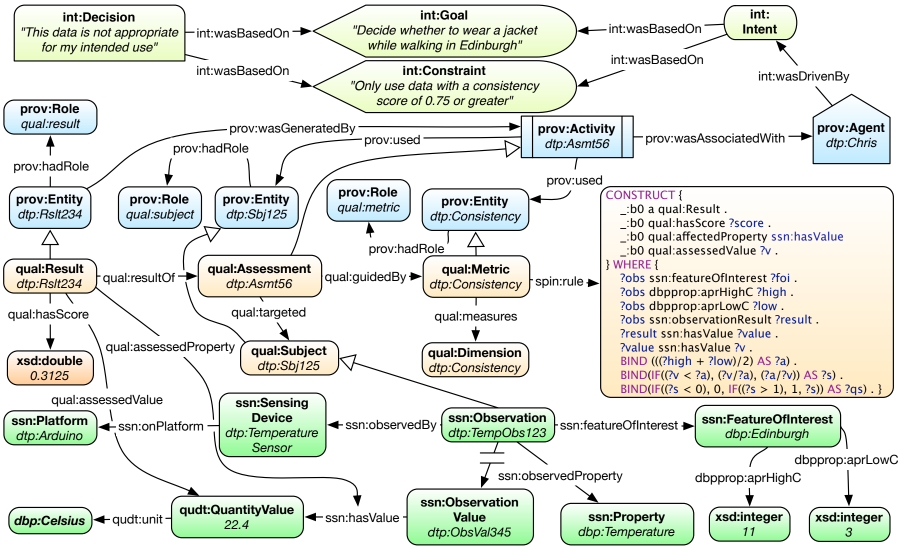
Using this framework, we have produced a number of datasets containing sensor observations describing the environmental conditions in a number of use cases in April 2013. *CarCommuter* (D1) describes a car journey between Ellon and Aberdeen, UK; *CityWalk* (D2) describes a pedestrian journey in Edinburgh, UK; *TrainJourney* (D3) describes a train journey between Aberdeen and Edinburgh; *CoastalWalk* (D4) describes a recreational walk on a beach near Aberdeen; and *WeatherStation* (D5) describes a 24 hour period in a fixed location near Aberdeen. We have developed a web service[5] that enables the visualisation of each dataset and provides the URL of each dataset's SPARQL endpoint. Clicking on individual observations within a visualisation triggers assessment of the selected observation as a *Subject*. The example in Figure 2 extends the logic of a *Metric* using SPIN[6] to calculate the average temperatures for the feature of interest

[4] http://www.dbpedia.org

[5] http://sensornet.abdn.ac.uk:8080/SensorBox

[6] http://www.spinrdf.org

at the time the observation was produced. In this example, the observation describes the temperature of Edinburgh, UK where the average of the high and low temperature in April, according to DBPedia, is 7°C. The metric then states that observation quality, in terms of *Consistency*, decreases the further its value is from 7°C. In this example, the observation has a value of 22.4°C and therefore is annotated with a quality score of 0.3125 (indicating a low quality observation). It should be noted that this represents only one method of computing a consistency score for this observation; other agents may have others depending on their intended use for the data.



```
CONSTRUCT {
    _:b0 a qual:Result .
    _:b0 qual:hasScore ?score .
    _:b0 qual:affectedProperty ssn:hasValue
    _:b0 qual:assessedValue ?v .
} WHERE {
    ?obs ssn:featureOfInterest ?foi .
    ?obs dbpprop:aprHighC ?high .
    ?obs dbpprop:aprLowC ?low .
    ?obs ssn:observationResult ?result .
    ?result ssn:hasValue ?value .
    ?value ssn:hasValue ?v .
    BIND (((?high + ?low)/2) AS ?a) .
    BIND(IF((?v < ?a), (?v/?a), (?a/?v)) AS ?s) .
    BIND(IF((?s < 0), 0, IF((?s > 1), 1, ?s)) AS ?qs) . }
```

Fig. 2. Using Qual-O to assess the quality of SSN sensor observations.

As noted earlier in the paper, documenting QA provenance can enable agents to better understand the outcomes of quality assessment. Reasoning about quality using Qual-O allows a reasoner to infer QA provenance as an assessment is performed. Such a provenance record can link a *Subject Entity* with the *Metric Entity* used to assess it via the *Assessment Activity*. Furthermore, an *Agent*'s intent can also be captured (boxes with *int* namespace in Figure 2). In this example, the agent has a constraint that they can *only use data with a minimum consistency score of 0.75*. Quality assessment has produced a consistency score of 0.3125 for this example observation and so it is reasonable to conclude that this agent will not use this observation to achieve its goal, *decide whether to wear a jacket while walking in Edinburgh*.

4 Evaluation

To investigate the performance of our quality assessment framework, we conducted a series of experiments using a Sun Fire X4100 M2 with two dual-core AMD Opteron 2218 CPUs and 32GB of RAM running CentOS 5.8, Java SE1.6, JENA 2.10 and SPIN 1.3. Each experiment used a set of quality metrics (Table 1) to assess 600 observations across datasets 2, 4 and 5. For each metric we created two SPIN rules: one that examines the context around observations described using only SSN; the other uses subject provenance to assess quality based on the observations the subject was derived from. Experiment 1 measured the reasoning time required to apply quality metrics, expressed using Qual-O, to individual observations. Figure 3 shows that using metric set 2 resulted in a significant increase in reasoning time. This is only to be expected as the amount of metadata to examine has increased. However, set 2 was able to identify quality issues that set 1 could not as these issues were only present in the provenance record, e.g. that an observation was derived from another, low quality, observation. Experiment 2 compared the overhead of performing quality assessment (M1) with the overhead of performing quality assessment and capturing its provenance (M2). These results (Figure 4) demonstrate an increase in the reasoning time required to document quality assessment provenance due to the reasoner having to infer more triples during the assessment.

Table 1. The set of quality metrics used to evaluate Qual-O.

Dimension	**Description**
Consistency	Average April temperature[7] for Aberdeen and Edinburgh is around 7°C.
Consistency	Average humidity[8] for Aberdeen and Edinburgh is around 79%.
Believability	GPS observations should have been produced using at least 3 satellites.
Accuracy	GPS observations should have an error margin less than 100 metres.
Believability	There are no areas of Scotland below sea level and dry.
Accuracy	At least 4 GPS satellites are required to measure altitude.

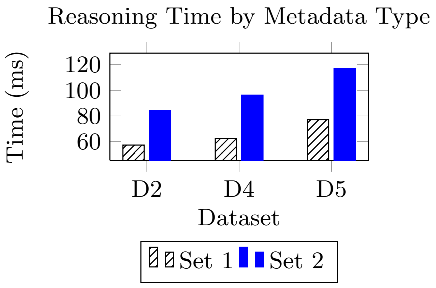


Fig. 3. Average time required by the SSN & PROV metrics.

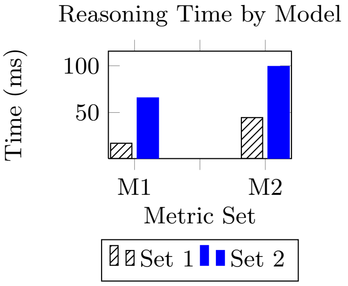


Fig. 4. Average reasoning time to perform QA and infer QA provenance.

[7] http://dbpedia.org/page/Aberdeen and http://dbpedia.org/page/Edinburgh

[8] http://www.currentresults.com/Weather/UnitedKingdom/humidity-annual.php

5 Conclusions & Future Work

Through the experimental results presented here and the experience gained deploying our framework in a real-world application [4], we have been able to demonstrate that quality assessment can be performed on sensor observations by examining observations (described using SSN) and their provenance (described using PROV). This should enable agents to make decisions about which datasets (sensor or otherwise) are reliable based on the quality results produced by our framework. Moreover, we have shown that inclusion of observation provenance during quality assessment causes an increase in required reasoning time. While this is a considerable increase, we argue that this is offset by the richer descriptions of assessment provenance that can be achieved using this method. Agents should be able to examine this provenance to better understand how assessments were performed and potentially make decisions about re-use of existing assessment results. Our future work will therefore investigate how existing quality results can be re-used to potentially reduce the overhead associated with QA. This work will include investigating how agents can use specific elements of QA provenance to make re-use decisions, e.g. if QA was performed by an agent that they trust. Further, we aim to strengthen our evaluation by repeating our experiments with different sets of metrics, different hardware specifications, and larger observation models.

Acknowledgements 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

References

1. Baillie, C., Edwards, P., Pignotti, E.: Quality reasoning in the semantic web. In: The Semantic Web - ISWC 2012. Lecture Notes in Computer Science, vol. 7650, pp. 383–390 (November 2012)
2. Bizer, C., Cygniak, R.: Quality-driven information filtering using the wiqa policy framework. Journal of Web Semantics 7, 1–10 (2009)
3. Compton, M., et al.: The SSN ontology of the W3C Semantic Sensor Network Incubator Group, vol. 17, pp. 25–32. Web Semantics: Science, Services and Agents on the World Wide Web (2012)
4. Corsar, D., Edwards, P., Baillie, C., Markovic, M., Papangelis, K., Nelson, J.: Short paper: Citizen sensing within a real time passenger information system. In: 6th International Workshop on Semantic Sensor Networks (2013)
5. 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)
6. 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 (June 2009)
7. Pignotti, E., Edwards, P., Gotts, N., Polhill, G.: Enhancing workflow with a semantic description of scientific intent. Journal of Web Semantics 9, 222–244 (2010)
8. Wang, R., Strong, D.: Beyond accuracy: what data quality means to data consumers. In: Journal of Management Information Systems. vol. 12, pp. 5–33 (1996)