Short Paper: Addressing the Challenges of Semantic Citizen-Sensing

David Corsar, Peter Edwards, Nagendra Velaga, John Nelson, and Jeff Pan

dot.rural Digital Economy Hub,
University of Aberdeen,
Aberdeen, UK.
dcorsar,p.edwards,n.r.velaga,j.d.nelson,jeff.z.pan}@abdn.ac.uk
http://www.dotrural.ac.uk

Abstract. The challenges of the sensor web have been well documented, and the use of appropriate semantic web technologies promises to offer potential solutions to some of these challenges (for example, how to represent sensor data, integrate it with other data sets, publish it, and reason with the data streams). To date a large amount of work in this area has focused on sensor networks based on "traditional" hardware sensors. In recent years, citizen sensing has became a relatively well-established approach for incorporating humans as sensors within a system. Often facilitated via some mobile platform, citizen sensing may incorporate observational data generated by hardware (e.g. a GPS device) or directly by the human observer. Such human observations can easily be imperfect (e.g. erroneous or fake), and sensor properties that would typically be used to detect and reason about such data, such as measurements of accuracy and sampling rate do not exist. In this paper we discuss our work as part of the Informed Rural Passenger project, in which the passengers themselves are our main source for transport related sensing (such as vehicle occupancy levels, available facilities). We discuss the challenges of incorporating and using such observational data in a real world system, and describe how we are using semantic web technologies, combined with models of provenance to address them.

Keywords: Citizen-Sensing, Semantic Sensing, Semantic Citizen-Sensing, Provenance, Semantic Web

1 Introduction

The challenges of the sensor web have been well documented in, for example, [21], [22] and [8]. Documented challenges include: modeling, querying, and reasoning with large scale sensor data [8, 11, 17, 15]; identification of, and integration with other relevant data sets, at scale [8, 11, 18, 24, 7]; characterizing and managing sensor data quality [8]; and supporting rapid application development [8].

The use of semantic web technologies offer potential solutions to some of these challenges. Ontologies, such as the W3C SSN XG ontology[1] provide mod-

[1] http://www.w3.org/2005/Incubator/ssn/XGR-ssn

els for sensors, sensor networks, and observations; and linked data [5] enables integration of sensor data with other data sets [4, 13, 18]. Sensors typically produce streams of data, and so there is potential for using technologies such as RDF stream querying [6, 3] (as explored in [15]) and RDF stream reasoners (e.g. [2, 23]) to support the use of that data. Further, Application Programming Interfaces (APIs), such as the Linked Data API[2] offer support for rapid application development.

To date a large amount of work in this area has focused on sensor networks based on "traditional" hardware sensors. In recent years, citizen sensing [19] has became a relatively well-established approach for incorporating humans as sensors within a system. Often facilitated via applications (apps) running on a mobile phone, citizen sensing may generate observational data by hardware (e.g. a GPS device) or directly by the human observer. Such human observations can easily be imperfect (e.g. erroneous, incomplete, or fraudulent), and so, as with any open system, this raises issues such as information quality (IQ) [10], reliability, trust, and reputation [14].

One further challenge of citizen sensing, is that for observations generated directly by the human observer, sensor properties that would typically be used to detect and reason about imperfect data, for example measurements of accuracy and sampling rate, do not exist. Similar problems exist with observations generated by the mobile phone's hardware: often the mobile APIs provide few details such as data sheets (describing sensor capabilities), settings used for observations, and, in some cases, which sensor generated an observation[3].

This lack of information makes it difficult to perform the necessary assessments of observations produced using citizen sensing. Semantic web technologies potentially have a role to play here by, for example, providing additional contextual information for use in assessment processes.

In this paper we describe an example real-world system which combines citizen sensing with semantic web technologies (section 2); discuss some of the challenges faced by this system (section 3); and describe how we are addressing those challenges (section 4).

2 Example System

As part of the Informed Rural Passenger (IRP) project[4], we are investigating the challenges of developing a trusted passenger information system (PIS) for rural areas. In our system the passengers themselves are our main source of transport related sensing, performed using a mobile app. The app enables passengers to contribute observations about their journey on public transport, including observations generated directly by the phone (e.g. location, presence of Wi-Fi) and by the passenger (e.g. occupancy level, and perceived vehicle temperature).

[2] http://code.google.com/p/linked-data-api/

[3] For example, the Apple iPhone iOS's location API uses either the cellular network, Wi-Fi, or GPS sensor to determine location, but does not indicate which was used.

[4] http://www.dotrural.ac.uk/irp

Using linked data principles, this data is then integrated with other relevant data sets, and used as the basis of a PIS, which provides passengers with details, including real-time bus locations, delays, and expected arrival times. This therefore gives the potential for any imperfect data passed as input to the system to adversely effect its outputs, reducing user trust in the system.

3 Challenges of Semantic Citizen Sensing

In developing the IRP PIS, we have identified a set of issues, which extend those defined for the sensor web, and, we believe, require to be addressed by any system which incorporates humans as a source of sensor data, in order to remain trusted by its users. These challenges are raised due to the potential generation of imperfect data by humans, and the lack of information for identifying and reasoning about it.

Challenge 1, is one of the most pressing: the need to characterise and manage constructs not just of data quality, but also of, for example, reliability, reputation, and trustworthiness, which can use the available types of data.

This gives rise to **challenge 2**: maximising the data available for making those assessments. Here, identifying and integrating the sensor data with appropriate external data sets can help address this challenge. Related to this are: **challenge 3**, selecting an appropriate model for describing the citizen sensors and their observations, the possible granularity of which is limited by the lack of information about them; and **challenge 4**, integrating the qualitative observations generated by humans with the machine generated quantitive observations.

In real-time information systems, short response times are vital; however, processes such as data integration and data assessments potentially conflict with this requirement. Further, the additional data generated by these processes adds to the amount that must be stored and processed. This gives rise to **challenge 5**: designing a system architecture which uses an appropriate combination of technologies (e.g. for storing and reasoning about the data), which enable the system to perform well while maintaining an acceptable response time.

Finally, **challenge 6** relates to ensuring user privacy, especially when sensitive data such as location is being collected and used as the basis of information passed to other users and/or services. Addressing this challenge is made more difficult by the integration with other data sets, which potentially provide additional data which can be used to violate a user's privacy.

4 Addressing These Challenges

Within the IRP project we are addressing the above challenges by, firstly exploring the data available within the application domain, and secondly investigating how it can be integrated to form an information ecosystem supporting a range of services which perform PIS functions and data assessments. Whilst the solutions below are outlined within the context of IRP, we believe they are generalisable to other systems incorporating humans as sensors.

The information ecosystem that we are developing to address challenge 2 integrates the passenger observations with various other types of data, including user profiles, social networks, and various types of transport information, such as: operator timetables; NaPTAN and NPTG[5] datasets[6]; road maps; and details of roadworks and travel disruptions[7], which themselves link to NPTG. We will also reverse geocode location points obtained from passengers and, where possible, link them to the nearest road/railway line and nearest settlement in the NPTG.

Fig. 1 outlines how we are addressing challenges 3 and 4, by basing our model of citizen sensors and observations on the SSN XG ontology, with domain extensions describing: mobile phones as sensor platforms, people as sensors, and the observations generated via our mobile app (Trips represent timetabled journeys).

Within the ecosystem, we also maintain a provenance record of the data. Provenance has previously been identified as essential in supporting reliability, discovery, trust [20], and quality assessment [10] of online information, and so may play an important role in assessments of data in the ecosystem. Fig. 1 shows how we use the Open Provenance Model Vocabulary[8] (OPMV) encoding of OPM [12] to provide explicit provenance information about observations. OPM defines provenance using causal relationships between *Artifacts*, *Processes*, and *Agents*, which we use, for example, to link observations generated by a phone's hardware to the passenger controlling the sensing.

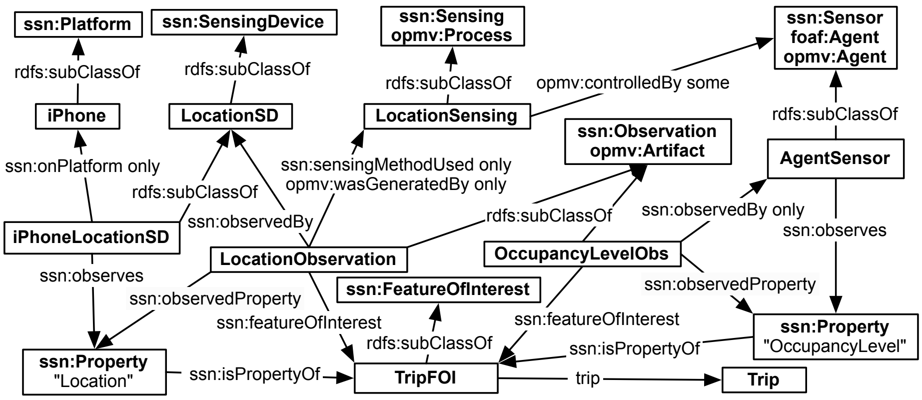

Fig. 1. SSN XG ontology extended for citizen sensing and provenance for IRP.

One reason for maintaining this information within the ecosystem is to support various types of data assessment, particularly of IQ and trustworthiness.

[5] http://data.gov.uk/linked-data

[6] NaPTAN provides details of all UK access points to public transport; NPTG provides details of all UK settlements and roads connected to the public transport network.

[7] Provided by http://trafficscotland.dataincubator.org/

[8] http://purl.org/net/opmv/ns-20101006

IQ assessments of data typically analyse various dimensions of the data, and so the additional information should be beneficial; for example, other members of our research team are currently investigating the role of provenance in IQ assessments of linked sensor data [1]. The multi-agent community have extensively studied models of trust and reputation [16, 14], which often rely on analysing past interactions between agents (i.e. analysing the provenance of interactions), while others combine trust, provenance and social networks [9]. As part of addressing challenge 1, we are currently investigating how these models can be applied within the ecosystem.

We will incorporate any data assessments and their results within the ecosystem as part of the provenance record (as subclasses of OPMV Process and Artifact classes respectively). This will allow services/applications (including those making new assessments) to make use of these assessments if appropriate.

The nature of the IRP project requires that it handles large quantities of data and still functions reliably in real time. To help support this and address challenge 5, passenger contributed observations are currently stored in a database, and exposed as linked data using the D2R server[9]. This setup takes advantage of the strengths of databases (such as scaling to large data sets, and handling multiple concurrent read, update, and delete operations). However, the disadvantage is that it does not exploit many of the advantages of semantic web technologies, such as the ontology based querying and reasoning.

5 Conclusions and Future Work

In this paper we have identified a set of challenges, which we believe, require to be addressed by any system that incorporates humans as a source of sensor data. We propose the use of semantic web technologies to help address these challenges, and illustrate their use in the development of a real-time PIS for rural areas.

We currently have three strands of future work addressing challenges 1, 5, and 6: developing a trust model for the ecosystem; evaluating the performance of different options for storing and reasoning about streaming linked sensor data, to determine if a combination can be found that provides (some of) the advantages of semantic web technologies without negatively impacting overall performance; and investigating how we can ensure user privacy.

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.: Assessing Quality in the Web of Linked Sensor Data. In: Proc. of AAAI-11 (2011)

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

2. Barbieri, D., Braga, D., Ceri, S., Della Valle, E., Grossniklaus, M.: Incremental Reasoning on Streams and Rich Background Knowledge. In: The Semantic Web: Research and Applications. vol. 6088, pp. 1–15. Springer Berlin Heidelberg, Berlin, Heidelberg (2010)
3. Barbieri, D.F., Braga, D., Ceri, S., Della Valle, E., Grossniklaus, M.: C-SPARQL: SPARQL for continuous querying. In: Proc. of the WWW'09. pp. 1061–1062. WWW '09, ACM, New York, NY, USA (2009)
4. Barnaghi, P., Presser, M.: Publishing Linked Sensor Data. In: Taylor et al. [22]
5. Berners-Lee, T.: Linked Data. IJSWIS 4(2), 1 (2006)
6. Bolles, A., Grawunder, M., Jacobi, J.: Streaming SPARQL - Extending SPARQL to Process Data Streams, Lecture Notes in Computer Science, vol. 5021, chap. 34, pp. 448–462. Springer Berlin Heidelberg, Berlin, Heidelberg (2008)
7. Compton, M., Neuhaus, H., Taylor, K., Tran, K.N.: Reasoning about Sensors and Compositions. In: Taylor et al. [21], pp. 33–48
8. Corcho, O., García-Castro, R.: Five challenges for the Semantic Sensor Web. Semantic Web Interoperability, Usability, Applicability 1(1), 121–125 (Jan 2010)
9. Golbeck, J.: Combining provenance with trust in social networks for semantic web content filtering. In: Proc. of IPAW 2006. vol. 4145, pp. 101–108. Springer (2006)
10. Hartig, O., Zhao, J.: Using Web Data Provenance for Quality Assessment. In: Proc. of Workshop on Semantic Web and Provenance Management at ISWC (2009)
11. Kessler, C., Janowicz, K.: Linking Sensor Data - Why, to What, and How? In: Taylor et al. [22]
12. Moreau, L., Clifford, B., Freire, J., Futrelle, J., Gil, Y., Groth, P., Kwasnikowska, N., Miles, S., Missier, P., Myers, J., Plale, B., Simmhan, Y., Stephan, E., den Bussche, J.V.: The open provenance model core specification (v1.1). Future Generation Computer Systems (July 2010)
13. Page, K., De Roure, D., Martinez, K., Sadler, J., Kit, O.: Linked Sensor Data: RESTfully serving RDF and GML. In: Taylor et al. [21], pp. 49–63
14. Ramchurn, S.D., Huynh, T.D., Jennings, N.R.: Trust in Multiagent Systems. The Knowledge Engineering Review 19(1), 1–25 (2004)
15. Rodríguez, A., McGrath, R., Liu, Y., Myers, J.: Semantic Management of Streaming Datas. In: Taylor et al. [21], pp. 135–147
16. Sabater, J., Sierra, C.: Review on computational trust and reputation models. Artificial Intelligence Review 24, 33–60 (2005)
17. Sabou, M., Kantorovitch, J., Nikolov, A., Tokmakoff, A., Zhou, X., Motta, E.: Position Paper on Realizing Smart Products: Challenges for Semantic Web Technologies. In: Taylor et al. [21], pp. 135–147
18. Sequeda, J., Corcho, O.: Linked Stream Data: A Position Paper. In: Taylor et al. [21], pp. 148–157
19. Sheth, A.: Citizen Sensing, Social Signals, and Enriching Human Experience. IEEE Internet Computing 13(4), 87–92 (2009)
20. Simmhan, Y.L., Plale, B., Gannon, D.: A survey of data provenance in e-science. ACM SIGMOD Record 34(3), 31–36 (2005)
21. Taylor, K., Ayyagari, A., De Roure, D. (eds.): Proceedings of the 2nd International Workshop on Semantic Sensor Networks (SSN09) (2009)
22. Taylor, K., Ayyagari, A., De Roure, D. (eds.): Proceedings of the 3rd International Workshop on Semantic Sensor Networks (SSN10) (2010)
23. Thomas, E., Pan, J.Z., Ren, Y.: TrOWL: Tractable OWL 2 Reasoning Infrastructure. In: the Proc. of ESWC2010 (2010)
24. Tran, K.N., Compton, M., Wu, J., Gor, R.: Short Paper: Semantic Sensor Composition. In: Taylor et al. [22]