DEMO: A Rural Passenger Information System Utilising Linked Data & The Crowd

David Corsar, Chris Baillie, Milan Markovic, Peter Edwards, John Nelson, Nagendra Velaga, Mark Beecroft, Somayajulu Sripada, Jeff Z. Pan & Konstantinos Papangelis
dot.rural Digital Economy Hub
University of Aberdeen, Aberdeen, UK
{dcorsar, c.baillie, m.markovic, p.edwards, j.d.nelson, n.r.velaga, m.beecroft, yaji.sripada, jeff.z.pan, k.papangelis}@abdn.ac.uk

ABSTRACT

This paper describes a real-time passenger information system based on crowdsourcing and linked data.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; J.7 [**Computers in other systems**]: Real time

General Terms

Design, Experimentation, Performance

Keywords

provenance, information quality, crowdsourcing, linked data, semantic infrastructure, transport

1. INTRODUCTION

Real-time passenger information (RTPI) systems use a range of information sources (including real-time vehicle locations, timetables, and details of disruptions), to provide passengers with information such as estimated vehicle arrival times and notification of delays or cancellations. Although common in urban areas, few systems exist in non-urban areas, for a variety of reasons, including the lack of infrastructure for obtaining and providing real-time information [8]. In the Informed Rural Passenger Project[1] (IRP), we are developing *GetThere*, a RTPI system for rural areas which uses data from multiple sources (including users) and delivers information via multiple channels (smartphone app, SMS messages, web sites); this system is being initially deployed in the Scottish Borders in partnership with First Group.

2. ARCHITECTURE

At the core of our system is a semantic information infrastructure built around the principles of linked data [4]. Figure 1 provides an overview of the various data and software components: the domain data layer provides a distributed domain knowledge base; the annotation services layer features services that reason about the domain data layer for specific purposes, with any annotations they produce being stored in the annotation data layer; the application services layer consists of services that use the other components to provide application specific functionalities (e.g. finding travel disruptions affecting a public transport service).

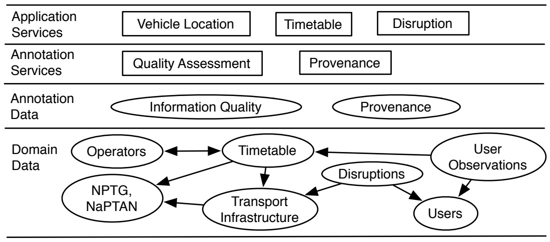


Figure 1: Real-time passenger information ecosystem.

The domain data layer uses linked data to integrate relevant datasets, each provided/hosted by different parties, and includes both open data and repositories only accessible within the ecosystem. The travel/transport datasets currently used include the UK Government's NPTG and NaPTAN datasets[2], which provide names and geolocations for settlements served by public transport, and details of all public transport access points (e.g. bus stops, train stations); a dataset describing public transport operators based on Traveline's National Operators Code (NOC) Excel spreadsheet; and a timetable dataset providing details of public transport routes and schedules, provided by operators and defined using the transit vocabulary[3].

The transport infrastructure dataset details the road network, based on maps from openstreetmap.org. The disruptions dataset describes crowdsourced reports of travel disruption (e.g. road closures), and is modelled using our own travel disruption ontology. The user observations dataset uses the W3C Semantic Sensor Network ontology[4] to describe observations (e.g. of vehicle occupancy level, vehicle location) crowdsourced from passengers using the *GetThere* app. The user dataset contains details of user accounts and

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

[2] Accessed via the SPARQL endpoint at http://gov.tso.co.uk/transport/sparql

[3] http://vocab.org/transit

[4] http://www.w3.org/2005/Incubator/ssn/ssnx/ssn

profiles, modelled using the SIOC[5] and FOAF[6] ontologies, along with descriptions of their journeys made while using the *GetThere* app. These datasets are accessed via a Fuseki SPARQL endpoint, and stored in a TDB store, running on an Amazon EC2 instance.

Given the open nature of the domain data, issues such as information quality, trust, and reputation naturally arise [7]. Examples range from malicious users and inaccurate devices to out-of-date information (e.g. timetables). Our ecosystem aims to mitigate these issues through a number of annotation services that can reason about the information stored within the domain data layer.

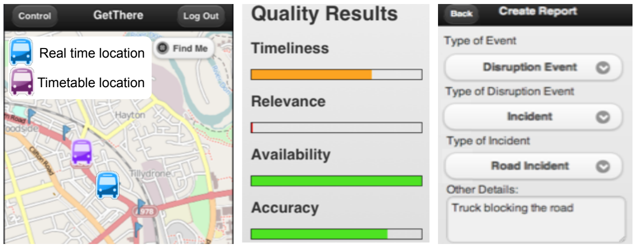
Evaluating the quality of data published on the Web has been identified as essential if agents (people or software) are to identify reliable information [1]. Therefore, as part of our software ecosystem, we have developed a service that can evaluate the quality of data within the ecosystem. This service and its application within IRP are discussed in detail in [1, 2]; briefly, this service employs a SPIN reasoner guided by a number of SPARQL rules [6] to examine the metadata associated with, for example, user observations. These rules describe a number of *quality metrics* that define how data should be evaluated against a number of *quality dimensions* [3]. Our system currently evaluates location observations from users against four quality dimensions including **accuracy** (accurate location observations have an associated error less than 25 metres) and **relevance** (relevant observations are no farther than 250 metres from the expected route of travel). At present, the results of quality assessment are displayed as a set of colour-coded bar graphs (see Figure 2).

Provenance provides a record of the entities, activities, and people involved in producing a piece of data, which can be used to perform assessments about its quality, reliability, or trustworthiness[7]. Given these potential uses, the ecosystem features a provenance annotation service that provides functionalities for generating and accessing the provenance of data within the ecosystem. This includes inferring provenance for user observations, and maintaining provenance records for data hosted by third parties [5], which can be used by services to support data assessment.

3. DISCUSSION

The *GetThere* system uses mobile devices to crowdsource information from, and provide information to, users. We are deploying the system in the Scottish Borders, a largely rural area where mobile signal coverage is unreliable; the ecosystem is therefore designed to be flexible in terms of how it acquires and transmits information (e.g. via the smartphone app or SMS messaging as appropriate).

We use open datasets to provide domain data where possible, which has presented us with various challenges. These include: information that may be out-of-date (e.g. the NPTG/ NaPTAN linked dataset is from March 2010); and integrating data published in various heterogeneous formats (e.g. webpages, spreadsheets, and linked data). The annotation layers are designed to address the former issue; to address the latter, we chose to use linked data throughout the ecosystem. This has required us to produce linked data versions of several datasets, such as Traveline's NOC and openstreetmap.org's road network. However, we argue the benefits linked data provides in terms of simplifying integration (through the use of dereferencable URIs) and maintenance (updates just have to be made in one location) outweigh the effort required to produce the linked data versions.



Type of Event: Disruption Event

Type of Disruption Event: Incident

Type of Incident: Road Incident

Other Details: Truck blocking the road

Figure 2: Screenshots of the *GetThere* smartphone app showing (left to right): vehicle locations; the results of invoking the quality assessment service; and creating a disruption report.

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] C. Baillie, P. Edwards, and E. Pignotti. Quality reasoning in the semantic web. In *Proc. of The 11th International Semantic Web Conference*. Springer-Verlag, to appear.

[2] C. Baillie, P. Edwards, and E. Pignotti. A role for provenance in quality assessment. In *Digital Futures 2012 The Third Digital Futures Conference*, to appear.

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

[4] C. Bizer, T. Heath, and T. Berners-Lee. Linked data - the story so far. *International Journal on Semantic Web and Information Systems*, 5:1–22, March 2009.

[5] D. Corsar, P. Edwards, N. R. Velaga, J. D. Nelson, and J. Z. Pan. Exploring provenance in a linked data ecosystem. In *Proc. of the 4th International Provenance and Annotation Workshop, IPAW 2012*, June 2012.

[6] C. Furber and M. Hepp. Swiqa - a semantic web information quality assessment framework. In *19th European Conference on Information Systems*, pages 922–933, 2011.

[7] S. D. Ramchurn, T. D. Huynh, and N. R. Jennings. Trust in multiagent systems. *The Knowledge Engineering Review*, 19(1):1–25, 2004.

[8] N. R. Velaga, M. Beecroft, J. D. Nelson, D. Corsar, and P. Edwards. Transport poverty meets the digital divide: accessibility and connectivity in rural communities. *Journal of Transport Geography*, 21(0):102 – 112, 2012.

[5] http://rdfs.org/sioc/spec/

[6] http://xmlns.com/foaf/spec/

[7] http://www.w3.org/TR/prov-dm/