Accepted Manuscript

Linking open data and the crowd for real-time passenger information

David Corsar, Peter Edwards, John Nelson, Chris Baillie,
Konstantinos Papangelis, Nagendra Velaga

PII: S1570-8268(17)30013-6
DOI: http://dx.doi.org/10.1016/j.websem.2017.02.002
Reference: WEBSEM 429

To appear in: *Web Semantics: Science, Services and Agents on the World Wide Web*

Received date : 26 July 2016
Revised date : 13 January 2017
Accepted date : 21 February 2017

Please cite this article as: D. Corsar, et al., Linking open data and the crowd for real-time passenger information, *Web Semantics: Science, Services and Agents on the World Wide Web* (2017), http://dx.doi.org/10.1016/j.websem.2017.02.002

Linking Open Data and the Crowd for Real-Time Passenger Information

David Corsar[a], Peter Edwards[a], John Nelson[b], Chris Baillie[a], Konstantinos Papangelis[c], Nagendra Velaga[d]

[a] *Computing Science, University of Aberdeen, Aberdeen, UK*
[b] *Centre for Transport Research, University of Aberdeen, Aberdeen, UK*
[c] *Computer Science, Xi'an Jiaotong-Liverpool University, Xian, China*
[d] *Department of Civil Engineering, Indian Institute of Technology Bombay, Mumbai, India*

Abstract

The availability of real-time passenger information (RTPI) is a key factor in making public transport both accessible and attractive to users. Unfortunately, rural areas often lack the infrastructure necessary to provide such information, and the cost of deploying and maintaining the required technologies outside of urban areas is seen as prohibitive. In this paper we present the *GetThere* system developed to overcome such issues and to provide public transport users in rural areas with RTPI. An ontological framework for representing mobility information is described, along with the Linked Data approach used to integrate heterogeneous data from multiple sources including government, transport operators, and the public. To mitigate possible issues with the veracity of this data, a quality assessment framework was developed that utilises data provenance. We also discuss our experiences working with Semantic Web technologies in this domain, and present results from both a user trial and a performance evaluation of the system.

Keywords: Semantic Web, ontology, quality, provenance, transport, citizen-sensing

1. Introduction

Real-time passenger information (RTPI) systems play a key role in the perceived quality of public transport services, influencing their attractiveness and accessibility [1, 2]. By providing passengers with real-time vehicle locations and estimates of arrival and departure times, RTPI systems allow people to plan and make decisions regarding their journeys. To achieve this, such systems must integrate heterogenous information such as service timetables, details of routes, and GPS-based vehicle locations, all of which may be provided by different agencies [3].

This paper reports on the work of the Informed Rural Passenger[1] project which investigated the suitability of a Semantic Web approach to data integration and use [4, 5] for smart mobility applications. RTPI systems are rare in rural areas, due to a lack of supporting infrastructure and the cost of maintaining technologies such as vehicle tracking in rural environments [6]. To overcome these issues we developed a mobility information ecosystem that used Linked Data to integrate open transport data with data received via citizen sensing [7]. The latter refers to use of humans as data providers using web-connected mobile devices [7]. In our work this involved asking bus passengers in rural areas to use the *GetThere* smartphone app to share the location of buses on which they were travelling. The app could also be used to access information about bus services stored in the ecosystem.

Insights into the information requirements of rural public transport users were obtained through a series of interviews and structured focus groups conducted in the Scottish Borders area of the UK [8]. Subsequent analysis of the interview and focus group transcripts, and a review of existing RTPI systems identified the following desirable capabilities for the *GetThere* smartphone app: C1) list available public transport services; C2) provide timetable (schedule) information for those services;

Email addresses: dcorsar@abdn.ac.uk (David Corsar), p.edwards@abdn.ac.uk (Peter Edwards), j.d.nelson@abdn.ac.uk (John Nelson), c.baillie@abdn.ac.uk (Chris Baillie), k.papangelis@xjtlu.edu.cn (Konstantinos Papangelis), n.r.velaga@iitb.ac.uk (Nagendra Velaga)

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

C3) provide (real-time) vehicle locations; and C4) provide individuals with information that is timely, accurate, and personalised to them, particularly during disruptions to the transport network.

Based on these requirements we defined the following specification for the computational infrastructure to support the *GetThere* app: I1) model public transport services and timetables, the transport infrastructure (e.g. roads, public transport access points), and vehicle locations; I2) use these models to integrate heterogeneous mobility data from multiple providers, including, where possible, open data sources to reduce data acquisition costs; I3) record and use data provenance to reason about quality issues arising from the use of data from external providers [9, 10]; and I4) use the integrated data to provide the desired RTPI capabilities.

This paper describes the semantic infrastructure developed to fulfil these requirements and its application in user trials. Semantic Web technologies are required to support the integration of heterogenous mobility data that is annotated with metadata providing schema and provenance information. Our contributions are: an ontological framework for representing mobility information suitable for multiple transport domains and geographic areas; use of W3C recommended and best practice ontologies in a system available to the general public; an approach for integrating highly dynamic data from citizen sensing with other static data sources; experience of reusing linked open data in a deployed system; a framework capable of assessing the quality of transport data; and an evaluation considering the effects of the information provided by the system on a sample population.

The remainder of this paper is structured as follows: Section 2 discusses related work; Section 3 describes the mobility information ecosystem; Section 4 describes the *GetThere* smartphone app; Section 5 discusses a system performance evaluation; Section 6 discusses a user study; Section 7 reflects on our experiences working with Semantic Web technologies in this domain, while Section 8 summarises conclusions and plans for future work.

2. Related Work

Garrigos & Zapater [11] describe a Semantic Web infrastructure for managing real-time traffic information, in which RDF describing vehicle entry and exit times is obtained from sensors on a section of road. This is used to calculate congestion levels and control vehicular access (for example, if the congestion level is high, large vehicles are prohibited). Few details are provided describing the ontologies used and the implementation of the vehicular access control reasoning. While this application of semantic technologies differs from our work, they do identify a key requirement relevant to our work: the importance of providing accurate information to users.

Samper et al [12] use ontologies to integrate traffic information, and support users with search and visualisation tasks. They define a road traffic ontology, covering road and vehicle classifications, location, geography, events, people, and routes. However, application of the ontology is limited to representing information during a multi-agent based traffic simulation, and it is not publicly available.

Plu & Scharffe [13] describe the publication of the Passim and NEPTUNE datasets as linked data. Passim lists French passenger transport services, and NEPTUNE describes French transport lines (stops, timetables, etc.). The publication process involved defining ontologies to model each dataset, and use of the DataLift platform[2] to convert data from CSV (Passim) and XML (NEPTUNE) to RDF, which is published as Linked Data; however, no uses of the data are described.

The Tiramisu system [14] uses a citizen sensing approach to acquire transport information. Bus passengers in three urban areas of the USA use a smartphone app to continuously provide their location and details such as vehicle occupancy. This information is, in-turn, used to provide others with estimated bus arrival times. However, Tiramisu does not consider issues arising from the quality of contributions from the crowd, increasing the risk of imperfect (incorrect, incomplete, or erroneous) information being provided to users.

3. Mobility Information Ecosystem

Figure 1 outlines our mobility information ecosystem. Several ontologies are combined to describe datasets, which are used by web services to support client applications, such as the *GetThere* Android app. Semantic Web technologies are used to support data integration, and distributed data storage and access.

[2] `http://datalift.org/`

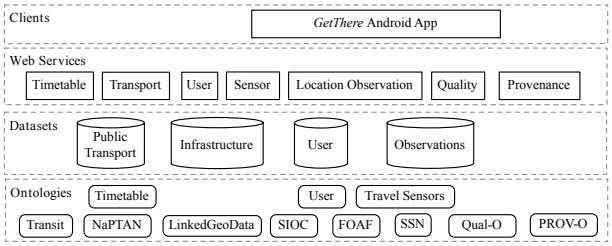

Figure 1: The semantic mobility information ecosystem.

Prefix	Namespace
trn	http://vocab.org/transit
ldg	http://linkedgeodata.org
irpt	https://w3id.org/abdn/irp/transport
irpu	https://w3id.org/abdn/irp/user
sioc	http://rdfs.org/sioc/spec
foaf	http://xmlns.com/foaf/spec
ssn	http://purl.oclc.org/NET/ssnx/ssn
irps	https://w3id.org/abdn/irp/sensors
geo	http://www.w3.org/2003/01/geo/wgs84_pos
qo	http://purl.org/qual/qual-o
prov	http://www.w3.org/ns/prov#

Table 1: Prefixes and namespaces of reused and defined ontologies.

3.1. Ontological Framework

The ontological framework models the data required by the *GetThere* app, including details of public transport services, the transport infrastructure, and vehicle locations. When developing the framework, we followed best practise guidelines and reused existing ontologies where possible [15], using sources such as Linked Open Vocabularies[3] and the Linked Open Data cloud[4]; new ontologies were created only for concepts that were not previously defined. Sub-class relationships between the ontologies were defined by systematically reviewing each concept definition to ensure semantic consistency in the alignment; additional relationships were defined as necessary to link the ontologies. The ontologies are listed in the "Ontologies" layer of Figure 1; Figure 2 expands this to outline the main ontological classes and their inter-relationships.

Public transport services and timetable information are represented using the transit ontology, which was selected as it was based on the General Transit Feed Specification[5], a common format for representing this type of information. The ontology describes public transport *trn:Service*s that operate on *trn:Route*s. Services are described as a list of *trn:ServiceStop*s with associated arrival and departure times. The LinkedGeoData ontology which represents OpenStreetMap[6] data, is used to represent the transport network. Roads are represented as a list of *lgd:Way*s, each with a start and end *lgd:Node* that is associated with a geo-location. The Transport ontology uses this to model the *irpt:BusServiceMap* of roads travelled by buses on each *trn:Service*.

The User ontology extends SIOC and FOAF to describe user profiles and their *irpu:Journey*s on public transport services. The citizen sensing aspects (including vehicle locations) are represented by integrating user details with extensions to the W3C Semantic Sensor Network ontology (SSNO), the de facto ontology standard for sensing applications [16]. SSNO describes *ssn:Sensor*s, their capabilities, and *ssn:Observation*s where a sensor has produced a value for a property of a *ssn:FeatureOfInterest* (thing being observed).

The Sensors ontology extends SSNO to model users as sensors, mobile devices as sensor platforms with attached sensors, the sensing methods they implement, and the types of observations they produce [17]. Example observation types include vehicle location reported by the phone's GPS sensor (captured using the Geo ontology) and occupancy level as observed by the passenger. This highly dynamic data is integrated with other data by modelling the user's *irpu:Journey* as the feature of interest for such observations. The journey refers to both the user (*foaf:Agent*) and the *trn:Route* being travelled, which, in turn, references the bus route map and timetable.

The W3C Provenance Working Group[7], define provenance as "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" [18]. Given the range of data providers within the system and the dynamic nature of the data, recording provenance is critical to ensure that timely and accurate information is provided to users. Provenance is captured using the W3C PROV-O ontology which defines three concepts: *prov:Entity* (things); *prov:Activity*, which occur over time and

[3] http://lov.okfn.org/dataset/lov/
[4] http://lod-cloud.net/
[5] https://developers.google.com/transit/gtfs/
[6] http://www.openstreetmap.org
[7] http://www.w3.org/2011/prov/

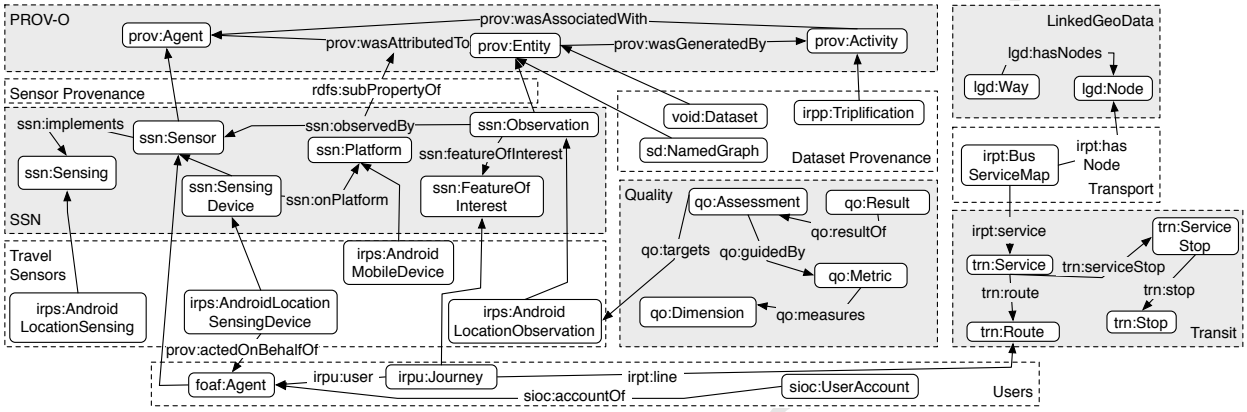

Figure 2: The ontological framework developed to support *GetThere*. Solid rectangles represent classes with labels indicating the class name; labelled arrows represent relationships between classes; unlabelled arrows represent sub-class relationships; dotted rectangles indicate the ontology defining the concepts; and shading indicates ontologies that were reused for this work.

act upon or with entities; and *prov:Agent*s, responsible for activities occurring and entities existing. Extensions of PROV-O in our system capture the processing of sensor data by web services (see Section 3.3 and [17]), and when and how open datasets were published by their owners and imported into the ecosystem. The latter enables web services to select the most recent data when, for example, providing timetable information to *GetThere* users [19].

Understanding the "fitness for use" or quality of data within the system is necessary to enable client applications to provide timely and accurate information to users. Quality is a multi-dimensional construct consisting of a number of quality dimensions assessed using appropriate quality metrics, each of which provides a quality value. These concepts are modelled via the Qual-O ontology [20], which is used to annotate data with quality indicators; the metrics used in our application are discussed in Section 3.3.

3.2. Datasets

Data expressed using the ontological framework instantiates the ecosystem for a particular geographic area and transport domain (bus, rail, air, etc.). We have created the datasets listed in the "Datasets" layer of Figure 1 necessary to support user trials with bus services in the Scottish Borders, UK (see Section 6). A review of open data portals identified existing datasets describing transport infrastructure and public transport access points. Although Linked Data versions of these datasets were available, for reasons discussed below, it was necessary to develop triplification programs that create the specific versions used in our work[8]. As the models of the source and target (RDF) data are semantically identical for each dataset, the triplification processes are straightfoward format conversions. The generated RDF is stored in a new named graph; along with VoID[9] profiles describing each dataset, the provenance of each named graph is also recorded, as described in [19]. Here, the named graph is a *prov:Entity* which was generated by a triplification *prov:Activity*, which has links to the source files and program used. The timestamp of the conversion is also recorded, which can be queried, for example, to select the most recently imported timetable information.

The Infrastructure dataset was created by converting relevant OpenStreetMap XML data (licenced with the Open Database Licence 1.0[10])[11] into RDF expressed using the LinkedGeoData ontology. A triplification script was developed rather than using XSLT to limit the conversion to ways travelled by the included bus services[12]. Should the OpenStreetMap data be updated, the corresponding dataset would also require updating; however,

[8] All triplification programs are available from `https://github.com/dcorsar/ecosystem.timetable`; all of our code is licenced under LGPL V2.1.

[9] `http://rdfs.org/ns/void`

[10] `http://wiki.osmfoundation.org/wiki/Licence`

[11] Extracted using `http://extract.bbbike.org/`

[12] LinkedGeoData, a Linked Data version of the OpenStreetMap data was not used as it did not include details of the road network.

this was not necessary during the deployment described here.

The NaPTAN[13] dataset (licenced with the Open Government Licence (OGL)[14]), provides details of public transport access points (i.e. airports, railway stations, and bus stops) in the UK. The related NPTG dataset[15], also licenced with OGL, provides details of UK transport regions, administrative areas, and localities. While a version of NaPTAN and NPTG are available as Linked Data[16] they were found to be over three years old. To ensure the latest data were used, bus stop details for each stop point (id, name, and location) were extracted from the timetable data (discussed below) and added to a bus stops named graph in the Public Transport dataset. Similarly, a version of NPTG was generated from the source CSV files and stored in a named graph in the Public Transport dataset. The NPTG dataset was created once during this work, as the source data did not change.

Although timetable data were not openly available, this information was provided for this work as ATCO CIF[17] formatted text files that were updated weekly by local government officials. As existing triplification tools could not process these files, a script was developed to generate RDF (expressed using the Transit ontology) from these files and to add it to a timetable named graph in the Public Transport dataset. This dataset also contains a named graph defining a route map for each bus service. These were generated by first manually creating a list of the way ids[18] for each service using route descriptions provided by the bus operator. A triplication script then used this list to generate an RDF version of the map associated with the relevant bus services.

3.3. Services

Several RESTful web services[19] have been developed (listed in Figure 1, "Web Services" layer) that are accessed by the *GetThere* app to provide RTPI. Each web service uses SPARQL v1.1 to query and/or update the relevant datasets in order to provide the RTPI functionalities. This design allows services to use datasets regardless of where they are hosted, and, should an endpoint fail, select and use a mirrored version if one exists.

The *timetable* web service retrieves details of public transport services, schedules, routes, stops, and (timetabled) vehicle locations from the Public Transport dataset. This compliments the *transport* web service, which retrieves details of regions, administrative areas, and localities from the NPTG data. The *user* web service provides an API for managing user profiles, which are stored in the user dataset. The *sensor* web service[20] provides an API for managing observations, sensor outputs, and observation values expressed using SSNO. This is extended by the *location observation* web service which handles creation of new user bus journeys, storage of locations uploaded from the *GetThere* app, and retrieval of the latest locations for vehicles on a specific route. During early field trials it became clear that GPS locations obtained via the *GetThere* app rarely placed the user's smartphone on the actual bus route due to the error associated with the reported GPS location (generally between 5 and 100 metres). This was addressed by the introduction of a map matching algorithm [21], which queries the infrastructure dataset for details of the surrounding road network and calculates an estimate of the user's actual location on the road. If map-matching is successful (it will fail if there are no roads within 100 metres of the reported location), a map-matched location observation (described using extensions of SSNO) is created and added to the observation dataset. The provenance of the map-matched observation, i.e. that it was derived from a GPS observation is also recorded. Such information can be retrieved using the *provenance* web service, which stores and retrieves provenance information using SPARQL queries generated by a prov-API[21].

The *quality* web service[22] assesses data using a SPIN reasoner [22] to apply data quality metrics encoded as SPARQL rules expressed against the relevant ontologies. Such quality metrics are created manually by individuals/developers to reflect the

[13] `http://www.dft.gov.uk/naptan/`

[14] `http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/`

[15] `http://www.dft.gov.uk/nptg/`

[16] `http://openuplabs.tso.co.uk/sparql/gov-transport`

[17] ATCO CIF is a UK standard timetable interchange format defined at `http://www.travelinedata.org.uk/CIF/atco-cif-spec.pdf`

[18] These were retrieved by browsing the OpenStreetMap way web pages, such as `http://www.openstreetmap.org/browse/way/53770302`

[19] Available at `https://github.com/dcorsar/irp-ecosystem-transport`

[20] `https://github.com/dcorsar/sensor-service`

[21] `https://github.com/dcorsar/prov-api/`

[22] `https://github.com/cbaillie/ecosystem-quality`

types of quality assessment they desire for their application. Data items are annotated with the resulting quality scores which can be displayed to users or used by other services to (de)select data. The dimensions and metrics used to evaluate location observations were developed following several field trials of the system; they are: *timeliness* - timely observations are no more than 1 minute old; *accuracy* - accurate observations have a GPS error margin of no more than 25 metres; *relevance* - relevant observations are no further than 100 metres from the expected route of travel; and *availability* - observations with a high availability score should have no more than a 1 minute delay between being created on a user's mobile device and being available from the infrastructure.

It has been argued that such metrics should consider provenance information as part of quality assessment [20]. For example, a quality assessment that does not examine provenance will always decide that a map-matched observation is highly relevant, as it describes a location directly on a bus route. However, the metric shown in Figure 3 examines the provenance of such an observation, which reveals that it was derived from a GPS location that was some distance from the route, and so the map-matched observation is assigned a lower relevance score.

```
CONSTRUCT {
   ?obs a qo:Subject.
   _:b0 a qo:Result.
   _:b0 qo:assessedProperty irps:distanceMoved.
   _:b0 qo:assessedValue ?distance.
   _:b0 qo:affectedInstance ?obs.
   _:b0 qo:hasScore ?qs.
   _:b0 qo:resultOf _:b1.
   _:b1 a qo:Assessment.
   _:b1 qo:guidedBy ?this.
   _:b1 qo:targets ?obs.
} WHERE {
   ?obs a irps:MapMatchedLocationObservation.
   ?obs (prov:wasDerivedFrom)+ ?pObs.
   ?pObs a irps:LocationObservation.
   ?pObs ssn:observationResult ?so.
   ?so ssn:hasValue ?ov.
   ?ov irps:distanceMoved ?distance.
   BIND ((1 - (?distance / 100)) AS ?q) .
   BIND (IF((?q < 0), 0, IF((?q > 1), 1, ?q)) AS ?qs) .
}
```

Figure 3: SPIN rule implementing the relevance metric for *GetThere*.

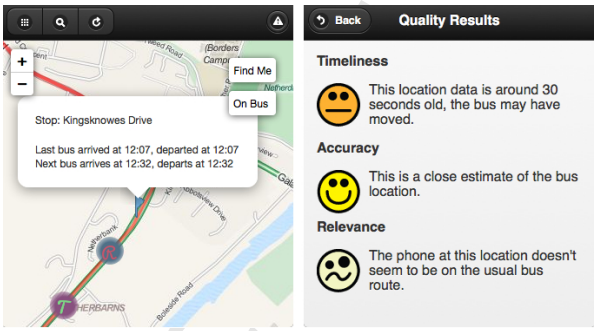

Figure 4: Screenshots of the *GetThere* smartphone app showing vehicle locations (left) and the results of invoking the *quality assessment* service (right).

4. The *GetThere* Smartphone App

The mobility information ecosystem has been designed to support a range of applications; at present it is used by the *GetThere* Android smartphone app (see Figure 4). The app presents users with a list of available bus routes; after selecting a route (and direction, either inbound or outbound), vehicle locations are displayed. These locations include both estimates based on the timetable and real-time locations crowd-sourced from other users on that route (Figure 4, left image). Bus stops along the route are also shown. A user can access timetable information at a particular stop by tapping on it. On boarding a bus, they tap a button to have their location uploaded to the *location observation* web service every minute. This upload frequency was determined based on field trials to balance the impact on the phone's battery with provision of sufficient information for waiting passengers to monitor the location of a bus, and estimate its arrival time at their location.

Users can also view quality assessment results for a real-time vehicle location by tapping on its icon. The result of each assessed quality dimension is shown via a graphical visualisation and accompanying textual description (Figure 4, right image). For example, timeliness is described by an amber neutral face indicating an intermediate quality score; accuracy represented by the smiling face indicating a high score; the confused face describes relevance, which has been assigned a low score as the GPS location was particularly distant from the route. It is left to the user to decide what, if any, action they should take by considering the data presented by the app in relation to their present situation; for example, a user provided with the quality results

shown in Figure 4 while waiting for a bus may decide to consult the bus operator's website to ascertain if the bus has been diverted from its normal route due to, for example, a road closure.

5. Performance Evaluation

To gain an understanding of system performance with a large number of users, a number of simulation experiments were performed. In particular, we were interested in the response times users would experience when they requested real-time information via the app; in accordance with guidelines described in [23] this should ideally be under a second, to allow "the user's flow of thought to stay uninterrupted" and at most within 10 seconds to maintain the user's attention.

The simulations used an identical setup to that employed for the later user trial (Section 6). The ecosystem was deployed on a server with an Intel Xeon E5-2407 @ 2.20GHz CPU, 100Gb SSD HDD, and 16Gb of memory. Fuseki with TDB stores was used to host the following datasets (sizes in brackets indicate the maximum Java heap size): public transport (3Gb), infrastructure (4Gb), and users (1Gb); the observations dataset was hosted by Sesame with a MySQL repository. Both Sesame and the ecosystem services were hosted in a Tomcat instance (with 2Gb heap size). The allocated heap sizes are based on previous testing and reflect the usage patterns of the datasets.

Two types of agent were used in the simulations: location providers, and information requesters. Location providers simulated passengers on vehicles, uploading locations to the server every 60 seconds (the same as the *GetThere* app). The uploaded locations were based on locations derived from the timetable. A number of location provider agents were simulated travelling on a number of routes, with new journeys started every 30 minutes. Each simulation also featured a number of information requester agents; these queried the ecosystem every minute for the locations of vehicles on a specific route; each requester agent made 1440 requests, simulating use of the system for 24 hours. The response times were recorded for each request (both location upload and inquiry). Table 2 summarises the results of the simulations: each row details the number of location provider agents simulated on each vehicle, the number of routes simulated, the number of information requesters simulated for each route, the total number of requests made for real-time vehicle locations, and the minimum, median, mean, maximum, 95th percentile and 99th percentile response times for those requests.

6. User Trials

A user study was conducted in order to determine: (i) if the semantic infrastructure could support a deployed RTPI system; and (ii) to investigate the effects of the information provided by the *GetThere* app on public transport users in rural areas. The trial took place in the Scottish Borders with 15 participants recruited from the Scottish Borders Campus of the Heriot-Watt University. Users were selected based on the frequency of their public transport journeys, bus services used, and origin-destination points. At the outset individuals were interviewed to understand their attitudes regarding various aspects of the bus service, such as: perceived wait time, feelings of satisfaction with the bus service, control over their journey, security, and willingness-to-pay for the bus service and information about it.

The infrastructure was configured with datasets for seven bus services, namely the First South East and Central Scotland 62, 72, 73, X95/95/95A, 396, and 397. The trial lasted three weeks, and during this period participants were asked to use the app to view and contribute information during their bus journeys. The participants contributed a total of 1887 real-time locations during 119 journeys, and made 347 requests for bus locations. Following the trial, participants were re-interviewed using the earlier questions. Verbatim transcripts of the interviews were analysed by four researchers who clustered each participant's responses around the different effects of RTPI (discussed below).

The results indicated that 14 participants felt the information from *GetThere* increased their sense of control during their journey and journey planning; 12 participants felt it made the bus service easier to use; 12 were willing to pay for the information; 11 felt the bus service had improved; 11 perceived a reduction in the time spent waiting at bus stops; and 10 felt an increase in their satisfaction with the bus service. During the second set of interviews, participants made several positive comments about the information provided by the app. For example, one participant observed that the information provided by the app was more accurate than the printed timetables available at bus stops, causing

No. of Providers	No. of Routes	No. of Requesters	Total Requests	Response Time (seconds) Minimum	Median	Mean	Maximum	95th Percentile	99th Percentile
1	6	10	172800	0.08	0.54	0.110	8.786	0.25	1.44
1	8	15	345600	0.01	0.09	0.35	32.15	1.63	4.04
2	6	10	172800	0.01	0.21	0.92	19.65	4.18	6.30
2	8	15	345600	0.004	0.26	4.33	2843.14	22.05	28.76
3	6	10	172800	0.01	0.48	1.97	39.58	7.90	16.04
3	8	15	345600	0.01	0.77	6.39	161.45	28.43	35.28
4	6	10	172800	0.001	0.90	3.65	120.92	14.31	24.32
4	8	15	345600	0.001	1.822	9.94	1024.94	33.61	73.22

Table 2: Summary of real-time vehicle location requests recorded during the performance simulations.

them to stop using the printed information. Another participant stated that while the timetable information (in the app) was good, they preferred seeing a real-time location as they could trust that as a more accurate representation of the bus location. Another user described using the real-time bus location on the app to verify sightings of the bus provided by family members elsewhere along the route. Further, recognition of the value of information provided by the app led to two participants becoming a source of bus route and timetable information for others in their kinship network.

7. Discussion

We will now review the implementation of the *GetThere* smartphone app in terms of the desirable capabilities identified in Section 1. Capabilities C1, C2, and C3 are implemented using open data and location data contributed by *GetThere* users; the information provided to users is personalised (C4) to the extent that it focuses only on the bus service(s) of interest to them. To improve the accuracy and timeliness of this information, the provenance of each dataset was recorded to ensure that only the latest data is used by *GetThere*. Considering the infrastructure requirements I1-I4, the ontological framework models the transport infrastructure, public transport services, and vehicle locations (I1), and has been shown to support the integration of heterogenous mobility data from multiple providers (I2). Data provenance is captured, and used by the quality assessment service to support reasoning about data quality (I3); and the user trials of *GetThere* demonstrate use of the ecosystem to provide an RTPI system (I4).

The *GetThere* system benefited from the use of Semantic Web technologies, in particular from the use of RDF as a common data model, ontologies to provide formal vocabularies, SPARQL for data retrieval and update, and the SPIN rule engine for quality reasoning. The main benefits of using RDF over other technologies, such as NoSQL, for data integration, were the native support for named graphs, and annotation of data with schema information and provenance meta-data. To achieve the equivalent functionalities with NoSQL technologies would have required development of custom approaches for each NoSQL engine used. We also benefited from the standard representational formalism across all the data types, provided by the use of ontologies; otherwise an alternative approach would have been required to unify XML Schema, CSV, and formatted text files. Further, we were able to explicitly define the relationships between concepts, particularly those defined by independent schemas (such as the link between locations contributed by users and a bus service), in a manner than can be easily shared and reused. Exposing the data via SPARQL endpoints simplified development of the web services, as these utilise SPARQL v1.1 requests to query and/or update data in any of the datasets, regardless of how data persistence is implemented. In contrast, use of alternative data storage technologies (that do not support SPARQL) would have required each web service to use different APIs to access different datasets, or to use a proxy web-service developed specifically to access a dataset; either of these approaches would have increased the development effort. Finally, using the SPIN reasoner allows the quality service to be more easily maintained and extended by simply updat-

ing SPARQL rules. This is in contrast to performing the reasoning in software code that would have been required if Semantic Web technologies were not used. This also allows the results to be more transparent by, for example, associating details of the reasoning activity and rules that were used with the inferred data.

Our experience of attempting to reuse datasets from the Web of Linked Data highlights the importance of meta-data describing the provenance of such datasets. Ideally this should include details of when and how the dataset was created, including if it is derived from an alternative version published elsewhere. Ideally any source versions will also be associated with meta-data that allows users to determine if the Linked Data reflects the latest version of the dataset. The licences associated with the open datasets that we reused permit republishing of the data, and so our RDF versions could be published as open data with appropriate provenance information acknowledging the original source. Regarding the ontologies we reused, the availability of documentation describing each ontology was a key factor in their selection, particularly when examples were available illustrating their intended usage.

The types of information modelled by the ontological framework (details of public transport services, timetables, location observations, and infrastructure) are not specific to the bus domain. The transit ontology was designed for public transport (including modes such as rail and ferry), and the Node and Way concepts of OpenStreetMap can also used to model railway lines and ferry routes. Similarly, the concepts are not specific to a particular geographic area[23]. Based on our experience working within the transport domain, we believe that the ontological framework, and supporting services could easily be utilised for other modes of transport and geographic areas.

The performance evaluation assessed the scalability of our solution with a considerably more intensive workload than the user trials. The increase in response time, particularly beyond the 10 second limit for maintaining the user's attention, indicates that if a similar workload were to be expected during deployment, it would be necessary to use additional techniques such as load-balancing or use of semantic streams [24] to maintain an acceptable user experience.

While privacy issues were not raised by participants of the user trial, by asking users to share their location with the *GetThere* app there is the potential of them being exposed to privacy risks. Work exploring potential risks to privacy in systems that utilise personal data in semantic ecosystems (including *GetThere*) is reported in [25].

With regards to potential limitations of *GetThere*, the compromise of uploading locations every minute (as discussed in Section 4), may reduce the quality of real-time information provided to users. Additionally, a limitation of the crowd-sourcing approach is that real-time information is only available for buses on which passengers are willing to share their location, meaning area-wide coverage cannot be guaranteed.

8. Conclusions

In this paper we have described a semantic mobility information ecosystem developed to provide RTPI, and a user-based evaluation of the *GetThere* smartphone app. The ontological framework provides a generic model for mobility information, that is not specific to the bus domain or the Scottish Borders area; as such, in the future we plan to explore its use to support further applications, for example to provide RTPI for other modes of public transport. We also plan to explore how social media can be incorporated into the infrastructure, as a channel for both provision of travel information to passengers, and as a means to obtain information about the transport network. Twitter[24] in particular is increasingly being used by public transport operators to publish details of disruptions to bus services, and we plan to explore the use of Semantic Web technologies to infer descriptions of reported disruptions and to assess the veracity of the reported information. Overall, we believe this work illustrates that Semantic Web technologies are ready to play an important role in addressing the data management challenges faced by smart mobility applications that deliver benefit to citizens.

Acknowledgements The research described here was supported by the award made by the RCUK Digital Economy programme to the dot.rural Digital Economy Hub; award reference: EP/G066051/1.

[23] During preliminary testing of *GetThere*, we successfully deployed the system for another geographic area by creating datasets for the county of Aberdeenshire.

[24] `http://www.twitter.com`

References

[1] A. Gooze, K. Watkins, A. Borning, Benefits of Real-Time Information and the Impacts of Data Accuracy on the Rider Experience, in: Proc. of the Transportation Research Board 92nd Annual Meeting, 2013.

[2] J. Grotenhuisa, B. Wiegmansa, P. Rietveldb, The desired quality of integrated multimodal travel information in public transport: Customer needs for time and effort savings, Transport Policy 14 (1) (2007) 27–38. doi:10.1016/j.tranpol.2006.07.001.

[3] K. Ganesh, M. Thrivikraman, J. Kuri, H. Dagale, G. Sudhakar, S. Sanyal, Implementation of a real time passenger information system, Intl. Journal of Engineering, Sciences, and Management 2 (2) (2012) 15–27.

[4] T. Berners-Lee, Putting government data online (2009).

[5] V. Lopez, S. Kotoulas, M. Sbodio, M. Stephenson, A. Gkoulalas-Divanis, P. Aonghusa, Queriocity: A linked data platform for urban information management, in: The Semantic Web – ISWC 2012, Vol. 7650 of LNCS, Springer Berlin Heidelberg, 2012, pp. 148–163.

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

[7] A. Sheth, Citizen Sensing, Social Signals, and Enriching Human Experience, Internet Computing, IEEE 13 (4) (2009) 87–92.

[8] K. Papangelis, A. Chamberlain, N. Velaga, D. Corsar, S. Sripada, J. Nelson, M. Beecroft, People, plans and place: Understanding and supporting responses to rural public transport disruption, in: Proc. of the 11th Intl. Conference on the Design of Cooperative Systems, 27-30 May 2014, Nice (France), Springer International Publishing, 2014, pp. 1–15.

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

[10] O. Hartig, J. Zhao, Using web data provenance for quality assessment, in: J. Freire, P. Missier, S. Sahoo (Eds.), Proc. of Workshop on Semantic Web in Provenance Management, Vol. 526, CEUR Workshop Proceedings, Washington D.C., USA, 2009.

[11] D. Garrigos, J. Zapater, Semantic infrastructure system with feedback: Application on traffic, in: Telematics and Information Systems (EATIS), 2012 6th Euro American Conference on, 2012, pp. 1–5.

[12] J. Samper, V. Tomas, J. Martinez, L. van den Berg, An ontological infrastructure for traveller information systems, in: Intelligent Transportation Systems Conference, 2006. ITSC '06. IEEE, 2006, pp. 1197–1202.

[13] J. Plu, F. Scharffe, Publishing and linking transport data on the web, in: Proc. of the First Intl. Workshop On Open Data, 2012, pp. 62–69.

[14] A. Steinfeld, J. Zimmerman, A. Tomasic, D. Yoo, R. D. Aziz, Mobile transit rider information via universal design and crowd-sourcing, in: 90th Annual Meeting of the Transportation Research Board, 2011.

[15] B. Hyland, G. Atemezing, B. Villazon-Terrazas, Best practices for publishing linked data, W3C Working Group Note, http://www.w3.org/TR/ld-bp/ (2009).

[16] L. Lefort, C. Henson, K. Taylor, Semantic sensor network xg final report, W3C Incubator Group Report (June 2011).

[17] D. Corsar, P. Edwards, C. Baillie, M. Markovic, K. Papangelis, J. Nelson, Short Paper: Citizen Sensing within a Real Time Passenger Information System, in: Proc. of the 6th Intl. Workshop on Semantic Sensor Networks, Vol. 1063, CEUR Workshop Proceedings, 2013, pp. 77–82.

[18] L. Moreau, P. Missier, Prov-dm: The prov data model, W3C Recommendation, http://www.w3.org/TR/prov-dm/ (April 2012).

[19] D. Corsar, P. Edwards, N. Velaga, J. Nelson, J. Z. Pan, Exploring provenance in a linked data ecosystem, in: P. Groth, J. Frew (Eds.), Provenance and Annotation of Data and Processes, Vol. 7525, Springer Berlin Heidelberg, 2012, pp. 226–228. doi:10.1007/978-3-642-34222-6_21.

[20] C. Baillie, P. Edwards, E. Pignotti, Qual: A provenance-aware quality model, Journal of Data and Information Quality 5 (3) (2015) 12:1–12:22. doi:10.1145/2700413.

[21] N. Velaga, J. Nelson, P. Edwards, D. Corsar, S. Sripada, N. Sharma, M. Beecroft, Development of a map-matching algorithm for rural passenger information systems via mobile phones and crowd-sourcing, Journal of Computing in Civil Engineering (2013) 732–742.

[22] C. Furber, M. Hepp, Swiqa - a semantic web information quality assessment framework, in: 19th European Conference on Information Systems, 2011, pp. 922–933.

[23] J. Nielsen, Response Times: The 3 Important Limits, https://www.nngroup.com/articles/response-times-3-important-limits/, accessed March 2016 (Jan 1993).

[24] A. Margara, J. Urbani, F. van Harmelen, H. Bal, Streaming the web: Reasoning over dynamic data, Journal of Web Semantics 25 (2014) 24–44. doi:http://dx.doi.org/10.1016/j.websem.2014.02.001.

[25] D. Corsar, P. Edwards, J. D. Nelson, Personal privacy and the web of linked data, in: Proceedings of Privacy Online 2013, a workshop co-located with the 12th International Semantic Web Conference (ISWC 2013), 2013.