# Towards a Semantic Layer to Support Road and Public Transport User Decision-making

Nicholas Gould School of Science and the Environment Manchester Metropolitan University Manchester, UK n.gould@mmu.ac.uk

Abstract— Municipalities now collect large amounts of data that could be potentially used to guide road users and public transport users in their decision-making. However, these data sources are heterogeneous and a semantic layer, expressed in an ontology, is proposed to match data sources, traveler needs, and the context of their journeys in order to support decision-making.

## Keywords—ontology; big data; road transport; public transport

## I. INTRODUCTION

Intelligent Transport Systems (ITS) are an integral part of the "Smart City" [1]. As well as economic benefits, ITS can provide environmental benefits by reducing emissions by reducing congestion and encouraging modal shift [2]. The quality of public transport can be increased by increasing the quality of passenger information systems [3] and this is the focus of this work. The concept of improving information can also be extended to road users; in the face of road congestion, routes can be changed or users can encouraged to switch to a different transport mode if sufficient warning of disturbance in the network can be provided. The data sources are multitude: passive sensors on the road network, live rail service data, timetables, road accident reports, roadwork schedules and social media. The challenge is to convert these heterogeneous data sources into useful information for travelers be they road users or train passengers. This position paper describes two separate, upcoming, projects, one related to road transport and the other to passenger rail transport, but both related in their desire to provide relevant, personalized, real-time information to the traveler.

The problem is not only of heterogeneous data sources but also disparate traveler needs. The information needs of an experienced rail commuter, usually on "auto-pilot", are different from those of the tourist using the rail station for the first time. Similarly, the local driver is aware that certain roads should be avoided at the time of the "school run" whereas a delivery driver may not be. A rail passenger travelling with children and heavy suitcases has different needs from those of a businessperson, late for a meeting. There is therefore a need for personalized travel information systems; systems that provide information such as "your usual driving route to work is currently congested but there David Atkin Transport for Greater Manchester Manchester, UK David.Atkin@tfgm.com

is a bus, leaving in 10 minutes, which will take you to your destination leaving from a stop 5 minutes walk away".

The problem can be illustrated if we consider the options facing the passenger arriving at Piccadilly railway station, a major transport hub in Manchester, England (Fig. 1). On arrival, the passenger has six modes to consider for the next leg of her journey: another train, walking, a taxi, light-rail (tram), bicycle hire, and a free shuttle bus. Live train times and the published train timetable can be used to help the choice of taking another train. The tram system, however, has no timetable but an advertised service frequency; the free shuttle bus has neither timetable or advertised frequency but a fixed route and a current position. The taxi option has a passenger queue length and a rank length; the departure time is a function of both of these. The personalized information systems should take into account the needs of the traveler and suggest the best choice, based on current information about each of these modes.

However, to be truly useful, traveler information systems need to react to disruption and not be merely travel planning services. The ability of a transport network to react to both planned events, such as football matches and road works, and unplanned incidents, such as traffic accidents and flooding, is seen as a key component of the Smart City [1]. Such events cause transport system failure (e.g. traffic congestion) and decreasing spatial access (e.g. for commuters, shoppers and hospital visitors), which has a major impact on the environment in the form of pollution, can restrict economic activity and also impacts on the quality of urban life. Transport system managers also need relevant, up-to-date, information to manage the road and public transport systems.

If the aim of the Smart City is to "improve the efficiency, equity and quality of life for its citizens in real time" [1] then this is particularly true of transport, where in train travel, for example, the lack of information provided in the event of disruption has been identified as a major irritation among passengers [4]. If more road users are to be encouraged to use public transport then real-time, personalized information systems are needed to provide the confidence that, if there is a disruption to their journey, they will receive timely and relevant information.



Fig. 1. Options facing the train traveler on arrival at Manchester Piccadilly and the factors affecting decision-making.

Much of this data is currently in transport mode specific silos and hence many of the current passenger information systems are restricted to a single mode such as rail travel. When disruption occurs, travelers may want to switch mode but this may mean switching to a different information system.

Any system that takes as input a set of heterogeneous data sources, a set of disparate user profiles and a heterogeneous set of events to influence a variety of responses requires a complex conceptual model. The aim of this research is to develop such a model, in particular, a model that can support reasoning. Any such model should be *open*, to allow its use by any application developer and data provider.

## A. Intelligent transport systems and big data

The relevant data are in a variety of forms: rail timetable information may be in a traditional RDBMS; tram service frequency and service start and stop times are "small data" but unformatted; road sensor data is high volume and high velocity. It might seem that the there is a "polyglot persistence" architecture here [5] but this term relates to enterprises or organizations, which have some degree of control over their own data. Here the data is coming from a number of different suppliers and may not persist. Data services may suffer temporary outages or be permanently withdrawn; sources that were once free may start extracting a charge or change their terms and conditions. It is therefore necessary to ensure that any model can embrace multiple data sources.

The aim is not to develop a traveler information system but to develop a semantic layer that sits between the data sources and any such system. If a data source is added or removed then the application using it does not need to be rewritten, the semantic layer just needs to be updated and the application can switch to a data source, described in the semantic layer, that provides semantically similar information. The relevance of big data to ITS can be illuminated by considering the five V's of big data: Volume, Variety, Velocity, Value and Veracity [6].

*Volume*: For example, the "TrafficMaster" database that holds historic road speed data, derived from GPS, for Greater Manchester, currently holds four billion records. Three terabytes of data have so far been collected from roadside Bluetooth sensors in Greater Manchester.

*Variety*: the example of Piccadilly station described earlier illustrates the problem of variety in the data sources but there is also variety in the nature of possible disruption. Planned roadworks have estimated start and end times whereas a road accident is unpredictable and its impact difficult to assess. Also, what is the impact on a train's departure time if its driver is arriving on another service that is running late?

Velocity: The volume of rail transport related data is less of a problem than its variety; there are only so many trains that can be accommodated by the regional rail network. The same cannot be said of the road network. At any time, thousands of vehicles can be on the Greater Manchester road network. Strategically placed roadside sensors record the passing of every vehicle. Currently, 120 Automatic Traffic Counters and 600 passive sensors (Bluetooth) return packets of data every 5 minutes. Real-time processing of the data is challenging; sensor data has to be cleaned by removing outliers, such as cyclists from general traffic and vehicles that have dwelt within a link, before the data can be used to describe how a section of road or the road network as a whole is performing. This means confidence in real-time or nearreal time data is low and it can only be viewed as indicative of road conditions. As a result the sensor data is primarily used for reporting past road network performance rather being used to represent the current situation. Big data processing techniques such as grid and cloud computing will be valuable in providing real-time road data.

*Veracity*: Attempts have been made to use social media to predict both current [7] and future road congestion [8] but such data have to be treated circumspectly. For example, a search of Twitter for road accident reports in Greater Manchester revealed one message that likened the contributor's life to a "road crash". Sensor data also has veracity issues when, for example, a signal from a pedestrian's mobile phone is taken for that of a car user.

*Value*: Some transport data sources have more value than others. For example, timetable information has more value in medium and long term planning but live train information has more value in the present.

## B. Intelligent transport systems and the Smart City

If any recent concept has attracted more skeptics than "Big Data" then it is the "Smart City". The Smart City is regarded by many as being dominated by large ICT corporations and there has been criticism of a top-down, centralized and software-driven approach [9]. There are also concerns of technological lock-in where ICT corporations promote the idea of a single solution that



Fig. 2. Measuring road speeds using two different techniques



Fig. 3. Comparing arrival times for two different modes

covers all aspects of the Smart City [10]. These discussions may be the realm of the social rather than computer scientist but in computing there is now a long history of open software, standards and data. Our aim of developing an open, shared, conceptual model supported by open standards should alleviate at least some of the problems of technological "lock-in".

Why should cities develop their own systems to monitor the road network, for example, when there are freely available sources such as Google Traffic, which map road speeds in realtime? Firstly, there are technological limitations; the Google Maps API does not currently allow for the export of raw data, which is vital for any system that hopes to combine information relating to multiple transport modes. Secondly, there are concerns over cities becoming reliant on private data sources for transport data [11]; what happens when the supplier withdraws the service?

# II. APPROACH

It is necessary to consider how to model the data sources, traveler needs and the context of travel, such as disruption.

# A. A semantic layer

To ensure the provision of relevant information it is necessary that information systems provide semantic interoperability [12]. The use of a semantic layer to support decision-making in ITS has been proposed to take into account traveler constraints and external context [13]. In particular, ontologies have been suggested as a way of resolving semantic heterogeneity in data sources [14]. The aim of this research is to extent this work by considering semantic relations beyond synonymy, partonomy and subsumption.

As an example, consider Fig. 2. Here we have two data sources; one is a stream of GPS location data from smart phones in moving vehicles, the other is a stream from fixed location roadside sensors that record the MAC addresses of passing smart phones in vehicles, using Bluetooth technology [15]. Although both techniques rely on the presence of smart phones in vehicles, the GPS-based method involves a moving sensor, whereas the Bluetooth method relies on a stationary sensor. Both techniques can be used to estimate road speeds but use subtly different methods (Fig. 2).

The next example (Fig. 3), considers two different modes, connecting rail to light-rail (tram). The requirement is to determine if the arrival of a train coincides with the arrival of a tram. To estimate the arrival time of a train, its identifier is required, then either the timetable database or the live train feed can be used, depending on availability. Since both systems use the same input and provide the same result, the model will need to provide a means of prioritizing data sources. Since the tram system has no published timetable, only a range of times can be given for its expected arrival. The tram operators publish a list of service frequencies at different times of the day, so the current time is needed to obtain the current frequency and from that, a maximum wait time can be calculated and then compared to the train arrival time.

One aim of the semantic layer, then, is to describe the inputs and outputs of the various data sources and how the outputs may be matched semantically.

## B. The use of personas

The previous example assumed that the train and tram were co-located, but in reality there will be a travel time between the two. We tend to think of transport hubs as simple nodes in the network, but they are not; they have internal structures that may affect particular customers in particular situations. For example, an arrival at platform 14 at Manchester Piccadilly (which is physically remote from the main station) is very different from an arrival at platform 7 (which is close to the main concourse). If my destination is close to Oxford Road station (a smaller Manchester station), it might be sensible to cross a few yards to platform 13 and then take another train to Oxford Road station, rather than walk down three separate flights of steps to the tram stop. This decision is dependent on my attitudes, circumstances and personality and these can be modelling using personas, fictional characters that are used to represent a range of real users [16].

The concept of a traveler's *persona* will be incorporated in the design of the semantic layer. This concept, which has its roots in user-centered design [16], will help to understand user needs, in particular what information is required given their circumstances [17]. A businessperson, late for a meeting, might prioritize speed over cost, whereas a tourist, on a budget, the reverse.

Much of transport planning and information is based on the perceived needs of the "average" commuter who works from 9am to 5pm in the same office, from Monday to Friday. However, in recent years that model has changed as more "flexible" work patterns have become widespread, with workers frequently having multiple jobs and working irregular hours in different locations. A study of commuters in Beijing [18] identified "extreme commuters" such as "night owls", who commute late at night, and "recurring itinerants" who make multiple journeys in the same day. Identifying the diversity in the transport user cohort can help determine information needs; the "night owl", for example, will be hindered by lack of transport system staff and fellow travellers to provide up-to-date information in the event of disruption.

Although, in this research, the focus is not on user-interface design but on conceptual model design, the development of personas can help define the different kinds of knowledge that must be represented. The transport domain is large and complex and the temptation is to capture all aspects of it. The use of personas can help to narrow the focus to what is necessary.

By considering transport users as individuals with particular needs, the use of personas to develop the model of transport counters the criticisms of the Smart City as being centralised and top-down. The road network manager or rail station manager have personas just like travellers. The use of personas helps to answer criticisms of big data that it lacks depth and detail and does not attempt to understand the individual's viewpoint [19].

# C. The context of travel

After data services and personas, the third part of the model will focus on context. Most existing transport information systems assume a disruption free journey but to provide a personalized real-time information system it is necessary to provide contextual information such as disruption, the current weather and the traveller's current situation (are they at a transport hub, for example, or a rural bus stop?)

We can advise a traveller to take a train if they are "near" to a railway station but concepts such as "nearness" are dependent not only on geometry but on the traveler's persona and the context of their journey; an infirm person's idea of nearness is different from that of a fit and healthy person; if a journey to the railway station is to be taken on foot then the concept of what is near will be influenced by weather conditions. It will be necessary to relate the different aspects of a traveller's persona with the data sources and the context of travel. The generic public transport framework develop by [3] can be made mode specific to allow it to be extended to road transport (Fig. 4). The persona of the traveler remains a constant (at least for any single journey), but the context of their journey and the relevant data sources may change depending on mode.

# D. The use of ontologies

Semantic interoperability is required to capture the context of transport data, the needs of the traveler and the context in which they find themselves. One way of encapsulating the semantics of a domain, such as transport, is to represent domain knowledge in an ontology [20]. Using an ontology to describe a domain can lead to intelligent knowledge retrieval using reasoning [21]. For example, we can use the ontology to infer that the output of the GPS system and roadside sensor in Fig. 2 are the same thing, even though they have different inputs. Once the ontology has been created, the first task is to determine which data sources are relevant for a given persona in a given context. i.e. which data sources will support the decision-making process? If a railway station is closed due to a fire, for example, then the rail timetable, for that station at least, is irrelevant.

The scope of the ontology needs to be prescribed. We might, not, for example, represent the nodes and edges of a road network in an ontology, but we might represent the knowledge that for a car journey to be possible between two locations they have to be on the same road network. Similarly, we would not store a rail timetable in an ontology, but we would store the knowledge that a rail timetable is relevant information for a rail journey, and also the nature of the information it provides.

The semantic annotation of data sources using ontologies is a well-established method but the encapsulation of personas is not. However, the formalization of personas in an ontology will ensure their consistency [22] and allow for the development of more complex personas than simply "commuter" or "tourist".

An ontology is defined as a shared conceptualization and its use will allow the development an open, generic model using a bottom-up approach.

# III. METHODOLOGY

The aim is to build the model in OWL (Web Ontology Language) using a "breadth-first" approach to consider all aspects of the domain. The model design will be driven by the need to satisfy the needs of personas: classes of system users, such as network manager, bus scheduler, commuter and bus driver.

The, potentially, very wide scope of the project will be prescribed by focusing on two particular personas, one currently using the road network and another currently using the rail network.



myHouse near HeatonRailStation hasOrigin Location RailStation hasDestination myOffice near PiccadillyRailStation

Fig. 5. Modelling the concept of a journey

# **IV.** CONCLUSIONS

Fig. 4. Multi-modal framework

For example, the road user persona might be described as a car driver who needs to arrive at a meeting by 10am, the context being that their preferred route is blocked by a road accident caused by icy conditions. The method of utilizing detailed personas is a "depth-first" approach that examines narrow vertical strips of the domain.

The aim is to capture the semantics of travel, which can be demonstrated in a simple example. We can define an object (instance) *myJourneyToWork*, which has an origin and destination, both objects in the *Location* class (Fig. 5, where dashed edges represent is-A relationships and labelled edges represent object properties). We can also assert that each location is "near" a RailStation. This is the context of the potential journey. If we define a *RailJourney* class as having an origin and a destination, both of which are near a rail station:

```
(hasDestination some
  (Location
    and (near some RailStation)))
and (hasOrigin some
    (Location
    and (near some RailStation)))
```

then we can infer that myJourneyToWork can be a RailJourney and from that, infer what information is needed. This is obviously a big simplification of reality; for example, the concept of nearness depends on the persona of the traveler and the context of travel. Whereas "near" is represented as a relationship in Fig. 5, it will have to be modelled, in the next iteration of the ontology, as a class to allow it to have its own relationships and attributes.

Once the ontology has been evaluated against a set of *competency questions* then next stage is to develop a prototype that will reason with the ontology, via the OWL API, using a reasoner such as HermiT, to make decisions that satisfy the needs of the persona in a given context.

The development of a semantic layer will help the integration of multiple data sources that are relevant to a particular traveler in a particular context. Rather than having to rewrite application code if a particular data source becomes temporarily or permanently unavailable, alternative sources can be utilized. All that is required is the description of the new data source in the semantic layer.

The novelty in this research is that it proposes a semantic layer that encapsulates traveler personas, disparate data sources and the context of travel in an ontology and that it aims to model both public transport and road transport. As [19] state "as the backlash against 'big data' increasingly stresses the importance of domain knowledge, the ability to build sound models from theoretical insights continues to carry weight in practice"; the semantic layer described above aims to be such a model.

## REFERENCES

- M. Batty, K. W. Axhausen, F. Giannotti, A. Pozdnoukhov, A. Bazzani, M. Wachowicz, G. Ouzounis, and Y. Portugali, "Smart cities of the future," The European Physical Journal Special Topics, vol. 214, no. 1, pp. 481-518, 2012.
- [2] S. Grant-Muller, and M. Usher, "Intelligent Transport Systems: The propensity for environmental and economic benefits," Technological Forecasting and Social Change, vol. 82, pp. 149-166, 2014.
- [3] S. Hörold, C. Mayas, and H. Krömker, "User-oriented information systems in public transport," Contemporary Ergonomics and Human Factors 2013, pp. 160-167: Taylor & Francis, 2013.
- [4] Transport Focus. "National Rail Passenger Survey: Spring 2015 Main Report," http://www.transportfocus.org.uk/research/publications/nationalrail-passenger-survey-nrps-spring-2015-main-report.
- [5] P. Sadalage, and M. Fowler, NoSQL distilled: a brief guide to the emerging world of polyglot persistence. Addison-Wesley, 2012.
- [6] Y. Demchenko, P. Grosso, C. De Laat, and P. Membrey, "Addressing big data issues in Scientific Data Infrastructure," in International Conference on Collaboration Technologies and Systems (CTS), San Diego, CA, 2013, pp. 48-55.
- [7] B. Pan, Y. Zheng, D. Wilkie, and C. Shahabi, "Crowd sensing of traffic anomalies based on human mobility and social media," in Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Orlando, Florida, 2013, pp. 344-353.

- [8] J. He, W. Shen, P. Divakaruni, L. Wynter, and R. Lawrence, "Improving Traffic Prediction with Tweet Semantics," in Twenty-Third International Joint Conference on Artificial Intelligence, Beijing, China, 2013.
- [9] O. Söderström, T. Paasche, and F. Klauser, "Smart cities as corporate storytelling," City, vol. 18, no. 3, pp. 307-320, 2014.
- [10] R. Kitchin, "Making sense of smart cities: addressing present shortcomings," Cambridge Journal of Regions, Economy and Society, vol. 8, no. 1, 2015.
- [11] E. Morozov, "Our cities shouldn't rely on Uber to devise new transport choices," The Observer, 1<sup>st</sup> February 2015.
- [12] A. P. Sheth, "Changing Focus on Interoperability in Information Systems: From System, Syntax, Structure to Semantics," Interoperating Geographic Information Systems, The Springer International Series in Engineering and Computer Science M. Goodchild, M. Egenhofer, R. Fegeas and C. Kottman, eds., pp. 5-29: Springer US, 1999.
- [13] A. Costanzo, and A. Faro, "Real time decision support systems for mobile users in intelligent cities," in 6th International Conference on Application of Information and Communication Technologies (AICT), Georgia, Tbilisi, 2012, pp. 1-5.
- [14] J. Zhai, J. Jiang, J. Yu, and J. Li, "Ontology-Based Integrated Information Platform for Digital City," in 4th International Conference on Wireless Communications, Networking and Mobile Computing, Dalian, China, 2008, pp. 1-4.
- [15] C. Bachmann, B. Abdulhai, M. J. Roorda, and B. Moshiri, "A comparative assessment of multi-sensor data fusion techniques for freeway traffic speed estimation using microsimulation modeling," Transportation Research Part C: Emerging Technologies, vol. 26, pp. 33-48, 2013.

- [16] J. Garrett, The Elements of user experience: user-centered design for the web and beyond, Berkeley, California: Pearson Education, 2011.
- [17] S. Hörold, C. Mayas, and H. Krömker, "Identifying the information needs of users in public transport," Advances in Human Aspects of Road and Rail Transportation, N. Stanton, ed., pp. 331-340: CRC Press, 2012.
- [18] Y. Long, X. Liu, J. Zhou, and Y. Chai, "Early Birds, Night Owls, and Tireless/Recurring Itinerants: An Exploratory Analysis of Extreme Transit Behaviors in Beijing, China," arXiv.org, http://arxiv.org/abs/1502.02056, 2015.
- [19] C. Dalton, and J. Thatcher. "What does a critical data studies look like, and why do we care? Seven points for a critical approach to 'big data'," http://societyandspace.com/material/commentaries/craig-dalton-and-jimthatcher-what-does-a-critical-data-studies-look-like-and-why-do-we-careseven-points-for-a-critical-approach-to-big-data/
- [20] M. Kavouras, and M. Kokla, Theories of geographic concepts : ontological approaches to semantic integration. Boca Raton: CRC Press, 2008.
- [21] V. R. Benjamins, D. Fensel, and A. Gomez Perez, "Knowledge Management through Ontologies," in Second International Conference on Practical Aspects Knowledge Management, Basel, Switzerland, 1998.
- [22] F. Anvari, and H. Tran, "Persona Ontology for User Centred Design Professionals," in International Conference on Information Management and Evaluation, Ho Chi Minh City, Vietnam, 2013, pp. 35-44.