Route Planning Using Linked Open Data

Pieter Colpaert

Ghent University - iMinds
Department of Electronics and Information Systems, Multimedia Lab
Gaston Crommenlaan 8, Bus 201
9050 Ledeberg-Ghent, Belgium
pieter.colpaert@ugent.be

Abstract. Intermodal route planners need to be provided with a lot of data from various sources: geographical data, speed limits, road blocks, time schedules, real-time vehicle locations, etc. These datasets need to be interoperable world-wide. Today, a lot of data integration needs to be done before this data can be be reused. Route planning becomes a data problem rather than a mathematical problem. Can the Web act as a global distributed dataspace for transport data? Could introducing Linked Open Data to this field make the data quality raise?

Keywords: #eswcphd2014Colpaert, Linked Open Data, Semantic Web, intermodal route planning.

1 Introduction

Intermodal route planning is a term used for planners which can advise the end-user to use multiple modes to get from one point to another. A transport mode is a type of transport, for example a train, tram, bus, car or bicycle. The amount of data that can be used to extract information from, is infinite. For instance, trying to answer the question "how long do I have to walk from one point to another?" can take into account the geolocation of the streets, the weather conditions at that time of the day, the steepness of the road, whether or not there is a sidewalk, criminality reports to check whether it is safe to walk through these streets, the accessibility of the road for e.g., wheelchairs or blind people, whether the street is blocked by works at that time, etc. We can imagine the complexities that arise if the user does not only want to walk, but that he also wants to get advice taking different transport modes into account. Advising an end-user can use an infinite amount of data that remains relevant for the problem. An *open world* approach is needed: a certain pool of data should be queried with the assumption that there is more data outside of this pool that may be relevant to the question.

Data quality, as defined by Orr et al. [14], is the measure of the agreement between the data views presented by an information system and that same data in the real world. When there is no agreement at all, the data quality is 0%, if it complies completely, it equals 100%. Orr et al. described in 1998 how the Feedback-Control System (FCS) (cfr. Fig. 1) affects data quality. This system describes a data life cycle: data is made available for reuse (A), the reuse is stimulated and supported (B), means to provide feedback are in place (C) and the feedback is also processed (D). The data quality increases at the speed of the slowest link in this process.

V. Presutti et al. (Eds.): ESWC 2014, LNCS 8465, pp. 827–833, 2014.

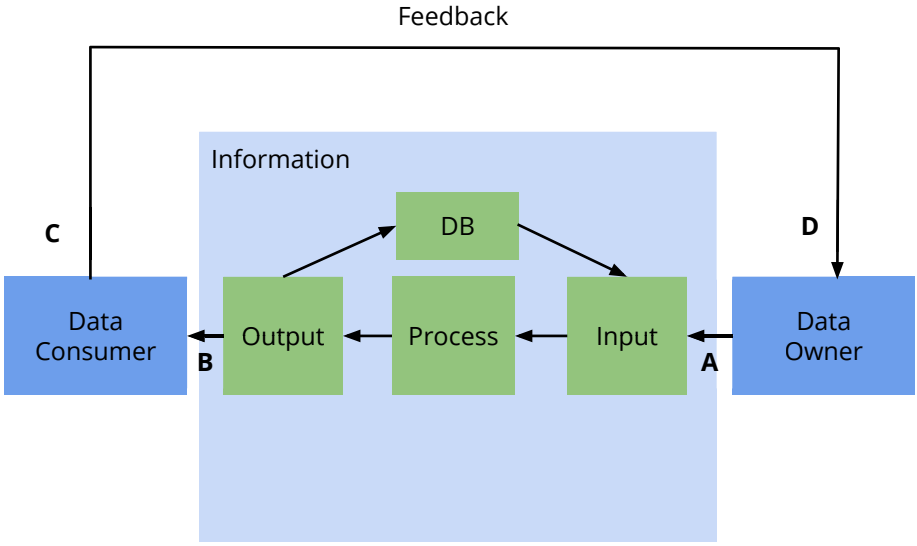

Fig. 1. The Feedback-Control System applied to data.

The trend towards Open Data is stimulating the reuse of the data. Can *Linked Open Data* indeed be used to raise the quality of transport data?

2 State of the Art

In The Netherlands in 1991, the mathematical problem of intermodal route planning has already been solved for data dumps of the Dutch railway system [16]. After that, quite a few projects came to exist, implementing intermodal route planning in their own way. The first section gives a concise overview of the state of the art of intermodal route planning systems. After discussing a couple of intermodal route planners and two algorithms for searching time tables, the state of the art of Linked Open Data is discussed. Finally, a couple of today's vocabularies used to publish or exchange transport data are discussed.

2.1 Intermodal Route Planners

Current intermodal route planners such as Open Trip Planner, Rapid Round based Routing [5] or Navitia.io use a *closed world* approach: they make the assumption that the data on the machine is complete and represents the real world perfectly. Both Open Trip Planner and Navitia.io provide a Web service to get answers to the question which routes should be followed to go from one point to another Rapid Round based Routing [5] is a C++-library which is memory efficient and has been used both from clientside as from serverside. The investment needed to add new datasets (both a new mode of transport or a new region) are high as manual intervention is needed to load the datasets in the store.

2.2 Route Planning Algorithms

For all kinds of route planning problems, Dijkstra is a popular solution. It can solve shortest path problems in a graph in $O(E + V \log V)$ with E the number of edges and V the number of vertices. It can easily be optimized by using heuristics (such as with the A* search algorithm) or the graph can be pruned while running the algorithm.

Route planning systems for public transit mostly use the RAPTOR algorithm [7]. All three route planners mentioned above are three different implementations of this algorithm. It is written to exploit the inherent structure of transport networks by operating in rounds and processing each route[1] of the network instead of popular graph based solutions based on Dijkstra.

A more recent approach is called the Connection Scan Algorithm (CSA) [8]. Just like RAPTOR, it is not graph-based. It is however not centered around routes, but around connections[2], which might be interesting as less data needs to be fetched per request. CSA is a very recent algorithm where parallel extensions, just like RAPTOR, seem promising [8].

2.3 Linked Open Data

Globally, data owners are slowly taking the decision to publish their data as Open Data. One way to go forward with Open Data is Linked Open Data (LOD), which uses RDF to structure the data and which uses the Web as a distributed dataspace. The global dataspace that comes to exist is called the LOD cloud.

There are tools available to raise the awareness of data owners. For example, "The 5 stars of Linked Open Data"[3] summarize the focus points of publishing Linked Open Data to the Web.

There are various steps to publishing Linked Open Data: data needs to be made discoverable, a suiting license need to be agreed upon, obstacles (such as privacy issues or copyright holder discussions) need to be overcome, etc. Various projects have introduced Linked Data Life Cycles [15,19,3,10]. These cycles describe the process to create and maintain Linked Open Data. More recently, also best practices to publish linked data on the web have been published by the W3C[4].

When the data is published, data is not per se discoverable. Data about the data (meta-data) needs to be published alongside the data. To make data discoverable for humans as an organisation, there is data portal software available, such as CKAN. To make data discoverable for machines, ontologies such as VoID [2] and DCAT [11] are available. There are various public Open Data Portals where datasets can be added, such as *http://datahub.io*.

While the Linked Open Data technology is available, there is however no maintained Linked Open Transport Data to be found that is published by transport agencies, at the time of writing.

[1] A route is a list of stop points a certain vehicle follows (e.g., a bus line).
[2] A connection in this context is a link to a next stop point.
[3] `http://5stardata.info`
[4] `http://www.w3.org/TR/2014/NOTE-ld-bp-20140109/`

2.4 Vocabularies

The General Transit Feed Specification (GTFS) for public transit[5] has been developed by Brian Ferris at Google for Google Maps. At the moment of writing, GTFS is the de facto standard for data dumps on public transit. In 2011, Ian Davis transformed this specification to an ontology, which is defined at *http://vocab.org/transit/terms/*.

Other specifications in the Open Transport field are for instance the Transmodel specification for public transport [4], SIRI for real-time public transit data exchange, Open511 for traffic events or DATEX for exchange of road transport data.

In 2009, the UK transformed 3 transport datasets towards RDF using the National Public Transport Access Network (NaPTAN) vocabulary. It is greatly inspired by Transmodel. NaPTAN defines the geographical hierarchy of the UK, defines the difference between a stop point (a place where a vehicle stops) and a stop area (a collection of stop points). It also defines different types of identification mechanisms for these stop areas and stop points and defines different types of stop points in the UK. The resulted dataset can still be queried at *http://transport.data.gov.uk/*.

3 Problem Statement and Contribution

Intermodal route planning is a data problem rather than a mathematical problem. Different datasets need to share the same identifiers (or make sure their identifiers are interoperable), need to be able to be queried nearly in real-time, need to be able to process machine readable feedback (as no dataset represents reality 100% correctly), need to track provenance, need to be made discoverable for both humans and machines, etc. These needs can all be put at a letter in Fig. 1:

- A: publishing data
- B: reusing data
- C: providing feedback
- D: processing feedback

The main contribution of this thesis is bringing together the field of intermodal route planning and Linked Open Data. The added value for data publishers researched is raising data quality. When Linked Open Data is introduced to intermodal route planning, algorithms need to be adapted to work with this technology.

4 Methodology

In order to achieve the best possible results within the available resources and timeframe, we suggest the following projects to be built: a high-level Open Transport vocabulary and an Open Transport Data Portal to make transport data discoverable and to stimulate a discussion on used vocabularies, tool chain, data reuse, etc. A third project will work together with the community at the Open Transport Data Portal to build a

[5] This thesis focuses on Open Transport, which also takes into account traffic, road signs, parking lots, taxis, bicycle routes, etc.

referential database for stop identifiers. A fourth project are various implementation of data publishing mechanisms (such as SPARQL). A last project is a proof of concept in which the Web will be used as a global distributed database for an intermodal route planning algorithm.

Future development on vocabulary will be carried out by the Open Transport community of the Open Knowledge Foundation. The methodology used for creating this vocabulary is inspired on "Process and methodology for developing semantic agreements" by the JoinUp project [1].

5 Preliminary Results

5.1 Algorithms

Dijkstra and its optimizations are to be avoided in most cases as it is hard to parallelize [7,12,13], yet it can be considered for subproblems.

For public transport, RAPTOR and CSA seem promising. Both algorithms can be considered, depending on what data each service is going to provide. For RAPTOR, starting at a certain stop, all routes are needed which leave next at that stop. For CSA, starting at a certain stop, only the next stop is needed for all next departures in that stop. Both algorithms are very promising for parallelization. Further research will show what algorithm is best in this use case.

5.2 Open Data

The 5 stars of Open Data Portals A paper has been written for the MeTTeG2013 conference [6], a conference for e-government, which summarizes five focus points for an Open Data Portal. The paper was written in a interdisciplinary setting with the communications department of Ghent University.

Open Transport. At the Open Knowledge Foundation (OKF), I have started coordinating the Open Transport Working Group[6]. The focus of this working group are three projects:

1. The Open Transport Vocabulary[7] creates a minimal representation of all concepts that are needed for intermodal route planning.
2. Stations.io[8]: a knowledge base to link identifiers for stop areas and stop points. The project links various sources together by creating owl:sameAs links, links identifiers to their concepts: stop areas or stop points and creates links between the different stop points and stop areas.
3. The Open Transport Data Portal[9] makes data discoverable, will analyze the added datasets for inconsistencies, will validate the datasets, will host conversations around transport data reuse.

[6] `http://transport.okfn.org`
[7] `http://github.com/opentransport/vocabulary`
[8] `http://stations.io`
[9] `http://transport.datahub.io`

Towards Linked Open Data The DataTank [17] is a data adapter. It takes a data source, such as a CSV file, JSON file, a web-service, a website, a database, etc. as an input and transforms it into a RESTful interface. The DataTank also supports mapping these sources towards RDF using a mapping language. The mapping language that will be supported in the next Long Term Support release of The DataTank is an extension of R2RML, called RML [9].

Using The DataTank we have create a RESTful interface to access transport data in Belgium called iRail. This interface can be found at *http://data.iRail.be/*.

Distributed Version Control for triples. R&Wbase [18] has been presented at the WWW2013 conference at the Linked Data On the Web (LDOW) workshop. It adds a version to each triple in a triple store which supports named graphs.

6 Evaluation Plan

As the thesis is still in a very early stage, the evaluation plan is still very vague. Data quality, defined as the agreement of the data with the real-world, needs to be assessed before applying Linked Open Data technologies and after. In order to get results, the FCS model will be used (see Fig. 1). Data quality will be derived from: the amount of data reuse, the amount of data feedback, the amount of data feedback that is being processed and the amount of data about a certain *real-world object* that is published. When these four factors increase, the data quality will increase [14] and the thesis will be validated.

While the data quality should increase, the intermodal route planning algorithms should remain fully functional without regressions in response time. Different architectures will be set-up which reuse the linked datasets and feed the right data in the algorithms.

7 Conclusion

This paper's contribution is to bring together the field of intermodal route planning with Linked Open Data. Linked Open Data tools are applied to data problems with intermodal route planning. Research focuses on raising the data quality. Four criteria need to be assessed – data reuse, data feedback, processing data feedback and publishing data –. Further research as part of this thesis shows whether or not the Web can then be used as a global distributed database for intermodal route planners.

Acknowledgements. This thesis is supervised by prof. Rik Van de Walle and co-supervised by dr. Erik Mannens and dr. Ruben Verborgh. The research activities described in this paper were funded by Ghent University, iMinds, the Flemish department of Economics, Science and Innovation (EWI), the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT) and the European Union.

References

1. Process and methodology for developing semantic agreements (June 2013)
2. Alexander, K., Cyganiak, R., Hausenblas, M., Zhao, J.: Describing linked datasets with the VoID vocabulary. W3C Interest Group Note (March 2011)
3. Auer, S., et al.: Managing the Life-Cycle of Linked Data with the LOD2 Stack. In: Cudré-Mauroux, P., et al. (eds.) ISWC 2012, Part II. LNCS, vol. 7650, pp. 1–16. Springer, Heidelberg (2012)
4. Bourée, K., Staub, L., van der Peet, G.: The European reference data model for public transport: Transmodel as a basis for integrated public transport information systems. In: Towards an Intelligent Transport System (1994)
5. Byrd, A., Koch, T., de Konink, S.: Rapid round-based routing (2013)
6. Colpaert, P., Joye, S., Mechant, P., Mannens, E., de Walle, R.V.: The 5 stars of open data portals. In: Conference on Methodologies, Technologies and Tools enabling e-Government 2013 Conference Proceedings (2013)
7. Delling, D., Pajor, T., Werneck, R.F.F.: Round-based public transit routing (2012)
8. Dibbelt, J., Pajor, T., Strasser, B., Wagner, D.: Intriguingly simple and fast transit routing. In: Bonifaci, V., Demetrescu, C., Marchetti-Spaccamela, A. (eds.) SEA 2013. LNCS, vol. 7933, pp. 43–54. Springer, Heidelberg (2013)
9. Dimou, A., Van der Sande, M., Colpaert, P., Mannens, E., Van de Walle, R.: Extending R2RML to a source-independent mapping language for RDF
10. Hyland, B., Wood, D.: The joy of data - a cookbook for publishing linked government data on the web. In: Wood, D. (ed.) Linking Government Data, pp. 3–26. Springer, New York (2011)
11. Maali, F., Erickson, J.: Data Catalog Vocabulary (DCAT). W3C Working Draft (August 2013)
12. Madduri, K., Bader, D.A., Berry, J.W., Crobak, J.R.: Parallel shortest path algorithms for solving large-scale instances (2006)
13. Meyer, U., Sanders, P.: δ-stepping: a parallelizable shortest path algorithm. Journal of Algorithms 49(1), 114–152 (2003)
14. Orr, K.: Data quality and systems theory. Communications of the ACM 41(2), 66–71 (1998)
15. Scharffe, F., Bihanic, L., Képéklian, G., Atemezing, G., Troncy, R., Cotton, F., Gandon, F., Villata, S., Euzenat, J., Fan, Z., Bucher, B., Hamdi, F., Vandenbussche, P.-Y., Vatant, B.: Enabling linked data publication with the datalift platform (2012)
16. Tulp, E., Siklssy, L.: Searching time-table networks. Artificial Intelligence for Engineering, Design, Analysis and Manufacturing 5, 189–198 (1991)
17. Vander Sande, M., Colpaert, P., Van Deursen, D., Mannens, E., Van de Walle, R.: The datatank: an open data adapter with semantic output
18. Vander Sande, M., Colpaert, P., Verborgh, R., Coppens, S., Mannens, E., Van de Walle, R.: R&Wbase: Git for triples. In: Proceedings of the 6th Workshop on Linked Data on the Web (2013)
19. Villazón-Terrazas, B., Vilches-Blázquez, L., Corcho, O., Gómez-Pérez, A.: Methodological guidelines for publishing government linked data. In: Wood, D. (ed.) Linking Government Data, pp. 27–49. Springer, New York (2011)