ONETT: Systematic Knowledge Graph Generation for National Access Points

David Chaves-Fraga[1], Adolfo Antón[1], Jhon Toledo[1], and Oscar Corcho[1]

Ontology Engineering Group, Universidad Politécnica de Madrid, Spain

Abstract. In this paper, we describe our implemented approach for the usage and exploitation of declarative mappings for the publication of open transport data from transport authorities and operators into an ontology based on Transmodel. This allows a homogeneous representation of transport data across EU transport-related organisations and minimises the need to understand ad-hoc heterogeneous representation formats for transport data as currently published by them. We show how we create and use RML mappings for the specific case of transforming GTFS data into a Transmodel-based ontology. In the future, such data may be further transformed into other formats such as NeTEx.

Keywords: Transmodel · GTFS · NAP · RML

1 Introduction

Transport data is being currently published by transport authorities and operators in many different formats, some of which are well-known de-facto standards, such as the General Transit Feed Specification or GTFS, and some others are adhoc data formats whose structure is decided by the data publisher (e.g., current datasets and APIs published by Empresa Municipal de Transportes de Madrid in its open data portal[1], tram information in Zaragoza[2], etc.)

All of these datasets have similarities, associated to the fact that they are describing overlapping sets of information (schedules, stops, vehicles, lines, etc.). They are also made available, commonly, using tabular data formats. For example, GTFS feeds are essentially zip-compressed files containing sets of CSV files following the GTFS specification. And other data sources such as those mentioned above as examples provide the data either in CSV or JSON.

Having all this data available in a homogeneous manner would actually reduce the total cost of reusing data sources, especially across operators/authorities and cities/regions. That is, developers may be able to develop one application that would be deployable in any city in the world with minor adaptations. This is already happening with GTFS, which is not only being used by Google Maps to provide data about transport infrastructure, but also for route planning, but also by other route planners, such as Navita.io and OpenTripPlanner.

To achieve this homogeneity, there are several options that may be followed:

[1] https://opendata.emtmadrid.es/

[2] https://www.zaragoza.es/sede/servicio/catalogo/327

- Transport authorities and operators may agree on using the same data format and hence publish according to such data format. They know well the type of data that they handle, the quality properties on such data, etc., so they should be able to provide this data easily. To some extent, this is what is happening currently with GTFS, and what should happen in the near future in the European Union with NeTex, according to directive 2010/40/EU and regulation 2017/1926 (MMTIS).
- 3rd parties (as well as operators and authorities themselves) may be able to create transformation rules that allow transforming the original data sources into other generally-agreed formats, republishing such transformed data either in the original data portals, if allowed to do so, or in other servers. Transformations may be done programmatically (that is, with ad-hoc code) or declaratively (using mappings in existing languages like R2RML [2] or RML [3]).

In this paper, we present our work on ensuring that declarative mappings can be used for the purpose of transforming transport data published by transport authorities and operators into a homogeneous representation based on Transmodel (the reference data model for public transport at European level, which will be further described in section 2). This data can then be further transformed into NeTEx so as to comply with the EU regulations for the publication of transport-related data in National Access Points.

2 Transmodel Ontology and GTFS

In its drive to foster interoperability across Europe, the EU is requiring each Member State to allow access to transportation data via a National Access Point (NAP). According to the EU Regulation 2017/1926, all transportation authorities, transport operators and infrastructure managers must provide static and dynamic data in specific data formats (e.g., NeTEx, SIRI). - the EU Regulation applies to different transportation modes, including air, train, road vehicle, bus, ferry, metro, tram, shuttlebus, car-sharing, car-pooling and bike-sharing.

Transmodel is the European Reference Data Model for Public Transport. It provides a conceptual model of common public transport concepts and data structures that can be used to build many different kinds of public transport information system such as timetabling, fares, operational management, real-time data, journey planning. It is divided into eight different sections or Parts: Common Concepts (CC), Public Transport Network Topology (NT), Network Description (ND), Operations Monitoring & Control (OM), Fare Management (FM), Passenger Information (PI), Driver Management (DM), Management Information & Statistics (MI).

These parts or sections are usually developed by different standards or specific data formats. One of the most relevant implementations is NeTEx, which covers partially some features of the parts CC, NT, ND, FM and PI. NeTEx releases the 2017/1926 EU Regulation (May 2017) where the European Commission recognized NeTEx as a strategic standard for the cross-border exchange of data. The first step must be taken before December 2019 when every European country must provide data available in NeTEx format at National Access Points to allow EU-wide multi-modal travel information services.

The General Transit Feed Specification (GTFS) is a de-facto standard for representing public transport data, a collection of at least five required, two optional required and up to fifteen CSV files (with extension .txt and preferably encoded as UTF-8) contained within a compressed file to describe a transit scheduled operations system. The aim of GTFS is providing at least trip-planning functionality. It defines the headers and a set of rules that must be taken into account when the dataset is created. Each file, as well as its headers, can be mandatory or optional and they have relations among them. The specification supports the representation of several public transport features such as trips, routes, stops, times, fares or calendar.

In order to provide a better GTFS to NeTEx conversion and further full data interoperability, we start to build up a Transmodel Ontology. The development is released in a github repository[3] where every material generated is upload about the different activities carried out during the development (i.e., use cases, user stories, glossary of terms, etc.). Based on the Transmodel base URI proposed by the CEN Transmodel working group model[4] and its documentation[5] we develop the corresponding ontology following the NeOn methodology[6]. Before performing the transformation from GTFS to the ontology based format of Transmodel, we analyse the relationship between the two standards. For example, in Table 1 we show the relation between the properties of *Agency* in the GTFS model with the corresponding property in Transmodel (*Authority*). The relation among all the resources and properties of GTFS model and Transmodel is available online[6].

Table 1. Example of relation among GTFS properties and Transmodel Ontology

GTFS	Transmodel (Ontology)
Agency_name	https://w3id.org/transmodel/terms#authorityName
Agency_url	https://w3id.org/transmodel/terms#agencyUrl
Agency_timeZone	https://w3id.org/transmodel/terms#authorityTimezone
Agency_lang	https://w3id.org/transmodel/terms#authorityLang

3 The ONETT Demo

The Open NEtwork of public Transport application (ONETT)[7] uses Semantic Web technologies to perform a systematic knowledge graph generation in the

[3] https://github.com/oeg-upm/transmodel-ontology

[4] https://w3id.org/transmodel/terms#

[5] http://www.transmodel-cen.eu/

[6] https://github.com/osoc-es/onett-paper/tree/master/Gtfs2Transmodel

[7] https://osoc-es.github.io/onett/

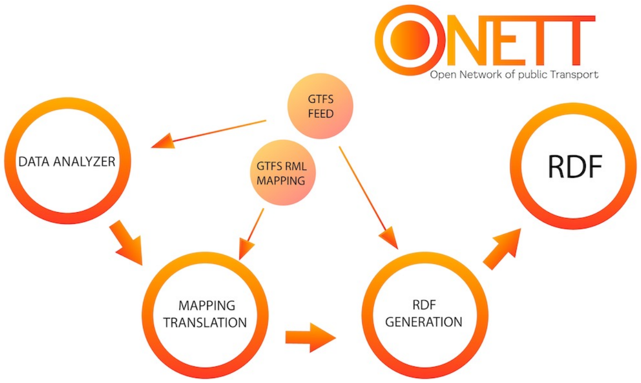

Fig. 1. ONETT. The ONETT workflow for the systematic generation of Knowledge Graph following Transmodel from GTFS feeds.

transport domain. More in detail, ONETT applies the concept of Ontology Based Data Access (OBDA) [5], which it aims at providing a unified view and common access to a set of data sources, using ontologies and mappings.

In this specific case, we create a general mapping between the full specification of GTFS[8] and ontology based Transmodel using the RML specification in its YARRRML [4] serialization. Before running the transformation, we have to perform a mapping translation [1] process to adapt the general mapping to the input data as it is not always going have the same structure and number of files due the naturalness of GTFS. Thanks to the simplicity of YARRRML serialization, the translation process is done in a efficient and simple manner. The workflow of the application is shown in Figure 1 where the SDM-RDFizer[9] engine for RML mappings is integrated in the application to perform the transformations of the input data in CSV to RDF. More in detail, the steps following by ONETT for generating the desirable RDF knowledge graph based on the Transmodel ontology from a GTFS feed are:

1. **Analyse the input data:** It decompresses and analyses the input GTFS feed to understand the files and the structure of each file (headers).
2. **Mapping translation:** It takes the general GTFS YARRRML mapping that represents the full specification and generates a new mapping corresponding to the input data.
3. **Knowledge Graph Generation:** It runs the SDM-RDFizer engine to transform the raw data to RDF.

These steps are a black box for the transport authorities that want to obtain the knowledge graph from their GTFS feeds. Using the web application the user only has to upload the compressed feed or provide a URL and ONETT

[8] https://github.com/osoc-es/onett-back/blob/master/mapping/mapping/gtfs2transmodel.yml
[9] https://github.com/SDM-TIB/SDM-RDFizer

generates automatically the corresponding knowledge graph. With this approach, we provide a useful tool to generate National Access Point complaint data from a de-facto standard and very popular data format in a systematic manner.

4 Conclusions and Future Work

The availability of homogeneous transport data from worldwide transport authorities and operators gives us the possibility of creating new types of applications related to transport (trip planners, fare calculators, ticket recommenders, etc.) that can be deployed easily in different regions or cities. In this paper, we have shown our approach to create such homogeneous transport data based on declarative mappings that can be used to generate transport knowledge graphs for any region or city in the world that is currently publishing data in GTFS. The mappings allow transforming GTFS data into RDF according to a TransModel-based ontology. Such data can be queried in a homogeneous manner so that the aforementioned applications can be created more easily.

Acknowledgements

This work is partially supported by EIT Digital under Grant Agreement "No. EIT/ EIT DIGITAL/SGA2019/1 through action: SNAP" and by the Spanish Ministerio de Economa, Industria y Competitividad and EU FEDER funds under DATOS 4.0: RETOS Y SOLUCIONES - UPM Spanish national project (TIN2016-78011-C4-4-R) and by an FPI grant (BES-2017-082511). Thank you to our open Summer of code 2019[10] students: Luis Pozo, Pablo Castellanos, Marta Retana.

References

1. Corcho, O., Priyatna, F., Chaves-Fraga, D.: Towards a New Generation of Ontology Based Data Access. In: Semantic Web Journal (2019)
2. Das, S., Sundara, S., Cyganiak, R.: R2RML: RDB to RDF Mapping Language, W3C Recommendation 27 September 2012. www.w3.org/TR/r2rml (2012)
3. Dimou, A., Vander Sande, M., Colpaert, P., Verborgh, R., Mannens, E., Van de Walle, R.: RML: A Generic Language for Integrated RDF Mappings of Heterogeneous Data. In: LDOW (2014)
4. Heyvaert, P., De Meester, B., Dimou, A., Verborgh, R.: Declarative Rules for Linked Data Generation at your Fingertips! In: Proceedings of the 15^{th} ESWC: Posters and Demos (2018)
5. Poggi, A., Lembo, D., Calvanese, D., De Giacomo, G., Lenzerini, M., Rosati, R.: Linking data to ontologies. In: Journal on data semantics X, pp. 133–173. Springer (2008)
6. Suárez-Figueroa, M.C., Gómez-Pérez, A., Fernández-López, M.: The neon methodology for ontology engineering. In: Ontology engineering in a networked world, pp. 9–34. Springer (2012)

[10] `https://2019.summerofcode.es/`