Integrating Public Transportation Data: Creation and Editing of GTFS Data

Mário Braga, Maribel Yasmina Santos and Adriano Moreira

ALGORITMI Research Center, University of Minho
Guimarães, Portugal,
{mario.braga, maribel.santos, adriano.moreira}@algoritmi.uminho.pt

Abstract. The current state of standardization related to representation and exchange of data about public transportation systems is still at its infancy, which leads to severe interoperability issues in projects that depend on the data from several diverse sources. In many cases, the interoperability issues arise from the use of rudimentary information systems, or even paper-based procedures, to manage operational data such as schedules and tariffs. In these cases, exchanging data with external systems is very difficult. This paper describes the development of a web-based application aiming to simplify the creation and editing of public transportation data that could be easily exchanged in a normalized format. This description is preceded by a discussion about a data model that could ease data interoperability. Here, the GTFS reference, with some adjustments, is used as a guideline for the definition of such transportation data model.

Keywords: GTFS, GIS-T, Interoperability, Transportation

1 Introduction

The growing migration to urban areas leads is leading to the increase of transportation demands, namely parking space, energetic consumption, ambient pollution and traffic congestion, in the transportation of people or goods. Public Transportation Systems (PTS) are often used to mitigate the impact of the aforementioned issues. Through the interaction with PTS, users may benefit from a more efficient, low cost and sustainable journey [1–3]. Services provided by PTS can be an added-value to the users, when enhanced with data from different public transportation authorities, and when combined with other information like weather or road congestion.

PTS for buses, subways, trains and other public vehicles are currently available in many cities and countries. These systems, usually complex, combine concepts and technologies from Geographic Information Systems (GIS) and Transportation Information Systems in what is usually known as Geographic Information Systems for Transportation (GIS-T)[4]. While the domain of GIS may be considered to be in a mature state, the Transportation Information Systems

are still moving towards standardization. The current state in the transportation systems or in their processes slow down projects like the TICE.Mobilidade[1] where data exchange is so critical.

The TICE.Mobilidade project, a large Portuguese project integrating more than 25 partners, aims to develop a loosely-coupled PTS that provides applications and services that allow users to reduce the costs of their daily travels and promotes the use of PTS. Although this platform seems similar to others PTS, there are some characteristics that distinguishes it from other systems. The core component in the project is a platform that then aggregates data from different types of public transportation sources and provides it to the software development community so they can develop their own applications. This core platform, named One.Stop.Transport (OST), faces many challenges: it has to provide mechanisms to import data from different transportation authorities; store the fed data under an homogeneous data model and; provide mechanisms that allow the development and deployment of applications developed by third party entities. Although all the aforementioned challenges have a critical impact to the success of the OST platform, this paper will only focus on the approaches conducted to develop mechanisms that help the interoperability between the OST platform and public transportation authorities.

As will be described in the remaining of this paper, the OST platform adopted the General Transit Feed Specification (GTFS) as a fed protocol, and this decision must be complemented with the development of tools that help public authorities to manage their data. This paper presents a data model capable of storing and supporting the edition of multiple GTFS feeds, and also a web-based application that makes use of that data model to store the gradual creation of GTFS data.

This paper is organized as follows. Section 2 presents the GTFS model. Section 3 describes the OST platform and its main components. Section 4 presents the proposed extension to the GTFS model, while section 5 introduces the proposed editor application. Section 5 concludes with some remarks about the presented work.

2   General Transit Feed Specification

Many transportation authorities use legacy data models that have little or even no preoccupation with interoperability with other systems. Interoperability is now a crucial area of study in PTS and many efforts have been conducted towards the development of standards. In 2006, Google launched a transit service that merges public transit information and maps in order to provide multi-modal trip planning. Users can easily use this service by picking origin and destination locations in a map and having, as a result, a multi-modal travel itinerary including time, price and trajectory. While the service represents a trivial PTS functionality, the way how Google feeds this service soon became a *de facto* standard.

---

[1]http://tice.mobilidade.ipn.pt/

Looking towards data exchanging and interoperability, Google provides a reference format named General Transit Feed Specification (GTFS) through which public or private transportation authorities can exchanges their data with the Google services. The GTFS rapidly increased its popularity[2] leading many public authorities (which use GTFS format in the Google Transit service) to release their transit information for third-party developers [5]. As a consequence, many transit applications now support the exchange of data in GTFS format. As an example, the Travel Assistance Device, a mobile application that alerts the users when the destination stop is getting near, uses GTFS as the input format to load the stops and schedules [6]. The Graphserver, an open-source multi-modal trip planer, supports the GTFS format in order to load public transportation data to its multi-modal algorithm. The GTFS OpenStreetMap Sync [7] and the Open-TripPlanner[3], both open-source software systems that synchronize GTFS with OpenStreetMaps, allow users to plan a trip that can combine multiple means of transportation. This is also observed in other domains, for instance in intelligent mobile advertising applications or context-aware systems, which use GTFS data in order to display specific advertising [8] or even in the assessment/benchmark of public transportation systems [9].

In the initial version, GTFS only supported the exchange of static data for schedules, stops, agencies, trips, fares and zones, but in order to cover the demand for real-time information, Google launched in 2011 an extension named GTFS real-time[4] that can be used to make real-time information available to the users. With the availability of real-time data, public transportation users spend less time waiting, feel more safe and likely to use public transportation services [10]. This new feature reinforced GTFS as an exchange reference in order to become, in fact, a standard for public authorities to exchange their data.

2.1 Taxonomy

The GTFS format is composed of thirteen entries hold in Comma-separated values (CSV) files in which six of them are mandatory and the remaining are optional. The six mandatory entries hold the information about the agency name, contacts and language (Agency.csv); route short and long name, descriptions, color and type (Routes.csv); trip short name, headsign and direction (Trips.csv); stop name, latitude and longitude and other information like support for wheelchair boarding (Stops.csv); all the information related to the schedules like arrival and departure time (Stop_times.csv) and service days when the trips occur (Calendars.csv). The optional entries increase the value of the public transportation data by adding information about the feed itself (Feed_info.csv); attributes and rules to apply a fare (Fare_attributes.csv and Fare_rules.csv); data that define the trajectory of a trip (Shapes.csv); data that define connections between routes (Transfers.csv); an entry to define exception dates when a service

---

[2]A total of 703 feeds in 2013.

[3]http://opentripplanner.com/

[4]https://developers.google.com/transit/gtfs-realtime/

is explicitly active or inactive motivated by a certain event (Calendar_dates.csv) and; an entry (Frequencies.csv) that represents the time between trips (departure at first stop and arrival at last stop).

Although GTFS have been developed to be used as a sharing format, it can be conceptualized as a data model. All the entries of the GTFS clearly define all the relations and proprieties needed to construct an Entity-Relation (ER) diagram. As shown in Figure 1, this approach supported the development of a model with the mentioned entities and a new, named, Zones that holds the ID that links stops to a certain fare rule.
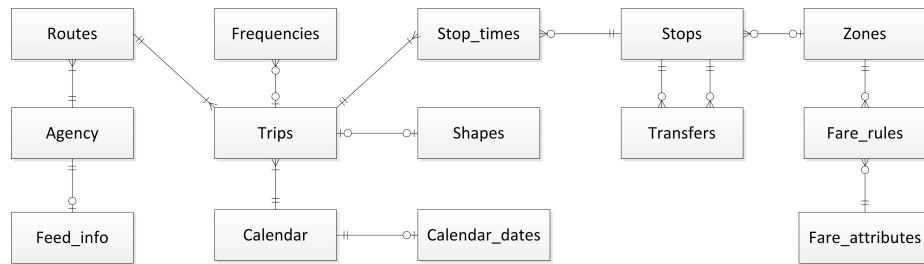
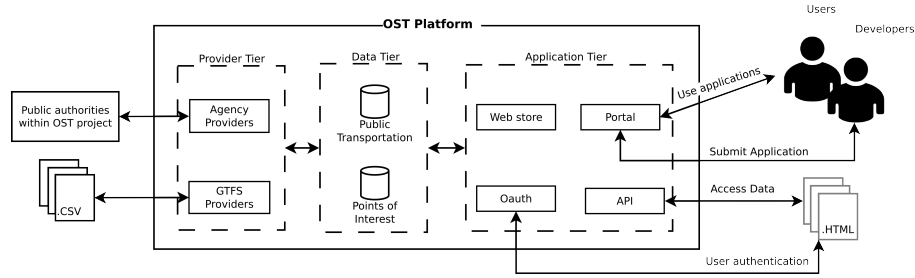**Fig. 1.** GTFS Entity–Relationship model

2.2 Tools

For public transportation authorities to validate their own GTFS data, Google provides the *feedvalidator* and *schedule_viewer* applications. The *feedvalidator* inspects the syntax of the feed and ensures that the files and the content match the specifications defined in the GTFS reference. The *scheduler_viewer* application helps users to see the data in a different way providing stop locations and shapes over a map, filter trips or services that occur in a certain date, or provide access to the departures at a certain stop.

Notwithstanding, both applications make use of GTFS-ready data. In order to create new data, public transportation authorities have to make use of other non-official tools or develop their own tool. Currently, there are some alternatives in this matter, as there are some tools[5] available for a fee, and there are open-source projects[6,7] that can be deployed and personalized in such a way that public authorities can create or modify their GTFS files. Yet, there are few applications that allow the creation of GTFS files in a smooth way.

---

[5]http://www.transiteditor.com/
[6]https://github.com/OneBusAway/onebusaway-gtfs-modules/wiki
[7]http://code.google.com/p/transitdatafeeder/

Fig. 2. OST Platform architecture

3   The OST Platform

The OST platform intends to provide a sustainable environment where users themselves may develop applications to fulfill their needs or expectations. To achieve so, the platform provides ways to import and aggregate data, and also provides data that users can use in their own web-based applications.

The OST platform relies on a loosely-coupled architecture composed by three tiers (see Figure 2). The Provider Tier contains services that load data into the OST platform. Here, the platform supports two different approaches. The first approach connects directly to transportation authorities that take part in the project. This approach showed to demand specific modules specifically crafted to each authority and very difficult to reuse in other authorities. In order to reach other authorities, an alternative approach was followed by using the GTFS reference as a feed protocol to import public transportation data. In this second approach, the transportation authorities are responsible for the creation and submission of the GTFS files to the OST platform.

The Data Tier holds the data models that store data for public transportation and other types of data that enhance the OST platform, such as Points of Interest. As to the last tier, named Application, it provides the interface to interact with other applications/systems or users. The Application tier is decomposable into four main components: a set of services for providing the data so that the software development community can proceed with the development of web-based applications (API); Services for authentication (Oauth); a main web application (Portal) that joins additional information in order to help the community to develop new applications and a place where users can view/access to the already deployed applications and; a Store that handle the hosting of web-based applications. The Store allows the submission of two different types of applications. In one hand, it allows the deployment of applications that will be hosted by the OST platform. On the other hand, it is possible to submit applications that are hosted outside of the OST platform. In this later case, there is no deployment under the OST platform, the only thing that has to be submitted to the Store is the URI to the application. Despite all efforts conducted by the project to improve interoperability with transportation authorities, the current state of the public transportation information systems, in particular the lack of

standardization, incapacitates the authorities to provide fully compliant GTFS files to the platform. The major issues are the inconsistency in the models, the quality of the data itself and, in some cases, the absence of mandatory information. To overcome these limitations, it was necessary to create a mechanism that allows the gradual creation, editing and storing of GTFS data up to the point when data completely fulfills the platform's needs. Since the OST platform integrates a set of services for: a) promoting web applications (Store); b) providing authentication mechanisms to OST users (Oauth) and; c) feeding the transportation model with GTFS files (GTFS providers), the adopted approach was to materialize this mechanism as a web-based application connected to a data model capable of storing the gradual creation of GTFS data.

4   GTFS Extension

As was described before, GTFS can be understood as an ER data model. However, because of its primary goal, the original conception of GTFS as a data model reveals some limitations when used in an application that integrates multiple data providers.

4.1   Primary Keys and Source IDs

GTFS is used as a format to allow single public transportation authorities to exchange their data. Theoretically, in this context, each provided ID is unique. However, different authorities can use the same IDs, making it difficult to join data from different providers into the same database. In other words, the data model can't use the provided IDs as primary/foreigner keys.

The issue with the original IDs is that they have substantial importance mainly because they are used as a reference to contextualize some information, for instance, many routes or services are known by their IDs and if they are not stored and showed to users, they may not understand the data that are provided. For instance, the route "1:Linha da Cidade" is known by route 1 and not by route "Linha da Cidade". To overcome this issue, a new attribute, named source_id, was added to the model to store the original ID of each record. This approach also assumes that in case of an update, the data provider can simply search by the original ID and proceed with the modification. Otherwise, it would be necessary to load again all the data.

4.2   Data Providers

The approach mentioned before was also used to detach data from different data providers. One of the requirements for a GTFS editor is that each user (data provider) can only access to its own data. By analyzing the GTFS ER model (see Figure 1), this could be done by using the Feed_info. The Feed_info holds all the information that describes the feed authority that fed the data and is directly linked to the Agency entity. By using this entity it is possible to

retrieve the agencies that the authority fed and backtrack all the remaining data that are related to that agency. Such an approach would work whenever all the data are linked and is possible to backtrack all data to the Feed_info. However, this entry is not mandatory and some public authorities may not even include the Feed_info entry in their feed. To overcome this problem, it was added a new attribute (user_id) associated to a new entity, which supports user authentication and additional information about the users.

4.3 Geographic Representation

Aiming to represent the geographic data in a more efficient way, the original representations used in GTFS to store geographic references were replaced by new formats. In its original structure, the geographic location of a stop is represented by two proprieties named latitude and longitude, and the trajectories to define a trip are represented by linking an ordered set of latitude and longitude attributes. Those attributes where replaced by the the creation of the Open Geospatial Consortium (OGC) Points and Linestrings fields, allowing the storage and processing of geographical objects. All the mentioned adaptations were applied to the previous ER and the obtained model is shown in Figure 3.
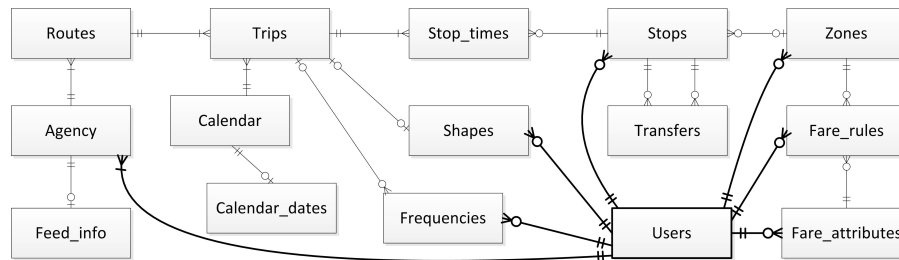

Fig. 3. Extended GTFS Entity–Relationship model

5 GTFS Editor Application

On top of the data model described in the previous section, a web-based application has been developed that enables users to access, edit and store their data in an interchangeable format. The application was conceptualized as a thin client application, in other words, it was divided in two different parts, one that resides in the server side and another that runs on the client side (see Figure 4). Although the editor emerged as an applications of the OST platform, it is not integrated in this platform. This way, users who use the Editor to store their data do not, automatically, make them available in the OST platform. The process of feeding the OST platform stays at the responsibility of the user itself who has to use the GTFS Editor in order to download their current GTFS data from

the Editor server and then upload them to the OST platform through the GTFS provider interface (as illustrated in Figure 2). In order to use the GTFS Editor web-application, users can access directly[8] or use the Portal[9] website on the the OST platform. Two types of authentication coexist, one with the application itself (for non-OST users), and or another one that uses the OST authentication services (Oauth).
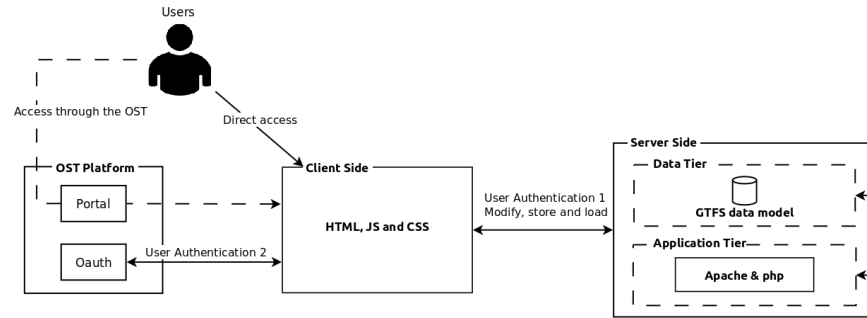
Fig. 4. Application architecture

After login, users can start the editing of data by loading their existing GTFS files (if any) or create new data. Each modification made to the data is stored into the server's database and, at any time or from any place, users can reload their data (Load last session) and resume the edition (Figure 5). The Editor tab (see Figure 5) allows users to enter the workspace where it is possible to create, view, modify and delete data. In the current version, the workspace contains five inner tabs that are related to the following GTFS entities: Agency, Route, Calendar, Stop, Shape and Trips. Here, the inner tab Trips clusters both Trips and Stop_Times entities. One of the concerns in the development process was to create a simple and yet user friendly interface. To achieve so, some processes integrate automatic mechanisms. For instance, the creation of schedules includes smart processes that automatically fill the stops based on the route and calendar information (Figure 6) and a function to estimate the arrival and departure times for a list of sequential stops[10]. In other cases, the application makes use of digital maps to assist the drawing or edition of geographic elements (Figure 7).

6 Conclusions

With this work, the GTFS reference has been used to inspire the development of a model that supports multiple GTFS feeds where the data are attached to users; support geographic data types and; save original IDs. The proposed model can also be used as a core data model to many GIS-T applications or services

---

[8]https://hera.dsi.uminho.pt/editorgtfs/web/

[9]https://www.ost.pt/

[10]The estimation is made by dividing the time of a trip by the total number of stops.
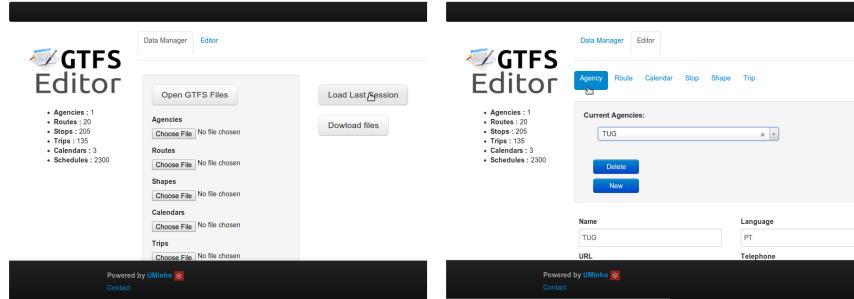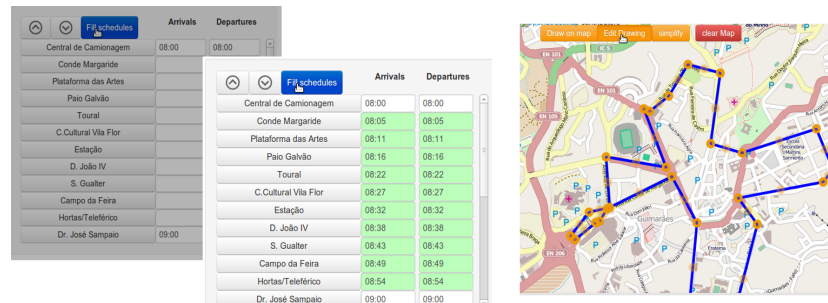
Fig. 5. Data Manager and Agency workspace tab

Fig. 6. Fill of stops taking as input the route and calendar

Fig. 7. Estimation of time arrivals/departures and edition of trips trajectories

that need to store public transportation information. This model can be seen as an extension of the GTFS reference, with little impact in the original GTFS format, so, the model can evolve to include GTFS real time feature without any compliance issue.

Aiming to facilitate the exchange of transportation data, a web-based editor has been developed. This application allowing users to create, store and edit their own GTFS files. In its current version, the editor supports the mandatory entries: Agency, Routes, Trips, Stop_times, Stops, Calendar and the optional entity Shapes. Future versions should support the remaining optional entities. The present work is crucial to the success of the OST platform since the majority

of transportation agencies still use proprietary data models and exchange formats. Most Portuguese public transportation authorities still avoid the adoption of normalized approaches, mainly because they are stuck in legacy systems or methodologies. The developed model and web-application can help those agencies to make progresses in its mechanisms for storing data but also in their current interoperability issues.

Acknowledgment

References

1. C. G. Chorus, E. J. Molin, and B. Van Wee, "Use and effects of advanced traveller information services (atis): a review of the literature," *Transport Reviews*, vol. 26, no. 2, pp. 127–149, 2006.
2. C. G. Chorus, J. L. Walker, and M. E. Ben-Akiva, "The value of travel information: A search-theoretic approach," *Journal of intelligent transportation systems*, vol. 14, no. 3, pp. 154–165, 2010.
3. T. F. Golob and A. C. Regan, "Impacts of information technology on personal travel and commercial vehicle operations: research challenges and opportunities," *Transportation Res. Part C: Emerging Technologies*, vol. 9, no. 2, pp. 87–121, 2001.
4. J.-C. Thill, "Geographic information systems for transportation in perspective," *Transportation Research Part C: Emerging Technologies*, vol. 8, no. 1, pp. 3–12, 2000.
5. B. Ferris, K. Watkins, and A. Borning, "OneBusAway: a transit traveller information system," in *International ICST Conference on Mobile Computing, Applications, and Services (Mobicase 2009)*, pp. 92–106, 2010.
6. S. J. Barbeau, P. L. Winters, N. L. Georggi, M. A. Labrador, and R. Perez, "Travel assistance device: utilising global positioning system-enabled mobile phones to aid transit riders with special needs," *Intelligent Transport Systems, IET*, vol. 4, no. 1, pp. 12–23, 2010.
7. K. Tran, S. Barbeau, E. Hillsman, and M. A. Labrador, "Go_synca framework to synchronize crowd-sourced mapping contributors from online communities and transit agency bus stop inventories," *International Journal of Intelligent Transportation Systems Research*, pp. 1–11, 2013.
8. C. Evans, P. Moore, and A. Thomas, "An intelligent mobile advertising system (iMAS): location-based advertising to individuals and business," in *2012 Sixth International Conference on Complex, Intelligent and Software Intensive Systems (CISIS)*, pp. 959 –964, July 2012.
9. Y. Hadas, "Assessing public transport systems connectivity based on google transit data," *Journal of Transport Geography*, vol. 33, pp. 105–116, 2013.
10. B. Ferris, K. Watkins, and A. Borning, "Onebusaway: results from providing real-time arrival information for public transit," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1807–1816, 2010.