Open Access

Open Journal of Web Technologies (OJWT)
Volume 7, Issue 1, 2020

www.ronpub.com/ojwt
ISSN 2199-188X

Adding Semantics to Enrich Public Transport and Accessibility Data from the Web

Paloma Cáceres, Almudena Sierra-Alonso, Belén Vela, José María Cavero, Miguel Ángel Garrido, Carlos E. Cuesta

High Technical School of Computer Engineering, Rey Juan Carlos University,
C/ Tulipán s/n, 28933-Móstoles, Madrid, Spain,
{paloma.caceres,almudena.sierra,belen.vela,josemaria.cavero,miguel.garrido.carlos.cuesta}@urjc.es

ABSTRACT

Web technologies and open data practices have now begun to promote new issues and services addressed to both final and specialized users. The smart cities initiative has also introduced new trends and ideas to offer to the public, one of which is the challenge of a more inclusive society that will provide the same opportunities for all. One of the major areas that could benefit from these new initiatives is public transport by, for example, providing open and accessible datasets, which include information by and about people with special needs. In this sense, the Google Transit Feed Specification (GTFS) defines a format to describe public transportation and associated geographic information. It includes details regarding accessibility and what people with special needs might require to get around using public transport. We are, however, of the opinion that this specification has a low granularity and is not sufficient, since it only takes into account only mobility needs. As suggestions for improvement, we propose to enrich GTFS data by combining public transport data from multiple Web sources with semantic metadata techniques. Those data are stored in a public semantic dataset. To define this dataset, we propose a systematic method to extract data from different sources and integrate them. This method is applied to obtain data about the metro system from the website of Metro Madrid and GTFS. Relevant SPARQL queries and two applications are developed to evaluate the usefulness of the dataset obtained.

TYPE OF PAPER AND KEYWORDS

Application Paper: *public transport, accessibility, semantic Web, open data, linked data, semantic annotation*

1 INTRODUCTION

The emergent smart city initiative is based on using Information and Communication Technologies (ICT) in order to provide the public with new services, taking into account aspects as relevant as, for example, transport. However, people with special needs have difficulties moving on public transport: it is difficult for them to identify a specific route that will satisfy their accessibility needs. In fact, Google Maps does not even offer this possibility. Despite the claim made by Verge[1] newspaper in December 2016 that "Google Maps now shows if a location is wheelchair accessible", it is not possible to obtain an accessible route on public transport

[1] http://www.theverge.com/2016/12/15/13968054/google-maps-twenty-percent-wheelchair-accessible

using Google Maps, nor does it provide accessibility information for users with special needs other than wheelchairs.

Information concerning accessibility is sometimes available on the Internet, but those data consumers who wish to develop services with this existing information have problems as regards extracting useful data: the relevant information is sometimes embedded in the HTML code of the Web pages, and it is consequently difficult to obtain and reuse it.

We have, therefore, carried out a study of the accessibility data related to the metro, which we obtained from the official Websites [4]. We analysed whether any accessibility information existed and the format of that information in order to determine the possibility of reusing it. This analysis was carried out by following Tim Berners-Lee's open data5-star[2] classification, which defines five information categories concerning the capacity of a computer to use those data (i.e. a PDF or HTML, and an XML file are processed in a different way; a star belongs to a PDF or HTML file and three stars belong to an XML file). The highest category corresponds to linked data. When referring to computing, Wikipedia defines linked data[3] as "structured data which is interlinked with other data so that it becomes more useful through semantic queries. It builds upon standard Web technologies, but rather than using them to serve web pages only for human readers, it extends them to share information in a way that can be read automatically by computers.

Tim Berners-Lee, director of the World Wide Web Consortium(W3C), coined the term linked data in a 2006 design note about the Semantic Web project. Linked data may also be open data, in which case it is usually described as linked open data(LOD)". Our study focuses on the information obtained from the official metro websites of fourteen European cities (Barcelona, Berlin, Hamburg, Istanbul, London, Madrid, Milan, Moscow, Munich, Oslo, Paris, Rome, Saint Petersburg and Vienna) [4]. Only 57% of the studied metro companies publish accessibility data. Moreover, we concluded the following: 0% of the cities have four or five stars for the data shown on their websites, that is, there are no linked open data on the metro companies' official websites; 21% of the cities provide only data as regards the three star category (CSV or XML format); and 71% of the cities provide open data in the case of the one-star category (HTML format). It is consequently possible to conclude that finding relevant information about public transport accessibility on the Internet and reusing this information is sometimes difficult.

As the Global Initiative for Inclusive Information and Communication Technologies[4] and World ENABLED[5] establish, one of the priorities as regards increasing accessibility is "Using open and accessible datasets that include information by and about persons with disabilities" [26][25]. In this respect, we propose to generate a public transport dataset that integrates the accessibility information from different web sources, with the following goals: to enrich the information concerning public transport with accessibility data and, provide open data to consumers who could reuse them to develop new specific mobility services. Furthermore, Berners Lee proposes "using the Web to connect related data that wasn't previously linked (also known as linked data)". Linked data extends the Web in order to share information that can be read automatically by computers by adding metadata (semantic data) to existing data and interlinking them with one another. Bearing in mind the need for public transport accessibility information for users and data consumers, and following the open and linked data initiative, we feel that it is necessary to generate an open and linked dataset of public transport information and its accessibility features so as to provide the public with data and services according their needs. The information is obtained from both open data sources and Web pages, and it is for these reasons that we need to carry out a data integration process.

In this work we, therefore, propose a systematic method to generate a first approach for a linked and open dataset obtained from data concerning different means of transport (multi-modal public transport) from different open data sources and Web pages. We have followed the proposed method in order to generate a specific dataset by means of a case study with real public transport data related to the city of Madrid in Spain. The data concern two means of public transport (the subway and light rail) and are obtained from an open data source that provides public transport data in the Google Transit Feed Specification[6] (GTFS) format and were obtained from different Web pages whose transport data are embedded in their HTML code. The generation process is carried out in an iterative manner, which allows us to enrich and modify the method so as to improve it in each step. We have also validated the dataset in two ways: we have first stored the data in a semantic repository and carried out a set of queries in order to verify the data quality, and we have then used the dataset in two Android apps. One of these generates metro and light rail routes according to the users' needs, while the other, which is based on gamification and crowdsourcing techniques, permits the users to capture an ACcessibility

[2] http://5stardata.info/en/

[3] https://en.wikipedia.org/wiki/Linked_data; https://www.w3.org/standards/semanticweb/data

[4] https://g3ict.org/resource_center

[5] http://worldenabled.org

[6] https://developers.google.com/transit/gtfs/

Element (ACE) represented on a map and to indicate whether or not an ACE is operational. This dataset is now available as open linked data for data consumers and can be downloaded from our Coruscant server [7]. The dataset can also be queried by means of a Fuseki[8] endpoint[9] created for this purpose. The dataset validation was carried out by means of a set of SPARQL queries and the development of two Android apps.

In summary, the main objective of this work is the definition of a linked open dataset containing data regarding public transport and its accessibility features. The format of this data allows data consumers to use it to generate new mobility services, thus improving the lives of citizens and public transport users. As mentioned previously, not even Google's giant provides accessible routes for all kinds of people with special needs who wish to use public transport. Our dataset solves this problem, since it integrates and links complementary information, resulting in a complete and unique dataset that increases the number of ACcessibility Elements (ACE). Moreover, the information originates from different Web sources, public transport agencies and organizations, and is intended for different means of public transport. This dataset is obtained by following the systematic method proposed herein. The work carried out for the definition and validation of this dataset adds the following contributions:

- Definition of a systematic method to generate semantically-enriched datasets with data from different sources.
- Application of this method to obtain a dataset with accessibility data from the Metro Madrid. This case study uses real-world data about public transport network and includes accessibility data.
- Sixteen significant SPARQL queries to check the correction of semantic annotation of the dataset.
- Two applications that use this dataset to evaluate its usefulness. In addition, these applications provide useful services for citizens. One of them calculates accessible routes by metro between two stations in the city of Madrid, and the other one allows you to add accessibility information using gamification.

The paper is structured, as follows: Section 2 presents the related works; Section 3 describes our specific proposal, including the background, the systematic method employed to add semantic data, a case study and a summary of the results of our case study; Section 4 presents the validation of our proposal, while our conclusions and future work are presented in Section 5.

2 RELATED WORK

Several works have followed the smart cities initiative with the intention of providing the public with services in order to improve their mobility. If this accessibility information is to be included, it is crucial to make improvements. In this respect, there are proposals such as Wheelmap[10] and/or WheelMate[11], whose objective is to include accessibility information on maps. Wheelchair Routing[12] is a project that develops a route planner especially for wheelchair users based on OpenStreetMap[13] (OSM) Data. AccessMap provides accessible routes to people with mobility needs[14]. With regard to Google's giant, Google Maps does not, as yet, provide an accessible route.

The GTFS (see Footnote 5) defines a format for public transportation and associated geographic information and does include a little information about accessibility and the special needs of certain public transport users. OpenTripPlanner (OTP) is another project that provides services for passenger information and transport network analysis [22]. It computes routes combining multiple means of transportation (pedestrian, bicycle, buses, metro, car, etc.) built from OpenStreetMap (see Footnote 13) and GTFS (see Footnote 6) data. OTP also takes (transport) wheelchair accessibility into account. But none of these proposals considers disabilities other than those related to mobility.

Furthermore, "Ciudades Patrimonio de la Humanidad" provides accessible routes for not only users with mobility needs but also the blind or hearing-impaired people[15]. However, the users cannot establish origin or destination points that have not been considered by the application.

The European Union is making a great effort to improve public transport [13] and has founded projects such as ACCESS 2 ALL [1], Mediate [19] or CIVITAS [10]. One of the objectives of all of them is to analyse how to provide an answer to all citizens' accessibility needs: ACCESS 2 ALL makes a complete study of the special user needs and proposes customised services for route guidance coupled with localisation methods. The Mediate Project has identified a set of indicators with which to describe accessibility, has developed a tool that can measure the accessibility of urban transport and has

7 http://coruscant.my.to:8080/download/metro.xml

8 https://jena.apache.org/documentation/fuseki2/

9 http://coruscant.my.to:3030/dataset.html?tab=query&ds=/dataset_metro

10 https://wheelmap.org

11 http://myhealthapps.net/app/details/13/wheelmate

12 https://wiki.openstreetmap.org/wiki/Wheelchair_routing

13 https://www.openstreetmap.org

14 https://www.accessmap.io

15 http://www.ciudadespatrimonio.org/accesibilidad/?idioma=en

published a Good Practice Guide for accessibility. One of the subprojects of CIVITAS is the search for mobility strategies for vulnerable groups. These projects seek to establish a theoretical framework covering all aspects of citizens' mobility, but do not offer solutions in the form of user applications.

We estimate that the number of systems whose objective is to supply information about accessibility is increasing. However, most of them consider only mobility needs. To the best of our knowledge, there are currently no software applications that analyse the status of the public transport network in order to estimate the availability of accessibility features.

In order to provide the public with services with which to improve their mobility, it is necessary to have useful data, that is, data in a format that can be processed by a computer. Linked Open Data would appear to be the best solution by which to unify the format of the data and make them available to all [7]. In these sense, there are works that propose to use semantic technologies to annotate and link data so as to obtain a set of processable data. We can classify the proposals analysed in two groups: those, which define a generation process for linked data, and those, which specify vocabularies that describe features of public transport.

With regard to the first group, Belk et al. [1] proposes to add semantic mark-ups in the context of Adaptive Interactive Systems; Hyland and Wood [15] describe a process with which to generate government data and Hidalgo et al. [14] describe a set of methodological guidelines that can be used to publish library data on the Web. There are also specific proposals focused on the transport domain: Najdenov et al. [21] describe how to take into account events on roads, while Keller et al. [17] and, Plu and Scharffe [23] incorporate transport linked data on the Web by employing the Transmodel standard [27] as a basis. The mPASS project [24] analyses urban accessibility in general so as to allow the planning of accessible routes. However, it does not detail how people with special needs can use public transport. ASK-It [30] presents a methodology with which to define the requirements of people with functional limitations and represent them by means of an ontology that classifies those functional limitations and relates them to user actions Services for those users are then developed. But, none of them specifies the accessibility features of public transport.

With regard to the various vocabularies related to public transportation, several semantic models represent specific aspects of the public transport domain. Although some of them include accessibility features for people with special needs or disabilities, to the best of our knowledge, there are currently no projects that fully define accessibility limitations and their relationship to the features of public transport. Some of the most representative of them are presented below.

NAPTAN [20] is a vocabulary with which to identify, in a unique manner, the national public transport access nodes of the United Kingdom. It does not incorporate any aspect of accessibility. The Ofi-ontology[16] proposal, meanwhile, makes it possible to represent whether a place is accessible to people with mobility problems by means of classes such as `AccessFacilities` and properties such as `is_wheelchair_accessible`, but it does not take into account other necessary elements in order to provide blind or deaf people with information. DBpedia[17] has a property denominated as `isHandicappedAccessible` to indicate whether or not a transit station is accessible. Li et al. [18] incorporate only step free and lift facilities as accessibility elements (mobility disabilities) into their Tube Facility ontology. Accessibility ontology [25] is similarly also focused on concepts related to supporting only mobility disability problems.

Ding et al. [8] propose an ontology with which to link datasets retrieved from four different sources in the UK. They enrich data and present an architecture so as to provide data for accessible travelling [9]. But they take only mobility impairment into account. TRANSIT [11] is a specific ontology for transit but it does not incorporate any aspects of accessibility to human transport. The European OASIS (Open architecture for Accessible Services Integration and Standardisation) project [12] does not incorporate relevant concepts for needs regarding accessibility to public transport.

In order to annotate and link data, it is necessary to establish relationships among the domain concepts: how the infrastructure elements are interconnected, which accessibility features every element has and which accessibility needs each disability has. We model this knowledge based on two standard models that describe public transport features: Transmodel [27] and IFOPT [16].

Transmodel [27] is a European Reference Data Model for Public Transport Information that provides both a model of public transport concepts and data structures that may be useful when building information systems related to the different kinds of public transport. This model defines data for the description of networks and information related to vehicle and driver scheduling, personnel (driver) disposition, operations monitoring and control, passenger information, fare collection and management information and statistics [27]. As our objective is to describe the accessibility of a specific means of transport (specifically, Metro de Madrid) we

[16] http://ip-kom.net/data/html_documentation_ontologien/ofi-ontology.html

[17] http://mappings.dbpedia.org/index.php/OntologyProperty:IsHandicappedAccessible

use the part of the Transmodel model relating to the description of the network. However, it does not provide any information regarding the accessibility elements of public transport.

Another standard, IFOPT metamodel [16], was conceived as an extension of Transmodel. It is a CEN[18] Technical Specification that defines a model (and also the identification principles) for the main fixed objects related to public access to public transport (e.g. stop points, stop areas, stations, connection links, entrances, etc.). These concepts are identified in Transmodel, but IFOPT details them and adds more: It already includes specific structures with which to describe accessibility data concerning the equipment of vehicles, stops and access areas.

IFOPT is divided into four parts: Administrative Area Model, Topographical Model, POI (Point Of Interest) Model and Stop Place Model. The last one "describes the detailed structure of a STOP PLACE (that is station, airport, etc.) including physical points of access to vehicles and the paths between the points, including ACCESSIBILITY" [16]. IFOPT uses many terms and concepts from Transmodel and adds many elements of stations, denominated as "fixed points". Some of those elements are quays, entrances, paths, stop areas, boarding positions, connection links, etc., and, of course, the equipment at the stop places: lifts, ramps, escalators, stairs, travellators, wheelchair areas, etc. The IFOPT standard, therefore, provides the model we require to express the accessibility of a metro station. But IFOPT goes further, since it defines the special needs a user may have: mobile, visual or auditory impairment; an aversion to lifts, escalators, confined spaces; a guide dog, oversize baggage, etc., and establishes the relationship among accessibility part constraints, the equipment required and some user needs.

3 A PROPOSAL TO GENERATE A SEMANTIC DATASET

The eventual objective of our work is to improve transit information in order to support new social accessibility services, which will have a great impact on society, such as calculating public transport routes that are accessible for all.

We are working to generate a semantic dataset that integrates data from different sources by unifying the terms used for the same concept in each of the sources to provide a unique source of data. This dataset is provided as a service for public transport domain applications with web-scale data. In this paper, we describe a systematic method with which to generate semantic data concerning multi-modal public transport obtained from multi-source Web data by adding new semantics in order to enrich the original information. Furthermore, data are related to each other to obtain a dataset of interlinked data in the sense described by Tim Berners-Lee, as we mentioned in the Introduction. We are generating the dataset by working with semantic web technologies such as RDF(S), RDF and SPARQL.

The dataset generated, which is denominated as Dataset_v2, is open and can be downloaded from our Coruscant server (see Footnote 7).

In this section, we first describe the background to this work, which includes our previous work and the semantic technologies required to carry it out. A description of the process employed to generate open and linked unified data by integrating data from different web sources and formats, is then provided, after which a case study is carried out with real data of the Metro Madrid public transport company. Finally, we include a summary of the data results attained after carrying out the case study.

3.1 Background

In this section, we introduce the underlying semantic technologies and our previous work regarding this proposal: The MAnto vocabulary.

3.1.1 Semantic Annotation

In order to obtain a useful dataset, these data must be public and linked. To link data, it is necessary to add metadata so as to provide the semantics of the data to be linked. This can be done by using various semantic Web technologies: RDF(S)[19] (Resource Description Framework Schema) or OWL[20] (Ontology Web Language), which make it possible to define the set of semantic terms (ontology); RDF/XML[21] (Resource Description Framework/eXtensible Markup Language), which permit the storage of the data annotated with those semantic terms; and SPARQL[22] (a query language with which to extract data from an RDF/XML file).

The use of RDF(S) (Resource Description Framework Schema) or OWL (Ontology Web Language) allows us to define a vocabulary or ontology that makes it possible to add metadata to the data. These metadata permit us to define the data semantics. RDF(S) and OWL have a set of classes and properties that provide basic elements for the description of ontologies (vocabularies of a specific domain, e.g. the public transport domain).

[18] "Comité Européen de Normalisation, CEN" in French.

[19] https://www.w3.org/TR/rdf-schema/

[20] https://www.w3.org/TR/owl-features/

[21] https://www.w3.org/TR/rdf-syntax-grammar/

[22] https://www.w3.org/TR/sparql11-query/

The MAnto ontology created to annotate data from multiple Web sources and formats is described in the following section. We use the MAnto vocabulary to annotate the data in a semantic and unified manner and to provide open and linked data.

3.1.2 The Early Version of the *MAnto* Ontology

Our initial work regarding public transport and its accessibility features was related to the public bus network. We based it on the IFOPT model and we had to extend it to support all the accessibility features available for the buses in the network. We, therefore, defined a conceptual model that specifically described the domain of vehicle equipment that could, for example, allow users to access a bus with a bicycle. This work resulted in the first version of MAnto (Mobility and Accessibility Ontology), which included the terminology that describes the vehicle equipment [3].

Another previous work was related to the public metro network [4]. We again based this on the IFOPT model and did not need to extend it. In this case, we generated a dataset based on data from the official Metro Madrid website[23], denominated as Dataset_v1[24] [29]. This work extended the first version of MAnto[25] and was then used to semantically annotate the original data (by adding metadata to data).

In summary, MAnto is based on the Identification of Fixed Objects in Public Transport (IFOPT) reference data model [16] and defines a set of terms that describe the accessibility characteristics of the means of transport in order to help passengers on their journeys through cities. For example, one element of public transport is a station (or, to employ the IFOPT terminology, a stop place). A stop place could or could not have a lift or escalator. MAnto vocabulary must, therefore, have *stopPlace*, *lift* and *escalator* as terms. These terms will be the metadata of the semantic annotation.

MAnto provides a comprehensive vocabulary to describe different modes of public transport and their accessibility elements. It describes not only physical elements of transport, such as a station, a platform, an access, etc., but also includes how stations are arranged in lines, the possible connections, the geographical locations at a certain point on a map, etc. But its main contribution is the incorporation of accessibility information so as to facilitate the use of public transport by people with special needs.

Figure 1 provides an example of a definition from our ontology: a stopPlace is (`rdf:type`) a resource element (`rdfs:resource`), which has two properties (`hasEscalator`, `hasLift`). These properties have a stopPlace as a source (`rdfs:domain`) and a boolean value as a destination (`rdfs:range`).

Table 1 shows a subset of terms of our MAnto ontology (those presented in Figure 1. Every term has to be written with the MAnto prefix ("`mao:`").

Once the vocabulary (or ontology) has been defined, we can annotate the data source, which is done using the RDF/XML language. The RDF/XML (Resource Description Framework/eXtensible Markup Language) is a standard model that makes it possible to describe data for data interchange on the Web in a semantic manner, using an ontology.

We shall employ the partial schema in Figure 1 and the MAnto terms in Table 1 to describe that the ALCORCON CENTRAL station has a lift and an escalator in the following manner: the "`est_4_211`" resource is a stopPlace, which is indicated with the `<rdf:type rdf:resource="stopPlace"/>` syntax. Its name is ALCORCON CENTRAL, which is indicated with the `<sch:name> ALCORCON CENTRAL </sch:name>` sintax (from the Schema[26] vocabulary). The station has an escalator (`<hasEscalator>TRUE</hasEscalator>`) and a lift (`<hasLift>TRUE</hasLift>`). The RDF/XML code generated is shown in Table 2.

Moreover, we can retrieve semantic data using SPARQL queries. SPARQL is a query language with which to extract data from a graph in RDF format. For example, we can obtain a station and its `hasLift` value from the partial RDF/XML file in Table 2 (see Table 3).

3.2 A Systematic Method: Adding Semantic to Public Transport and Accessibility Web Data

In this section, we develop a systematic method with which to generate an open and linked dataset from the data regarding different means of public transport and Web sources. This method makes it possible to enrich data by adding new semantics. The process is described in the following steps:

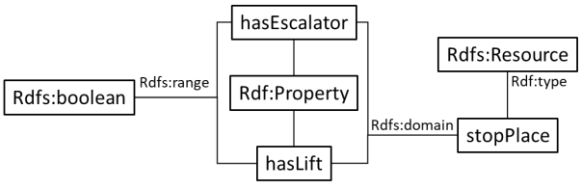


Figure 1: Example using RDF(S)

[23] http://www.metromadrid.es/en/

[24] https://raw.githubusercontent.com/vortic3/LinkedUnifiedDataset/master/Dataset_v1.xml

[25] https://vortic3.github.io/MANTO/

[26] https://schema.org/

Table 1: MAnto terminology subset

MAnto Term	Description
mao:stopPlace	A stop place
mao:hasLift	Lift to access to the station
mao:hasEscalator	Escalator to access to the station

Table 2: Semantic data using RDF/XML

```
<rdf:Description rdf:about="est_4_211">
   <rdf:type rdf:resource="stopPlace"/>
   <sch:name>ALCORCON CENTRAL</sch:name>
   <mao:hasEscalator>TRUE</mao:hasEscalator>
   <mao:hasLift>TRUE</mao:hasLift>
</rdf:Description>
```

Table 3: Example using SPARQL

SPARQL sentence:

```
SELECT ?station ?value
WHERE {
     ?station mao:hasLift ?value
}
```

Obtained data:

```
est_4_211   mao:hasLift   TRUE
```

Step 1. Study the information and accessibility features of the public transport network so as to include them in the dataset: create a glossary of terms concerning the terminology used in the data source, describing each one, and generate a class diagram with which to represent these terms and their relationships.

Step 2. Identify the data semantics of this information in order to align them with our reference vocabulary (MAnto): establish the mappings between the elements of the data source and the MAnto terms by means of a table. A mapping could be a direct or a processed transformation. In the latter case, when necessary, some pseudo-code has to be included.

Step 3. If the alignment is possible: (a) obtain the original data (develop a solution with which to obtain the data from the source); (b) semantically annotate them using existing terms from MAnto or, (c) from other vocabularies only when necessary. If the alignment is not possible: (d) analyse the difference between the original data and the reference vocabulary in order to extend it and, in this case, return to the second step.

Step 4. Integrate the dataset with other sources: (a) taking into account multimodal public transport: analyse new data and repeat from step 1 to step 3 when necessary; (b) taking into account previous public transport datasets.

3.3 The Case Study: Applying the Method to the GTFS Data Source

Details of the specific process that we have carried out in this work by means a case study are provided as follows: semantically annotate the real data of the Metro Madrid public transport company from the GTFS (steps 1, 2 and 3); unify and integrate them with the data previously annotated from Dataset_v1 (step 4).

Step 1. Study the information and accessibility features of the public transport network included in the GTFS: we have studied the whole GTFS. A GTFS feed is a collection of a minimum of six CSV files, with a .txt extension. One of these files is the *stops.txt* file, which describes a public transport feature (station, stop or entrance/exit to a station). The *stops.txt* file has the following structure: *stop_id*, *stop_code*, *stop_name*, *stop_desc*, *stop_lat*, *stop_lon*, *zone_id*, *stop_url*, *location_type*, *parent_station*, *stop_timezone*, *wheelchair_boarding*. We shall now describe each of the terms in this structure.

The *stop_id*, *stop_code*, *stop_name* and *stop_desc* columns represent the identification, code, name and description of a public transport feature (the station name is really the *stop_desc* column while the *stop_name* is a short name); *stop_lon* and *stop_lat* represent the longitude and latitude coordinates employed to geographically locate the feature; *zone_id* is a zone identifier (to establish different rate zones); *stop_url* permits the inclusion of a web page of the public transport feature; and *stop_timezone* indicates the time zone in which this stop, station, or station entrance is located.

A complete description of the information regarding the following items will be provided owing to their complexity: (a) the *location_type*, (b) *parent_station* and (c) *wheelchair_boarding* columns:

(a) *location_type* represents three different physical entities: station, stop and station entrance/exit (*location_type*=1, 0 or empty, 2, respectively). A station could have different stops and station entrances/exits. According to the GTFS, a station is defined as an area or physical structure that includes one or more stops, along with one or more entrances or exits. That is, one station will be composed of stops and entrances. Stop describes the area or place where passengers get on/off the vehicle to start, continue or end the trip. Station entrance/exit represents the physical point between a place of interest and where the trip starts or ends, or a transfer is made.

(b) The *parent_station* column allows us to establish the relationships between the three entities described by *location_type*. When *location_type* represents a stop or a station entrance/exit, the *parent_station* represents the station to which they belong. As mentioned above, a

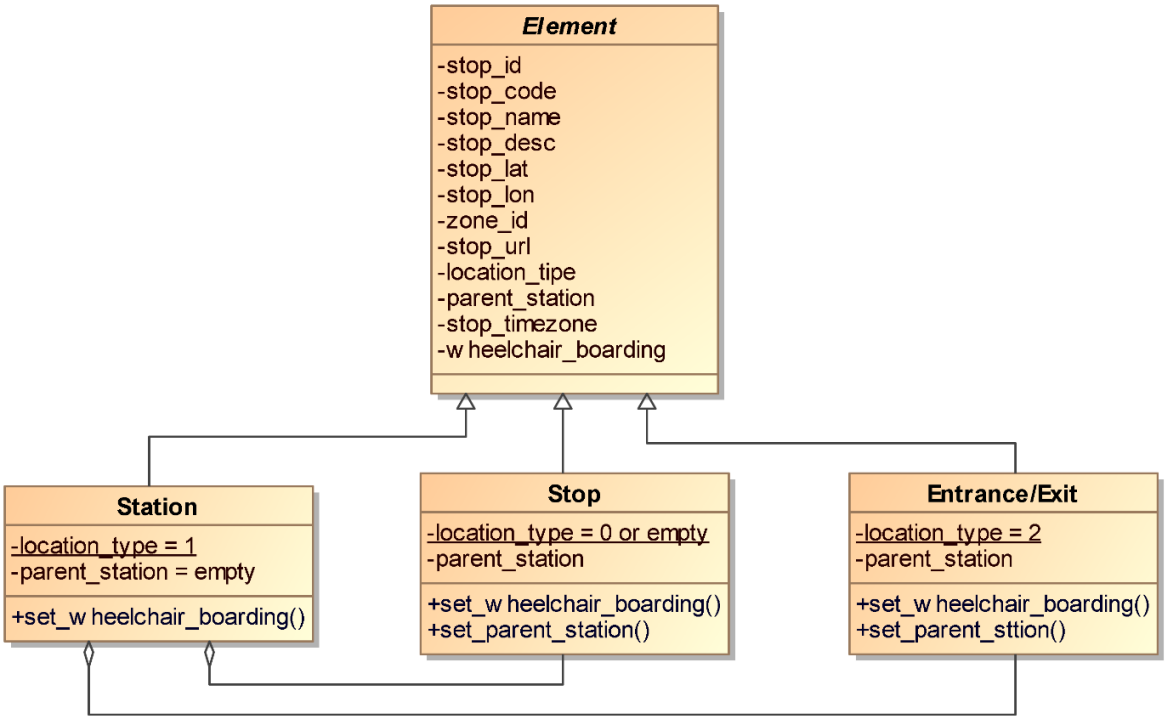


Figure 2: Conceptual Model of the GTFS

station is composed of stops and station entrances/exits and a stop or a station entrance/exit will, therefore, inherit some functionalities or capabilities from its parent station.

(c) The *wheelchair_boarding*, as the GTFS states, "identifies whether wheelchair boardings are possible from the specified stop, station, or station entrance." The *wheelchair_boarding* has values depending on the *location_type* column.

When *location_type* represents a station, *wheelchair_boarding* has the following values:

0 (or empty) - indicates that there is no accessibility information for the station.

1 - indicates that at least some vehicles at this station can be boarded by a person in a wheelchair.

2 - wheelchair boarding is not possible at this station.

When *location_type* represents a stop, the values are the following:

0 (or empty) - the stop will inherit its *wheelchair_boarding* value from the parent station, if specified in the parent.

1 - there is some sort of accessible path from outside the station to the specific stop/platform.

2 - there is no accessible path from outside the station to the specific stop/platform.

When *location_type* represents a station entrance/exit, *wheelchair_boarding* has the following values:

0 (or empty) - the station entrance/exit will inherit its *wheelchair_boarding* value from the parent station, if specified in the parent.

1 - the station entrance/exit is wheelchair accessible.

2 - there is no accessible path from the station entrance/exit to the stop.

In order to clarify this specification, we created a conceptual model represented with a UML class diagram with which to describe it (Figure 2). It describes the metro elements and their relationships following the GTFS specification. The *Element* class is abstract and declares the attributes, which define the spaces that are part of metro stations. These attributes are inherited and take a value when we create instances of the *Station*, *Stop* or *Entrance/Exi*t classes.

Three of these attributes have a special performance: when they are inherited, they take a value in a particular way:

- *location_type* is a static attribute in the *Station*, *Stop* or *Entrance/Exi*t classes. If an object is an instance of a *Station* class, its value is 1, and it is 1 for any instance of *Station*. *Location_type* is similarly 0 or does not value the *Stop* class, and is 2 for the *Entrance/Exit* class.

Table 4: Mappings between GTFS and MAnto

GTFS	MAnto
stop_code	mao:stopplacecode
zone_id	mao:tariffZone
If location_type is station	mao:stopPlace
wheelchair_boarding=0	mao:wheelchairAccess=UNKNOWN
wheelchair_boarding=1	mao:wheelchairAccess=TRUE
wheelchair_boarding=2	mao:wheelchairAccess=FALSE
If location_type is stop	mao:quay

wheelchair_boarding=0	mao:wheelchairAccess= set_wheelchair_boarding ();
	set_wheelchair_boarding ()
	{search station.stop_id = stop.parent_station
	then read (station.wheelchair_boarding);
	assign(stop.wheelchair_boarding,
	station.wheelchair_boarding)
	}

wheelchair_boarding=1	mao:wheelchairAccess=TRUE
wheelchair_boarding=2	mao:wheelchairAccess=FALSE
If location_type is station entrance/exit	mao:entrance

wheelchair_boarding=0	mao:wheelchairAccess= set_wheelchair_boarding ();
	set_wheelchair_boarding ()
	{search station.stop_id =
	station_entrance.parent_station
	then read (station.wheelchair_boarding);
	assign(station_entrance.wheelchair_boarding,
	station.wheelchair_boarding)
	}

wheelchair_boarding=1	mao:wheelchairAccess=TRUE
wheelchair_boarding=2	mao:wheelchairAccess=FALSE

Table 5: Combining vocabularies

GTFS	From other vocabularies
stop_name	sch:name (from schema.org) (*)
stop_lat	wgs84_pos:lat (from WGS84 Geo Positioning)
stop_lon	wgs84_pos:long (from WGS84 Geo Positioning)
parent_station	gn:parentFeature (from geonames.org)
stop_timezone	gn:timeZone (from geonames.org)

- *parent_station* indicates of which station it is part. This is designated by means of the *stop_id* attribute of the corresponding instance of the *Station* class
- *wheelchair_boarding* takes a value depending on the values of the *location_type* and *parent_station* attributes. This dependence is shown in Table 4.

At this point, we wish to underline that the accessibility information provided by the GTFS refers only to whether a user can access a public transport station or stop when in a wheelchair. The GTFS does not consider any other users' needs, such as a phobia as regards escalators or lifts.

Table 6: MAnto extension

GTFS	New terms
stop_name	mao:shortName (from IFOPT: Short name) (*)
stop_desc	mao:name (from IFOPT: name)
stop_url	mao:infoLink (from IFOPT:Info link)

Step 2. Identify the data semantics of the GTFS in order to align them with our reference vocabulary (MAnto): We have analysed the different columns and their semantic information obtained from *stops.txt* in order to align them with our MAnto ontology. As mentioned in step 1, *location_type* could be a station, stop or station entrance/exit. These different features will, nevertheless, all be annotated in *location_type*. Table 4 shows the mappings employed to annotate data by means of MAnto.

Step 3. If the alignment is possible: (a) obtain the original data, (b) semantically annotate them and (c) introduce specific terminology from other vocabularies only when necessary. If the alignment is not possible, then (d) analyse the difference between the original

GTFS data and the reference vocabulary (MAnto) in order to extend it and, in this case, return to the second step; it will, in all cases, be necessary to combine MAnto vocabulary with others:

In this step, (a) the original Web data from the GTFS have been obtained automatically from the Web source. We have implemented a crawler to download the Metro and Light Metro data from the Transit feeds of Madrid public transport web page[27].

(b) With regard to the alignment, the annotated data will be that corresponding to Table 4 (only five terms from GTFS can be directly transformed to MAnto).

(c) In this case, we need to annotate the name of the station, stop or station entrance/exit and the geographical or topological information so as to represent them on a map. We also need to know the parent station of a stop and of a station entrance/exit and the time zone. Table 5 shows the different vocabularies used in this step.

(d) When the alignment has not been possible, it is necessary to extend MAnto in order to define new terminology (see Table 6). In this case, we have defined three new terms. As will be noted in Table 5 and Table 6, we require a *name* term (*) with which to identify a station or stop name. We could use `sch:name` (from schema.org vocabulary) or `mao:shortName` from extended MAnto. Finally, we use the existing `sch:name` metadata.

Step 4. Integrate the dataset with other sources and datasets: (a) taking multimodal public transport sources into account; (b) taking into account previous public transport datasets.

(a) It is necessary to consider data from another means of public transport (in this case, *light rail*). These GTFS data are also obtained in the same way as those of the metro and we have, therefore, applied the same steps (from step 1 to step 3). The light rail annotated data have subsequently been added to the previously obtained GTFS metro data, specifying a new term (`mao:transportFor`) with which to distinguish the means of transport (*metro* and *rail metro*).

(b) Finally, it is necessary to integrate them into Dataset_v1. We combine the two datasets related to the station names (`sch:name`). It is consequently necessary to discard duplicated information and to add new information in order to enrich the final dataset (Dataset_v2). In this work, Dataset_v2 will be the integration of Dataset_v1 and the semantic data annotated from GTFS sources in this work.

3.4 A New Version of MAnto

As a result of the application of the proposed method to the case study, we have obtained an updated version of MAnto, which includes new terms with which to annotate data following the GTFS specification. Table 7 presents the terminology required to semantically annotate the data that we handle in this work.

Table 8 contains the code related to the terms of the new version of MAnto (marked in grey in Table 7), which is implemented by means of the Ontology Web Language[28] (OWL) and can be downloaded from the GitHub repository[29].

3.5 Summary

In summary, the process carried out to annotate the GTFS data and to integrate it into the Dataset_v1 results in a total number of 18.914 triples in Dataset_v2. Table 9 presents the related data.

As result of this work, we have extended MAnto vocabulary and the definitive code is next shown.

We have calculated the number of triples and their percentages in the following way: (a) from GTFS data: how many triples originate from metro or light rail; (b) from Dataset_v1: how many triples are added in order to integrate the accessibility features and connections within or with other transport media from the Metro Madrid website into the previous data.

We have also summarized the percentage of MAnto terms used in this work. We have grouped them by existing terms and extended terms. We have also grouped them by the specific origin of the data. Table 10 and Table 11, respectively, show this information.

4 VALIDATION

The validation of this proposal is carried out by: implementing a specific set of SPARQL queries in order to guarantee the data quality and whether the semantic is according to the contents (in subsection 4.1), and using the dataset in different apps, which provide metro accessibility information for users who have special needs (in subsection 4.2). We have verified the usefulness and correctness of the semantic data in both of the above cases.

4.1 Validation of Data Quality and Semantics

We have validated the quality and semantics of the data by developing sixteen queries in order to retrieve the associated data from the semantic dataset. We have

[27] https://transitfeeds.com/p/consorcio-regional-de-transportes-de-madrid

[28] https://www.w3.org/OWL/

[29] https://raw.githubusercontent.com/vortic3/LinkedUnifiedDataset/master/MAnto_Lite_ontology.rdf

Table 7: Terminology of MAnto

MAnto terms	Description
<mao:address>	The address of the stop place
<mao:entrance>	To indicate that the feature is an entrance
<mao:quay>	To indicate that the feature is a quay
<mao:stopPlace>	To indicate that the feature is a stop place (station)
<mao:hasEscalator>	To indicate whether or not a stop place has an escalator
<mao:hasLift>	To indicate whether or not a stop place has a lift
<mao:hasRamp>	To indicate whether or not a stop place has a ramp
<mao:hasTravelato>	To indicate whether or not a stop place has a travelator
<mao:stopplaceCode>	To preserve the code of the stop place from GTFS data
<mao:tariffZone>	To indicate the tariff zone in which this means of transport operates
<mao:wheelchairAccess>	To indicate whether or not a stop can be accessed by a person in a wheelchair
<mao:Transfer>	To indicate whether there are connections with other means of transport.
<mao:infoLink>	To show a link to additional information
<mao:transportFor>	To indicate the kind of means of transport (metro, bus,…)
<mao:connectionLink>	To indicate whether there are connections with other lines of this means of transport.

Table 8: Code OWL to define MAnto

```
<?xml version="1.0"?>
<rdf:RDF xmlns="http://com.vortic3.MANTO#"
     xml:base="http://com.vortic3.MANTO"
     xmlns:mao="http://com.vortic3.MANTO#"
     xmlns:owl="http://www.w3.org/2002/07/owl#"
     xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
     xmlns:sch="http://schema.org/"
     xmlns:xml="http://www.w3.org/XML/1998/namespace"
     xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
     xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">

<owl:Ontology rdf:about="http://com.vortic3.com/MANTO#"/>

<owl:DatatypeProperty rdf:about="http://com.vortic3.com/MANTO#address">
<rdfs:domain rdf:resource="http://com.vortic3.com/MANTO#Entrance"/>
<rdfs:domain rdf:resource="http://com.vortic3.com/MANTO#Quay"/>
<rdfs:domain rdf:resource="http://com.vortic3.com/MANTO#StopPlace"/>
<rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#string"/>
</owl:DatatypeProperty>

<owl:DatatypeProperty rdf:about="http://com.vortic3.com/MANTO#hasEscalator">
<rdfs:domain rdf:resource="http://com.vortic3.com/MANTO#StopPlace"/>
<rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#boolean"/>
</owl:DatatypeProperty>

<owl:DatatypeProperty rdf:about="http://com.vortic3.com/MANTO#hasLift">
<rdfs:domain rdf:resource="http://com.vortic3.com/MANTO#StopPlace"/>
<rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#boolean"/>
</owl:DatatypeProperty>

<owl:DatatypeProperty rdf:about="http://com.vortic3.com/MANTO#hasRamp">
<rdfs:domain rdf:resource="http://com.vortic3.com/MANTO#StopPlace"/>
<rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#boolean"/>
</owl:DatatypeProperty>

<owl:DatatypeProperty rdf:about="http://com.vortic3.com/MANTO#hasTravelator">
<rdfs:domain rdf:resource="http://com.vortic3.com/MANTO#StopPlace"/>
<rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#boolean"/>
</owl:DatatypeProperty>
</rdf:RDF>
```

Table 9: Dataset_v2: Number of triples

Source	Kind of data	N. of triples	%
From GTFS data	Metro	16,417	86.85
	Light rail	1,744	9.24
From Metro Madrid Web site (Dataset_v1)	Accessibility features	344	1.83
	Connections with other transport media	67	0.35
	Connections within the same transport media	327	1.74
From public transport media (Dataset_v2)	metro	17,124	90.54
	light rail	1,790	9.46

Table 10: Percentages of MAnto terms

%	From	Terms used
73.3%	Existing terms	<mao:stopplaceCode> <mao:tariffZone> <mao:stopPlace> <mao:quay> <mao:entrance> <mao:wheelchairAccess> <mao:hasLift> <mao:hasRamp> <mao:hasEscalator> <mao:hasTravelato> <mao:Transfer>
26.7%	New terms (MAnto extension)	<mao:address> <mao:infoLink> <mao:transportFor> <mao:connectionLink>

Table 11: Percentages of terms by origin

%	From	Terms used
40%	GTFS	<mao:stopplaceCode> <mao:tariffZone> <mao:stopPlace> <mao:quay> <mao:entrance> <mao:wheelchairAccess> <mao:address> <mao:infoLink>
5%	Different transport media	<mao:transportFor>
30%	Data integration	<mao:connectionLink> <mao:hasLift> <mao:hasRamp> <mao:hasEscalator> <mao:hasTravelato> <mao:Transfer>
25%	Other vocabularies	<sch:name> <wgs84_pos:lat> <wgs84_pos:long> <gn:parentFeature> <gn:timeZone>

created a Fuseki endpoint (see Footnote 9) for this purpose, which is open access. We have followed a systematic validation process: a query is executed and the associated result is then analysed: (a) if the result is correct, the annotation is valid, and the defined mapping is correct; (b) if the result is not correct, the data from the sources are not valid data or the annotation must be redefined: then, we have to analyse what happens and correct the data when possible.

The sixteen queries are grouped in four different categories: direct transformations (1), processed transformations (2), means of public transport (3) and accessibility elements (4).

Implement a query for each direct transformation from a GTFS term to a MAnto or another vocabulary term (these transformations refer to the terms included in Table 4, Table 5 and Table 6). Table 12 shows this validation process when an incorrect result is obtained. Table 13 shows this validation process when a correct result is obtained.

1. Implement a query for each GTFS term, which requires a specific transformation to be included in the final dataset. In our case, *location_type*, *parent_station* and *wheelchair_boarding* need a processed transformation. The results were correct. Table 14 shows an example.
2. Implement a query for each different means of public transport (with the aim of obtaining the whole data associated with that means). Table 15 shows the query and its results.
3. Implement a query to obtain the stop place triples with the *hasLift* label. (We repeat the query for the *hasEscalator*, *hasTravelator* and *hasRamp* labels). The result can be seen in Table 16.

With regard to Query 4, we compare the triples obtained from Dataset_v1 and Dataset_v2. Table 17 shows a fragment of the triples obtained. Note that the name associated with the stop place (underlined text in the code) is different in the two datasets ("`Vodafone Sol`" and "`Puerta del Sol`"). In this case, we need to refine the fitting between the two datasets.

After executing these queries from Dataset_v2, we discovered two kinds of errors:

(i) Some downloaded data do not follow the GTFS specification: data from the Metro and Light Metro data from the Transit feeds on the Madrid public transport website[30] do not follow this specification. We, therefore, detected that an address rather than a description appeared in the description column (*stop_desc*) of the stops.txt file (this error was found by means of the query shown in Table 12). The same occurred in the *stop_url* column. The number of errors encountered was 1,360 for each issue, signifying that the percentage of errors from the GTFS source is 97.39%.

[30] https://transitfeeds.com/p/consorcio-regional-de-transportes-de-madrid

Table 12: Validation query for direct transformation from GTFS to MAnto. Incorrect results

Query 1.a.
`SELECT ?station ?stop_desc`
`WHERE {`
`?station mao:name ?stop_desc}`
	×

Results	The `mao:name` content is not a "name", it is a postal address. We analyse the original data (GTFS source) and our direct transformation to find the error: the data source really contained postal address. Then, we redefine this transformation: now, stop_desc from GTFS will be `mao:address`.

Table 13: Validation query for direct transformation from GTFS to MAnto. Correct result

Query 1.b.
`SELECT ?station ?stop_desc`
`WHERE {`
`?station mao:address ?stop_desc}`
	√

Results	`mao:address` is a correct literal.
Dataset fragment	
`<http://com.vortic3.MANTO#acc_4_12_43><http://www.w3.org/1999/02/22-rdf-syntax-ns#type><http://com.vortic3.MANTO#entrance> .`	
`<http://com.vortic3.MANTO#acc_4_12_43><http://com.vortic3.MANTO#address> "Plaza de la Puerta del Sol 9".`	
`<http://com.vortic3.MANTO#acc_4_12_43><http://www.geonames.org/ontology#parentFeature><http://com.vortic3.MANTO#est_90_58> .`	
`<http://com.vortic3.MANTO#acc_4_12_43><http://com.vortic3.MANTO#wheelchaiAccess> "FALSE" .`	
`<http://com.vortic3.MANTO#acc_4_12_43><http://www.geonames.org/ontology#timeZone> "" .`	
`<http://com.vortic3.MANTO#acc_4_12_43><http://com.vortic3.MANTO#TariffZone> "" .`	
`<http://com.vortic3.MANTO#acc_4_12_43><http://com.vortic3.MANTO#stopplaceCode> "12" .`	
`<http://com.vortic3.MANTO#acc_4_12_43><http://www.w3.org/2003/01/geo/wgs84_pos#lat> "40.41671" .`	
`<http://com.vortic3.MANTO#acc_4_12_43><http://www.w3.org/2003/01/geo/wgs84_pos#long> "-3.70458" .`	
`<http://com.vortic3.MANTO#acc_4_12_43><https://schema.org/name> "Mayor" .`	

Table 14: Validation queries for calculated transformation from GTFS to MAnto

Query 2.
`SELECT ?station`
`WHERE {`
`?station rdf:type mao:stopPlace }`
	√

Results	If GTFS location_type = 1 then the data is a station (`mao:stopPlace` in MAnto).

Query 2.2
`SELECT ?station ?theValue`
`WHERE {`
`?station mao:wheelchairAccess ?theValue}`
	√

Results	If **wheelchair_boarding = 0** (or empty) then `mao:wheelchairAccess`="UNKNOWN" for a specific station, stop o station entrance/exit. If **wheelchair_boarding = 1** then `mao:wheelchairAccess`="TRUE" for a specific station, stop o station entrance/exit. If **wheelchair_boarding =2** then `mao:wheelchairAccess`="FALSE" for a specific station, stop o station entrance/exit.

Table 15: Validation query for each different means of public transport

Query 3.
`SELECT ?station ?PTMedia`
`WHERE {`
`?station mao:transportFor ?PTMedia}`
	√

Results	We have validated that each means of transport (`mao:transportFor`) had all its stations and their corresponding, we verify that all associated stations what quays are associated to a specific means.

Table 16: Validation query for a specific accessibility element: *hasLift*

Query 4.
`SELECT ?station ?Lift`
`WHERE {`
`?station mao:hasLift ?Lift}`
	×

Results	We obtained the stop places labelled with `mao:hasLift` from the Dataset_v2. Then we repeat the query from the Dataset_v1. Next we compared both results and we find that after adding the accessibility data from Dataset_v1 to Dataset_v2, the results from the Dataset_v2 are less triples than the results from the Dataset_v1: the name of the stop places are different between both datasets. We need to use specific techniques to match the name between of two datasets.

Table 17: Results to query for a specific accessibility element: *hasLift.*

From Dataset_v1:

```
<https://www.metromadrid.es//es/viaja_en_metro/red_de_metro/estaciones/Sol.html>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type><<http://com.vortic3.MANTO#stopPlace> .
<https://www.metromadrid.es//es/viaja_en_metro/red_de_metro/estaciones/Sol.html>
<http://xmlns.com/foaf/0.1/name> "Vodafone Sol" .
<https://www.metromadrid.es//es/viaja_en_metro/red_de_metro/estaciones/Sol.html>
<http://com.vortic3.MANTO#Transfer> " Cercanías Renfe" .
<https://www.metromadrid.es//es/viaja_en_metro/red_de_metro/estaciones/Sol.html>
<http://com.vortic3.MANTO#ofLine> _:genid1 .
<https://www.metromadrid.es//es/viaja_en_metro/red_de_metro/estaciones/Sol.html>
<http://com.vortic3.MANTO#hasLift> "TRUE" .
<https://www.metromadrid.es//es/viaja_en_metro/red_de_metro/estaciones/Sol.html>
<http://com.vortic3.MANTO#hasEscalator> "TRUE" .
```

From Dataset_v2:

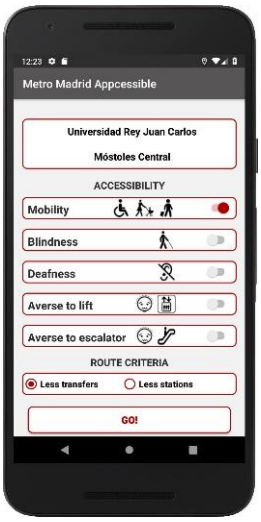
```
<http://com.vortic3.MANTO#est_90_58><http://www.w3.org/1999/02/22-rdf-syntax-
ns#type><http://com.vortic3.MANTO#stopPlace> .
<http://com.vortic3.MANTO#est_90_58><http://com.vortic3.MANTO#stopplaceCode> "58" .
<http://com.vortic3.MANTO#est_90_58>
<https://schema.org/name> "Puerta del Sol" .
<http://com.vortic3.MANTO#est_90_58><http://com.vortic3.MANTO#address> "Plaza de la Puerta del Sol
6" .
<http://com.vortic3.MANTO#est_90_58><http://com.vortic3.MANTO#TariffZone> "A" .
<http://com.vortic3.MANTO#est_90_58><http://www.geonames.org/ontology#timeZone> "Europe/Madrid" .
<http://com.vortic3.MANTO#est_90_58><http://www.w3.org/2003/01/geo/wgs84_pos#long> "-3.70331" .
<http://com.vortic3.MANTO#est_90_58><http://www.w3.org/2003/01/geo/wgs84_pos#lat> "40.41688" .
<http://com.vortic3.MANTO#est_90_58><http://com.vortic3.MANTO#wheelchaiAccess> "UNKNOWN" .
```

Table 18: Percentages of errors obtained from the Web source

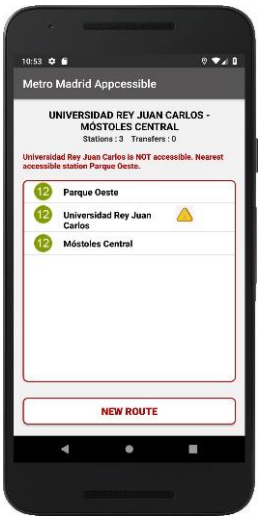
From	Description	Number	%
GTFS data source	stop_desc	1,360	97.39
	stop_url	1,360	
Multiple sources (Dataset_v1 and GTFS data integration)	Literal name of stop place	73	2.61

(ii) The integration of data is, in some cases, inconsistent: the station *names* from Dataset_v1 were different to the station *names* from the Transit feeds of Madrid public transport website and we did not match the data. It was, therefore, necessary to unify the station names (one of these errors is shown in Table 17). Some examples of this kind of errors were the names "Vodafone Sol" versus "Puerta del Sol" and "Plaza Castilla" versus "Plaza de Castilla". The number of errors encountered was 73 and the percentage of error from data integration is, therefore, 2.91%.

Table 18 provides a summary of the aforementioned percentages of errors. All the errors discovered were from the data sources and not a result of the data annotation process proposed in this work. In 97.39 % of the cases, the errors are due to data are not according to the GTFS specification. In 2.61 % of the cases, the errors discovered are related to the stop place names that



(a) Request a route



(b) Notify an event to the user

Figure 3: Metro Madrid Appcessible screen snapshots

(a) Request a route

(b) Details of the route

Figure 4: Metro Madrid Appcessible screen snapshots

appear differently written in GTFS and on the Metro Madrid website (for example, "Plaza Castilla" versus "Plaza de Castilla"). We have corrected the incorrect data and have attained high data quality, since no errors were found in our Dataset_v2. We cannot, however, guarantee the quality of the data if the sources change.

4.2 Validation of Usefulness by means of apps

We validated Dataset_v2 by developing two different Android apps, which use the dataset Dataset_v2 for different and specific purposes.

The first is an app denominated as Metro Madrid Appcessible that provides metro accessible routes according to the users' needs [5]. When the app starts up, it requests Dataset_v2 from our Coruscant server. The app calculates a route according to the users' needs. It is possible that no accessible route exists, but if one does, this app will be able to calculate it. The user interface in **Figure 3**a shows the information that the app requires in order to establish the route. First, the origin and the destination of the trip; second, the user's type of special need (defined by groups: people with special mobility needs, blind people, hearing-impaired people and different phobias); and third, a decision regarding the route (minimizing commutes or stations).

In this example, we have requested a route starting at the Universidad Rey Juan Carlos station. As indicated previously, the app then notifies a lift failure at this station and offers the closest accessible stations as an alternative (in this case, the closest accessible stations are "Parque Oeste" and "Mostoles Central"), as shown in **Figure 3**b.

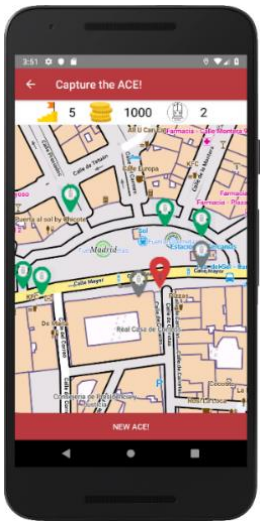


Figure 5. Access 'n'Go! screen snapshot

Figure 4 shows another example. In this case, we have requested a route starting at the Vodafone Sol station to Plaza de Castilla station considering a phobia as regards escalators (the user interface is shown in Figure 4a). In this case, a route exists. The details of this route are shown in Figure 4b.

We have tested the app with different test cases. These were selected in order to cover every kind of special need (mobility, blindness, deafness, aversion to lifts and aversion to escalators). All the test cases were satisfactory as regards providing an accessible route when possible.

The second is an Android app, based on gamification and crowdsourcing strategies, which is denominated as Access 'n' Go! [5]. This app obtains the public transport ACcessibility Elements (ACEs) from Dataset_v2 and represents them on a map. The users (who are, in fact, players) can then capture the ACEs, thus obtaining stations and lines in the public transport system, although the app also has other functionalities. ACEs are the entrances to a station (lifts, escalators, stairs or ramps) and have geographical coordinates (latitude and longitude), thus enabling their representation on a map. A player can capture an ACE when s/he is close to it. Figure 5 depicts the screen employed to capture ACEs. We consequently represent ACEs and the player's current position on the map with different colours: the player's current position is a red pin, the ACEs already captured are in grey, while the ACEs that have not been captured are green. The scores (points, levels and badges) are shown at the top of screen.

The results of the validation of this app are described in [5].

5 Conclusions and Future Work

The recent appearance of smart cities has led to a situation in which it will be necessary to achieve an inclusive society through the use of the new information and communication technologies, and one important challenge is to facilitate the mobility of all citizens. Open data practices facilitate new issues and services addressed to both final users and consumers.

The Google Transit Feed Specification, which defines a format for public transportation and associated geographic information, does not provide the constructions required to represent accessibility or special needs data at an adequate level of granularity.

In this paper, we, therefore, propose to combine semantic annotation techniques with public transport data (and their accessibility features) obtained from multiple Web sources. We have described a systematic method with which to unify and integrate diverse information so as to provide an open RDF public transport dataset. This could be employed in applications that will help the public move in a city.

Furthermore, we have verified the correctness of the data by using Dataset_v2 in two different apps. As Dataset_v2 includes the geographical information and accessibility characteristics of public transport networks, we can represent the data on a map and offer new accessibility information services to public transport users.

The first app, called Metro Madrid Appcessible, provides users with metro accessible routes according to their needs. Users can employ it to request a route, indicating their specific accessibility needs. The route is then represented on a map, and the user is guided along the route. The second app, called Access 'n' Go!, is based on gamification and crowdsourcing strategies. It obtains the accessibility elements (ACEs) from Dataset_v2, and represents them on a map. The players can then capture the ACEs in order to attain public transport stations and lines, among other functionalities. A player captures an ACE when s/he establishes whether or not an ACE is operational. Both apps have been validated [5], [5], and work properly with Dataset_v2.

As future work, we intend to improve the systematic process defined in this work in order to additionally integrate the information regarding the APIs of public transport networks. The process will also take into account the accessible routes generated by users with special needs, using crowdsourcing techniques. These routes will have to integrate these new issues into the final dataset.

Acknowledgements

This work is supported by the Access@City project (TIN2016-78103-C2-1-R), funded by the Spanish Ministry of Science, Innovation and Universities.

References

[1] Access2All, "Mobility Schemes Ensuring Accessibility of Public Transport for All Users", EU FP7 project, https://cordis.europa.eu/project/rcn/89881/factsheet/en, accessed 30th June, 2019.

[2] M. Belk, P. Germanakos, E. Papatheocharous, P. Andreou and G. Samaras, "Integrating Human Factors and Semantic Markups in Adaptive Interactive Systems". *Open Journal of Web Technologies (OJWT)*, vol. 1 no. 1. pp. 15-26, 2014, http://nbn-resolving.de/urn:nbn:de:101:1-2017052611313

[3] P. Cáceres, A. Sierra-Alonso, C.E. Cuesta, B. Vela and J.M. Cavero, "Modelling and Linking Accessibility Data in the Public Bus Network", *Journal of Universal Computer Science*, vol. 21, no. 6, pp. 777-795, 2015.

[4] P. Cáceres, A. Sierra-Alonso, C. E. Cuesta, J. M. Cavero and B. Vela, "Towards Smart Public Transport Data: A Specific Process to Generate Datasets Containing Public Transport Accessibility Information" *In Proceedings of the Second International Conference on Universal Accessibility in the Internet of Things and Smart Environments*, pp. 66-71, 2018.

[5] P. Cáceres, C. E. Cuesta, A. Sierra-Alonso, B. Vela, J. M. Cavero and M.A. Garrido. 2019. "Even Smarter Data: Using Crowdsourcing to Improve Accessibility in Real-Time", *In the Fourth*

International Conference on Universal Accessibility in the Internet of Things and Smart Environments, pp. 26-32, 2019.

[6] P. Cáceres, C. E. Cuesta, B. Vela, J. M. Cavero and A. Sierra-Alonso, "Smart Data at Play: Improving Accessibility in the Urban Transport System", *Behaviour and Information Technology*, https://doi.org/10.1080/0144929X. 2019.1652852. pp.1-14, Aug. 2019.

[7] C. Ding, M. Wald and G. Wills, "Using Open Accessibility Data for Accessible Travelling", *in Proceedings of the 14th International Conference on Computers Helping People with Special Needs*, France, pp. 1-4, July 2014.

[8] C. Ding, M. Wald and G. Wills, "Open Accessibility Data Interlinking", *in Proceedings of the 14th International Conference on Computers Helping People with Special Needs*, France, pp. 73-80, July 2014.

[9] C. Ding, M. Wald and G. Wills, "Linked Data-Driven Decision Support for Accessible Travelling", *in Web for All – Doctoral Consortium*, Italy, 2015.

[10] CIVITAS Initiative, "Cleaner and Better Transport in Cities", http://www.civitas.eu, 2012, accessed 20th July, 2019.

[11] I. Davis, "TRANSIT: A Vocabulary for Describing Transit Systems and Routes", http://vocab.org/transit, accessed 19th May 2019.

[12] European Commission, "OASIS, Open Architecture for Accessible Services Integration and Standardisation", http://www.oasis-project.eu/, accessed 19th May 2019.

[13] S. Gaggi, T. Fluhrer and T. Janitzek, "Innovation in Urban Mobility: Policy Making and Planning. Directorate- General for Mobility and Transport". Luxembourg: European Union, 2013, https://www.kowi.de/Portaldata/2/Resources/fp/tripurban-mobility.pdf, accessed 30th June, 2019.

[14] Y. Hidalgo Delgado, R. Estrada Nelson, B. Xu, B. M. Villazon Terrazas, A. Leiva Mederos and A. Tello, "Methodological Guidelines for Publishing Library Data as Linked Data", *in Proceedings of the 2nd Conference on Information Systems and Computer Science*, pp. 241-246, 2017.

[15] B. Hyland and D. Wood, "The Joy of Data - A Cookbook for Publishing Linked Government Data on the Web", *Springer*, 2011.

[16] IFOPT, "Identification of Fixed Objects in Public Transport" Standard CEN/TC 278, EN 28701, European Committee for Standardization, 2012.

[17] C. Keller, S. Brunk and T. Schlegel, "Introducing the Public Transport Domain to the Web of Data" *in Proceedings of the 15th International Conference on Web Information Systems Engineering*, pp 521-530, 2014.

[18] Y. Li, E.A. Draffan, H. Glaser, I. Millard, R. Newman, M. Wald, G. Wills and M. White, "RailGB: Using Open Accessibility Data to Help People with Disabilities", *International Semantic Web Conference*, Boston, USA.11 - 15 Nov 2012.

[19] Mediate, "Methodology for Describing the Accessibility of Transport in Europe", EU FP7 project https://trimis.ec.europa.eu/, 2011.

[20] NAPTAN, "The National Public Transport Access Node Database of the United Kingdom", http://naptan.app.dft.gov.uk/datarequest/help, accessed 19th May 2019.

[21] B. Najdenov, G. Petkovski, M. Jovanovik, R. Stojanov and D. Trajanov, "Automated Linked Data Generation from the Transport Administration Domain", *in Proceedings of 23rd Telecommunications Forum*, pp.827-830, 2015.

[22] OpenTRipPlanner, www.opentripplanner.org/, accessed 20th July, 2019.

[23] J. Plu and F. Scharffe, "Publishing and Linking Transport Data on the Web: Extended Version", *in Proceedings of the First International Workshop on Open Data*, pp. 62-69, 2012.

[24] C. Prandi, "Accessibility and Smart Data: The Case Study of mPASS", *in Proceedings of 13th International Web for All Conference*, Seoul, Republic of Korea, pp. 9-10, April 2014.

[25] B. D. Romero, M. J. Rodríguez, M. V. Hurtado, L. Ramos and H. M. Haddad, "Accessibility Ontology", https://sourceforge.net/projects/accessibilitydomainowlmodule/, accessed 19th May 2019.

[26] V. Santiago Pineda and J. Thurston, "Smart Cities for All: a Global Strategy for Digital Inclusion", Proposal by G3ict and World Enabled organizations, 2016.

[27] Transmodel, standardization, "Transmodel, Road Transport and Traffic Telematics. Public Transport. Ref. Data Model, EN 12896", http://www.transmodel.org/en/cadre1.html, accessed 19th May 2019.

[28] USA Government, "The home of the U.S. Government's Open Data", https://www.data.gov/, accessed 20th May 2019.

[29] B. Vela., J.M. Cavero, P. Cáceres and C.E. Cuesta, "A Semi-Automatic Data-Scraping Method for the Public Transport Domain", *IEEE Access*, Vol. 7, pp. 105627-105637, 2019.

[30] M. Wiethoff and S. Sommer, "Specification of Information Needs for the Development of a Mobile Communication Platform to Support Mobility of People with Functional Limitations", *in Proceedings of Universal Access in Human Computer Interaction Ambient Interaction*, pp.595–604, 2007

AUTHOR BIOGRAPHIES

Paloma Cáceres García de Marina received an MSc in Computer Science from the Universidad Politécnica of Madrid (Spain) in 1993 and a PhD in Computer Science from Rey Juan Carlos University (Madrid, Spain) in 2006. She is an associate professor at Rey Juan Carlos University. Her research interests are software engineering, web engineering, model driven development, data science, open and linked data, the semantic Web, and transport and accessibility. She is the author and co-author of several national and international papers related to these areas.

Almudena Sierra-Alonso received her Ph.D. in Computer Science from the Universidad Politécnica de Madrid, in 2000. She is an associate professor in the Computer Science and Statistics Department of Rey Juan Carlos University. Her research interests are focused on software engineering, data science, open and linked data, the semantic Web, and transport and accessibility. She is the author and co-author of several papers related to software engineering, the semantic web and teaching in Engineering Education (in areas such as operating systems and software engineering).

Belén Vela has an MSc in Computer Science (Carlos III University) and a PhD in Computer Science (Rey Juan Carlos University). She is currently an Associate Professor at the School of Computer Science (Rey Juan Carlos University). She belongs to the Vortic3 Research Group. Her research interests are focused on Software Engineering, Information System Engineering, Data Science, Open Data, Databases (NoSQL), Model Driven Development, Transport, Accessibility and Scientometrics. She has had numerous papers published in prestigious journals and conferences.

José María Cavero has an MsC in Computer Science from the Polytechnic University of Madrid and a PhD in Computer Science from Rey Juan Carlos University. He is an Associate Professor at the School of Computer Science (Rey Juan Carlos University). He belongs to the Vortic3 Research Group. His research interests are focused on Software Engineering, Information System Engineering, Data Science, Open Data, Databases (NoSQL), Transport, Accesibility and Scientometrics. He has several papers published in prestigious journals and conferences.

Miguel Ángel Garrido Blázquez received an MSc in Computer Science from Rey Juan Carlos University (Madrid, Spain) in 2014. He has worked as a computer technician in several entities for 10 years, and is currently an assistant professor and a Computer Science PhD student at Rey Juan Carlos University. His research areas are web engineering, model driven development, the semantic web, linked open data, transport and accessibility.

Carlos E. Cuesta is an Associate Professor of Software Engineering at Rey Juan Carlos University (Madrid, Spain). He has a PhD in Information Technologies (2002) from the University of Valladolid. He belongs to the VorTIC3 Research Group. His main research area is Software Architecture, including such topics as Self-Adaptive Systems and Systems-of-Systems, and relative to such fields as Concurrency Theory, Formal Methods, Agreement Technologies, Intelligent and Distributed Systems or Data Engineering. He is currently the Program Chair of the 12th European Conference on Software Architecture (ECSA 2018). He has published in premier conferences and journals, such as the International Journal of Information Technology and Decision Making, or Future Generation Computer Systems.