Open Public Transport Data in Macedonia

Conference Paper · April 2014

DOI: 10.13140/RG.2.2.29985.20326

CITATIONS
3

READS
4,394

4 authors:

Elena Mishevska
Ss. Cyril and Methodius University in Skopje
1 PUBLICATION **3** CITATIONS
SEE PROFILE

Bojan Najdenov
Ss. Cyril and Methodius University in Skopje
8 PUBLICATIONS **38** CITATIONS
SEE PROFILE

Milos Jovanovik
Ss. Cyril and Methodius University in Skopje
58 PUBLICATIONS **184** CITATIONS
SEE PROFILE

Dimitar Trajanov
Ss. Cyril and Methodius University in Skopje
151 PUBLICATIONS **537** CITATIONS
SEE PROFILE

Open Public Transport Data in Macedonia

Elena Mishevska, Bojan Najdenov, Milos Jovanovik, Dimitar Trajanov
Faculty of Computer Science and Engineering
University Ss. Cyril and Methodius
Skopje, Republic of Macedonia

***Abstract*—The need to represent data on the Web in a way that will make it easier to manage, has led to new solutions for data representation, visualization, storage and querying. The concepts of Open Data, Linked Data and the Semantic Web offer a significant improvement in information and data dissemination. These concepts aim towards making data on the Web machine-readable and enable interlinking between data from different datasets, published on different locations. This allows easier data retrieval by software agents, and enables use-case scenarios which are unavailable over isolated data silos. On the other hand, personal time management and daily commute navigation in urban areas are one of the biggest influencers on the quality of life of a person. Public transport data has high value for citizens and generates numerous use-cases. In this paper, we describe the process of obtaining data from the public transport company JSP Skopje, transforming them into the standardized Google Transit Feed Specification[1] format, enhancing them and creating 4 star Open Data. We reused the Transit Ontology[2] and the W3C Geospatial Vocabulary[3], and developed our own complementing ontology for annotation purposes. We published the generated RDF datasets in order to support the provided use-case scenarios from this domain via a public SPARQL endpoint.**

Keywords—Public Transport; Open Data; GTFS; RDF; Ontologies;

I. INTRODUCTION

The Open Data concept represents the idea that data generated by the governments, their institutions and other public entities, and which is public by its nature, should be published in an open, raw and machine-readable format. This data can then be used, reused, republished and redistributed, generally in applications which leverage the value of the data [1]. This allows for a variety of useful applications to be built with the published datasets, using and combining data in various ways. Further linking of this data with data from other datasets, extends the possibilities, by providing the opportunity to use data and information relevant for the use-case, but not part of the original dataset. Linked Data is about employing the Resource Description Framework (RDF) and the Hypertext Transfer Protocol (HTTP) to publish structured data on the Web and to connect data between different data sources, thus effectively allowing data from one dataset to be interlinked with data in another dataset [2].

Public transport is a very important issue in the lives of citizens. Therefore, public transport data caries values which are of great importance for the levels of quality of life. This was our main motivation to work with public transport datasets and try to create data representations which provide various use-case scenarios via REST services.

Another thing which motivated us to work on this subject was the Helsinki Open Transport Data Manifesto[4], which empowers the free flow of transport data across Europe and puts focus on the opportunities and benefits of opening up and sharing this reach resource to the stakeholders.

A. The Star Rating System for Data

In order to classify the data published on the Web by its availability and its usefulness, there is a standard star rating system. According to the rating system[5], every information which has been published online can be considered as Open Data, and is given a one star rating. This usually includes images, PDF documents, and other documents types as well.

Making the data machine readable earns it a two star rating. This usually includes Microsoft Excel spreadsheets. Publishing data in non-proprietary format, such as CSV, earns it a three star rating. Published Open Data datasets, which use Semantic Web standards (RDF, RDFS, OWL) for annotation of the entities, earn four stars. The last one, the fifth star, is given for datasets which contain links towards other, already published dataset on the Web, in order to provide context [3].

Almost all of the Google Transit data published in the GTFS format are Open Data by these definitions, since they are available online for public use. However, since the GTFS format of the data obligates data to be represented in strictly formed CSV files (explained in more details further in the paper), no effort has been made to bring public transportation data up in the star rating system.

[1] https://developers.google.com/transit/gtfs/reference
[2] http://vocab.org/transit/terms/
[3] http://www.w3.org/2005/Incubator/geo/XGR-geo/
[4] http://www.epsiplatform.eu/transport
[5] http://5stardata.info/

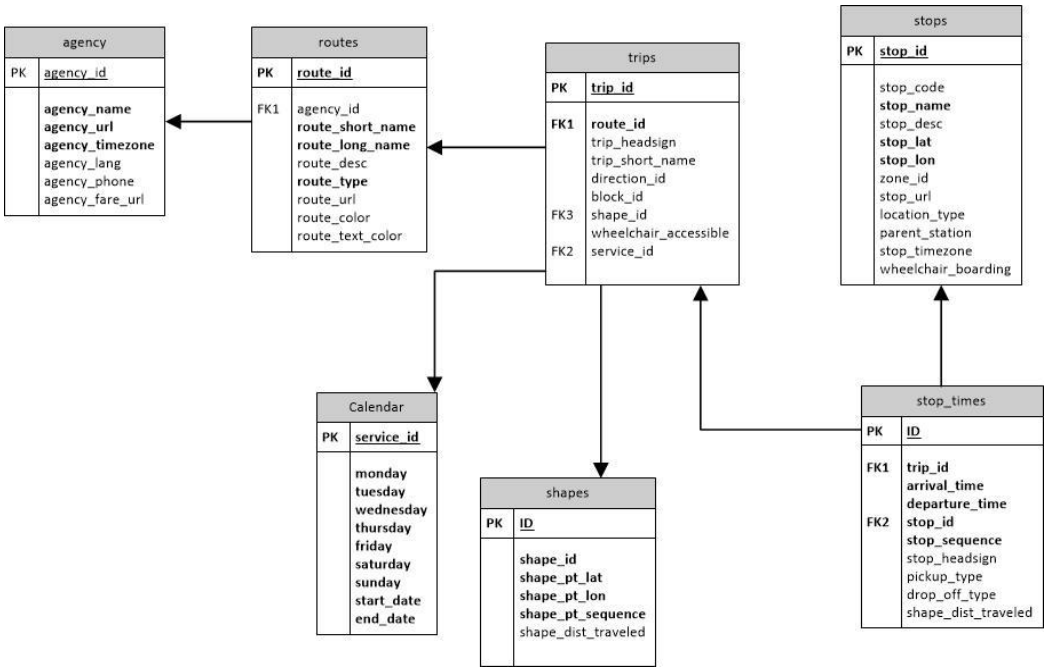

Figure 1. The GTFS Schema for the data from JSP Skopje.

We decided to leverage local public transportation data from Macedonia as high as possible in the rating system, by reusing existing ontologies and combining them with our own ontology for the annotation purposes. We used data collected from the JSP Skopje website, which has been transformed by another project into GTFS format in order to be used with the Google Transit system. The dataset is described in more details further in the paper.

II. Related Work

The Google Transit system is widely used around the globe from many transport companies from the six continents that use different kinds of transportation. Google Transit is available for companies from USA, Austria, Belgium, Australia, Ghana, Nigeria, Bulgaria, etc[6]. However, very little effort has been made to semantically annotate any of the Google Transit data so far, or any other transit data, not necessarily in the GTFS format.

Data from the New York Subway in GTFS format have been published as RDF files, but the files, as they are presented, have not been created using any ontologies and are not linked anywhere on the Web[7]. Also, a tool for transforming GTFS data into RDF exists. The tool is written in Perl, and uses a Turtle[8] syntax to map the RDF files. An Android application, called GetThere[9], which uses Linked Open Data has been developed in order to provide information about the location of busses in rural areas in the UK [4].

As for transport open and linked data, there are only a few datasets from Great Britain and France [5], but they concentrate on the physical locations of the stops or the connectivity of the different stops or cities. For now, no linked data datasets of any transit agency are available on the LOD cloud[10].

III. Google Transit Feed Specification

The data consists of multiple CSV format files (with the .txt extension), each representing a different piece of data, written in a strictly specified form. Not all of the schema tables are obligatory, and not all of the schema tables exist for different publishers [6].

JSP Skopje[11] is a public transportation company from Skopje, which provides public transit services on the territory of Macedonia's capital. The transport is done by busses which operate on different routes following predefined schedules. JSP Skopje publishes the bus schedule on their website in standard HTML format.

As part of another project at our Faculty, the data from the JSP Skopje website have been collected and transformed into the GTFS format, for the purposes of them being used as part of the Google Transit system. The JSP Skopje GTFS dataset consists of seven of the standard GTFS tables (Figure 1):

- agency - contains information about one or more transit agencies.
- stops - contains information about all the stops.
- routes - contains information about the routes of the transit agency.

[6] http://maps.google.com/landing/transit/cities/index.html
[7] http://www.cs.sunysb.edu/~pfodor/new_york_subway_data/
[8] http://www.w3.org/TR/turtle/
[9] http://gettherebus.com/

[10] http://lod-cloud.net/
[11] http://jsp.com.mk/

- trips - contains sequences of two or more stops that occur at a specific time.
- stop_times - contains vehicle arrival and departure times from individual stops for each trip.
- calendar – contains schedule information.
- shapes - the spatial representation of a route alignment so it can be accurately drawn on a map [7]

IV. TRANSFORMING THE GTFS DATA FROM JSP SKOPJE INTO FOUR STAR OPEN DATA

Besides having the transport data from JSP Skopje in GTFS format, we believe that creating a semantic annotation of the data and publishing it as four star Open Data on a SPARQL endpoint, will make it easier to use. This endpoint would serve as a REST service for developers to effectively use from their applications, built over the dataset and providing public transit system services for the city of Skopje.

In order to start the annotation process, we needed a suitable ontology.

A. Ontology

The most common approach and the best practice in providing an ontology, is reusing an already developed one. Additionally, one usually has to creating his own ontology for the properties which do not exist in other ontologies, in order to start with the annotation process.

Our search for a suitable ontology lead us to the Transit Ontology[12], developed for similar dataset to the one we had, a Google Transit dataset, but developed for a different scenario which was a bit different than ours. Regardless, the ontology provided enough properties we could match with our data. The ontology provided us with some of the classes and the properties we needed, and what was left for us was to find a solution for the remaining properties. Therefore, we reused some properties from the W3C Geospatial Ontology[13], and we created an ontology which covered the remaining properties.

Here, we provide a listing of the classes and properties we used from the Transit Ontology. The diagram of the ontology is shown on Figure 2.

The Transit Ontology we reused provided us the following classes:

- Transit Route – a public transportation route.
- Transit Stop – a location where passengers board or disembark from a transit vehicle.
- Transit Agency – an organization that oversees public transportation for a city or region.

The Object properties we used to link data from the previous classes were:

- Route – the route the trip of interest uses;
- Stop – a physical location connected to a service stop;
- Agency – the agency that operates the route of interest.

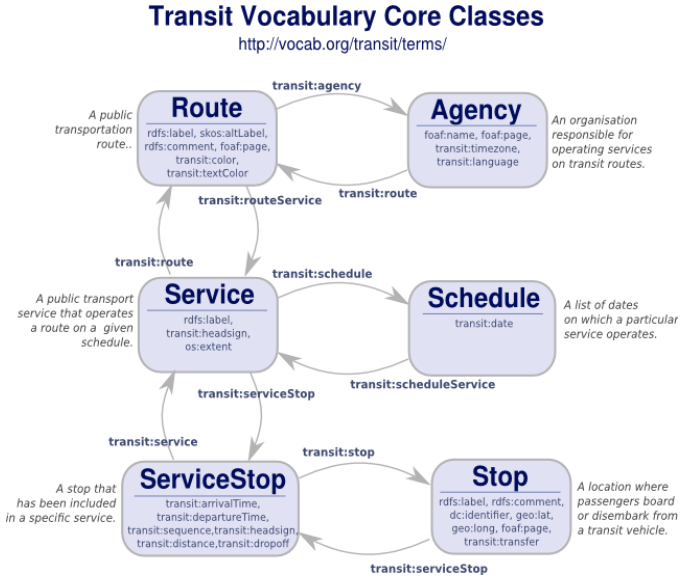

Figure 2. Transit Ontology Diagram.

The DataType properties we reused were:

- Timezone - the time zone where a person or organization is located.
- Language - the primary language used by a person or organization.
- Sequence - a sequence number for a stop along a route or service.
- Distance - the distance of this service stop from the first stop in sequence.
- Headsign - text that appears on a sign that identifies the service's destination to passengers.
- Color - a color associated with this route.
- Text Color - a legible color for text drawn against a background of the color associated with a route.

For the latitude and longitude data for the location of the stops and the locations of the different stop points, we used the 'latitude' and 'longitude' properties of the W3C Geospatial Ontology.

We defined the rest of the classes and properties in our own ontology, GTFS-ext, which extends the Transit Ontology and provides us with all of the classes and properties we needed in order to fully map the available GTFS data from JSP Skopje.

The Classes we added to our ontology are given in Table 1. The Object properties that we defined in order to link the classes can be seen in Table 2. Additionally, we created all the necessary Datatype properties which were not defined in the Transit Ontology, but were necessary for annotating our dataset, according to the GTFS Schema from Figure 1.

[12] http://vocab.org/transit/terms/
[13] http://www.w3.org/2005/Incubator/geo/XGR-geo-ont/

Table 1. Classes introduced in our GTFS-ext Ontology.

Class	Description
Shape	Specification about how the lines are represented on the map
Trip	Sequence of two or more stops that occur at a specific time
stop_times	Arrival and departure times from an individual stop
calendar	Provides schedule for a specific service

Table 2. Object Properties from our GTFS-ext Ontology.

Object Property	Description
service	References the service of a specified trip
st_times	Connects a trip with the specific departure and arrival times
shape	Shape associated with the referenced trip

B. Mapping the Data from CSV to RDF

The next step was mapping the data and transforming it from CSV to RDF. In order to achieve this, we used a Virtuoso Universal Server[14] instance, which provides mechanisms for transformation and management of various types of data. It serves as a Linked Data server and allows local and remote data querying with the Semantic Web query language, SPARQL. This is enabled over a public SPARQL endpoint which can be used as a REST service.

The mapping process was conducted into several stages. First, we imported the CSV files into relational databases in Virtuoso. Then, using R2RML[15] – a mapping language for transforming RDB data into RDF data – we created RDF Views over our relational databases.

The R2RML mapping was done using mapping files which contain information about the transformation of the RDB tables into RDF triples, each contains a subject, a predicate and an object. Each row of the relational database represents a unique entity, and each column a new triple with the entity as a subject. We used the previously discussed ontologies to annotate the data. Most of the entities from the tables are identified with the identifiers which were part of the GTFS CSV data. A small portion of the data (e.g. stop_times) is identified by the row number of the input file.

After all of the data was mapped, we ended up having seven individual graphs, one for each of the tables (Figure 1) containing the data. The next thing we needed to do was to link the graphs, i.e. create the links that connected the different pieces of information into usable information. RDF links take the form of RDF triples, where the subject of the triple is a URI reference in the namespace of one dataset, while the object of the triple is a URI reference in the other [8]. We achieved that using the SPARQL endpoint from the Virtuoso server and by using the SPARQL query language we combined and created the appropriate links for the different graphs. We created the following links:

- Linked the ‘routes’ graph with the ‘agency’ graph using the ‘agency’ object property.
- Linked the ‘trips’ graph with:
  - the ‘routes’ graph, using the ‘route’ property.
  - the ‘calendar’ graph, using the ‘service’ property.
  - the ‘shape’ graph, using the ‘shape’ property.
  - the ‘stop_times’ graph, using the ‘st_times’ property.
- Linked the ‘stop_times’ graph with the ‘stops’ graph using the ‘stop’ object property.

With that, we finished the process of semantic annotation of the JSP Skopje transit data.

The idea of linking the resulting four star RDF dataset with another similar datasets, in order to provide a wider range of additional use-cases, led to no success. We were unable to find other semantically annotated transit datasets related to ours in any way, which would make sense to interconnect.

V. USE-CASES

The basic idea behind the semantic annotation and leveraging the public transport data from JSP Skopje to four star Open Data, is creating a publicly available dataset which could be easily used and would provide a large variety of use-case scenarios involving public transport.

With our transformed dataset, we can provide information about the stop and the time the travelling party should go to, in order to get a bus that could take them to the desired location within the city of Skopje, what is the stop they need to get off, as well as the arrival time on that stop. This all would represent a search in our graph to find the departures on a specific stop, at a specific time which will match a route that suits our needs.

A. Use-Case 1

If, for example we want to find the departure times of a given bus route in a given period of time during a specific time of the year, we could generate a use-case similar to the following: we look for departure times of the route ‘R15’, from the ‘Ново Лисиче’ station, running between 09:00 – 10:00 AM, during weekdays in winter. For this, we can use the following SPARQL query:

```
prefix ont:
<http://linkeddata.finki.ukim.mk/lod/ontology/transit-
ont#>
prefix transit: <http://vocab.org/transit/terms/>
select distinct ?departure
where {
        graph
        <http://linkeddata.finki.ukim.mk/lod/data/routes#>
        { ?r ont:route_id "R15" }
        graph
<http://linkeddata.finki.ukim.mk/lod/data/calendar#>
        { ?s ont:service_id "DELNIK_ZIMEN" }
```

[14] http://virtuoso.openlinksw.com/
[15] http://www.w3.org/TR/r2rml/

```
        graph
<http://linkeddata.finki.ukim.mk/lod/data/trips#>
        { ?x transit:route ?r ; ont:service ?s ;
             transit:headsign "Карпош 4" ;
             ont:st_times ?st.
        }
        graph
<http://linkeddata.finki.ukim.mk/lod/data/stop_times#>
        { ?st ont:departure_time ?dep;
              transit:sequence 1. }
        FILTER regex(?dep,"(^09:)|(^9:)")
}
```

The results from the query are shown in Table 3.

This query obtains the URI of the route with the 'route_id' equal to 'R15' from the 'routes' graph and the URI of the service with the 'service_id' equal to 'DELNIK_ZIMEN' from the 'calendar' graph. Then, from the 'trips' graph, it finds the trips which correspond to the route and schedule we previously obtained and which have the 'transit:headsign' property set to 'Карпош 4' (the direction of the bus). After that, it finds the records from the 'stop_times' graph which correspond to the selected trips using the 'tr:st_times' property. Finally, it filters the properties that fulfill the condition that the departure time is between 09:00 - 10:00 AM.

Table 3. Results from the SPARQL query for Use-Case 1.

DEPARTURE
"09:55:00"

B. Use-Case 2

In another use-case scenario, we may want to find all the routes that go through a specific bus stop. This will require the usage of four graphs: the 'stops' graph, to find the stop URI, the 'stop_times' graph to find arrivals and departures from the specific stop and the trips that use that stop, the 'trips' graph to find the routes associated with the specified trip, and finally, the 'routes' graph to find the route's name. Let the stop we use be a more frequent one, for example the "МАЛ ОДМОР" stop.

To select the routes which go through the "МАЛ ОДМОР" stop along with their names, we use the following SPARQL query:

```
prefix transit: <http://vocab.org/transit/terms/>
prefix ont:
<http://linkeddata.finki.ukim.mk/lod/ontology/transit-
ont#>
select distinct ?r ?n
where {
        graph
<http://linkeddata.finki.ukim.mk/lod/data/stops#>
        { ?o ont:name "МАЛ ОДМОР" }
        graph
<http://linkeddata.finki.ukim.mk/lod/data/stop_times#>
        { ?s transit:stop ?o. }
        graph
<http://linkeddata.finki.ukim.mk/lod/data/trips#>
        { ?t ont:st_times ?s ;
             transit:route ?r.
        }
        graph
<http://linkeddata.finki.ukim.mk/lod/data/routes#>
        { ?r ont:name ?n }
}
```

The results from this query are shown in Table 4.

We will see how the SPARQL query achieves the result step by step. First, starting from the 'stops' graph, using the 'ont:name' property, it selects the appropriate URI of the stop and uses it in the 'stop_times' with the 'transit:stop' property to select the arrivals and departures on the selected stop. The 'transit:stop' property is a property of the transit ontology we reused. After that, the query passes the URIs of the arrival and the departure values to the 'trips' graph, in order to find the trips that arrive at certain times at the stop, using the 'ont:st_times' property and using the 'transit:route' properties selects the routes associated to the trips selected with the 'ont:st_times' property. Finally, in the 'routes' graph, the query uses the 'ont:name' property to find the already selected routes' names and presents the final results.

Table 4. Results from the SPARQL query for Use-Case 2.

R	**N**
http://vocab.org/transit/terms/Route/R4	"11 Октомври - Нас. Хром"
http://vocab.org/transit/terms/Route/R2	"Сарај - Автокоманда"
http://vocab.org/transit/terms/Route/R7	"Нас. Лисиче - Карпош 3"
http://vocab.org/transit/terms/Route/R15	"Ново Лисиче - Карпош 4"
http://vocab.org/transit/terms/Route/R22	"Транспортен центар - Волково"
http://vocab.org/transit/terms/Route/R59	"Карпош 3 - Гробишта Бутел"
http://vocab.org/transit/terms/Route/R19	"Шуто Оризари - Карпош 4"
http://vocab.org/transit/terms/Route/R24	"Кисела Вода - Тафталиџе"

C. Use-Case 3

We can have a scenario in which we want to count the number of bus trips which occur at a selected bus stop, during a specific schedule. For example, we can take 'NEDELA_ZIMEN' as a schedule, and 'ПАЛИМА' as a bus stop. This way, we will count the number of buses trips which will pass at this bus stop, during Sundays in winter.

The SPARQL query for this scenario is:

```
prefix ont:
<http://linkeddata.finki.ukim.mk/lod/ontology/transit-
ont#>
prefix transit: <http://vocab.org/transit/terms/>
select count(distinct ?t) as ?count
where {
        graph
<http://linkeddata.finki.ukim.mk/lod/data/calendar#>
        { ?s ont:service_id 'NEDELA_ZIMEN' }
        graph
<http://linkeddata.finki.ukim.mk/lod/data/trips#>
        { ?t ont:service ?s ; ont:st_times ?st . }
        graph
<http://linkeddata.finki.ukim.mk/lod/data/stop_times#>
        { ?st transit:stop ?stop }
```

```
        graph
<http://linkeddata.finki.ukim.mk/lod/data/stops#>
        { ?stop ont:name "ПАЛМА" }
}
```

The result from the query is shown in Table 5.

Table 5. Result from the SPARQL query for Use-Case 3.

COUNT
193

We could run a similar query, but for the Saturday winter bus schedule, and compare the results with the one in Table 5. The SPARQL query for this would only require us to the schedule ('service_id' property) to "SABOTA_ZIMEN". The result of this query is Count: 275, which leads us to the conclusion that public transport bus lines in Skopje are more frequent on Saturdays than Sundays, which is expected.

D. Public Data Endpoint

It is also important to notice that these SPARQL queries can be sent to a SPARQL endpoint[16] on our live Virtuoso instance, in a REST service manner. This means that applications which would potentially use this dataset can query the data with simple HTTP GET requests and obtain the data in a variety of RDF and non-RDF formats, such as RDF/XML, Turtle, JSON-LD, RDF/JSON, N3, JSON, CSV, HTML, etc.

The general format of the HTTP calls is:

```
http://linkeddata.finki.ukim.mk/sparql?
query=SPARQLQUERY&format=FORMAT
```

VI. CONCLUSION AND FUTURE WORK

The Open Data concept holds the key to organizing and collecting information from the public sector, and using it in various ways and use-case scenarios. By publishing and linking data in the right way, we can gain more useful information and extract properties which do not even have to exist in our dataset, but can carry information of enormous relevance. And today, data represents the new oil for the industry [9].

In a similar manner, publishing and interlinking transportation data can take public transportation, trip planning, even sightseeing to a whole new level.

In the paper we gave an overview of the process of transforming the public transport data from JSP Skopje, into four star Open Data. We worked with big datasets, and created RDF graphs which contain 6.005.619 triples, out of which 5.227.956 are from the 'stop_times' graph, 725.758 are from 'trips', 46.914 are from 'shapes', 4.649 are from 'stops', 170 are from 'routes', 165 are from 'calendar' and only 7 are from the 'agency' graph.

We also provided example use-cases, in hope to encourage multiple stakeholders to start publishing Open Data from the public transportation sector and also, to start thinking about creative ways to use the available data. The annotated data and the use-case scenarios are accessible via HTTP GET requests to our public SPARQL endpoint.

In the future, we would continue our work in this sector and semantically annotate datasets from more transit agencies in Macedonia, both intercity and international ones. We would interlink them to provide more useful use-case scenarios which would allow trip planning throughout the country and would create an opportunity for developing application which could implement the features provided by the interlinked datasets. Hopefully, this will raise the awareness of the possibilities the usage of Open Data provides, and especially Open Transport Data.

On the other hand, we also hope to encourage more transit agencies, not only from Macedonia, but also from all around the world, to publish semantic annotated data, which will allow interlinking the data, leading to better usage of the published data and allowing the development of even more powerful solutions and more useful applications.

ACKNOWLEDGMENT

The work in this paper was partially financed by the Faculty of Computer Science and Engineering, at the Ss. Cyril and Methodius University in Skopje, as part of the research project "Semantic Sky 2.0: Enterprise Knowledge Management".

REFERENCES

[1] T. Berners-Lee, N. Shadbolt, "There's gold to be mined from all our data", The Times, 2012.

[2] C. Bizer, T. Heath, K. Idehen, T. Berners-Lee, "Linked Data on the Web", LDOW, 2008.

[3] M. Jovanovik, B. Najdenov, D. Trajanov, "Linked Open Drug Data from the Health Insurance Fund of Macedonia", 10th Conference for Informatics and Information Technology (CIIT), 2013.

[4] D. Corsar, P. Edwards, C. Baillie, M. Markovic, K. Papangelis, J. Nelson, "GetThere: A Rural Passenger Information System Utilising Linked Data & Citizen Sensing", International Semantic Web Conference (ISWC), 2013.

[5] J. Plu, F. Scharffe, "Publishing and linking transport data on the Web", First International Workshop On Open Data (WOD), 2012.

[6] N. Kizoom, P. Miller. A Transmodel based XML schema for the Google Transit Feed Specification With a GTFS / Transmodel comparison. Crown, 2008.

[7] A. Antrim, S. J. Barbeau, "The Many Uses of GTFS Data – Opening the Door to Transit and Multimodal Applications", ITS World Congress (ITSWC), 2013.

[8] C. Bizer, T. Heath, T. Berners-Lee, "Linked Data - The Story So Far", Special Issue on Linked Data, International Journal on Semantic Web and Information Systems, 2009.

[9] V. Kundra, "Digital Fuel of the 21st Century: Innovation through Open Data and the Network Effect", Joan Shorenstein Center on the Press, Politics and Public Policy, 2012.

[16] http://linkeddata.finki.ukim.mk/sparql