Automated Linked Data Generation from the Transport Administration Domain

Conference Paper · November 2015

DOI: 10.1109/TELFOR.2015.7377593

CITATIONS
2

READS
151

5 authors, including:

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

Goran Petkovski
Ss. Cyril and Methodius University in Skopje
6 PUBLICATIONS **2** CITATIONS
SEE PROFILE

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

Riste Stojanov
Ss. Cyril and Methodius University in Skopje
33 PUBLICATIONS **97** CITATIONS
SEE PROFILE

Automated Linked Data Generation from the Transport Administration Domain

Bojan Najdenov, Goran Petkovski, Milos Jovanovik, Riste Stojanov and Dimitar Trajanov

***ABSTRACT* — The Linked Data approach in data publishing allows the users and their data-driven applications to have broader use cases which encompass various data sources, either publicly available on the Web, or in private repositories. The use of W3C standards in publishing such data enables uniform access across platforms.**

Transport information today has higher importance to the citizens and society than ever and accessing the right information at the right time can improve the quality of everyday life for many people in the world.

In this paper, we describe our approach of building a system for automated Linked Data generation from the transportation domain. We used the data from the Swedish Transport Administration (STA) as a specific case study. For the purpose of RDF annotation, we developed the Transport Administration Ontology (TAO). The resulting five-star data enables advanced use case scenarios over the original STA data, which we also demonstrate with our web application.

***Keywords* — Automated Systems, Linked Data, Open Data, Transport Administration Ontology, Swedish Transport Administration.**

1. INTRODUCTION

We live in a world where data is omnipresent and means everything to us as individuals, but also as parts of organizations and societies [1]. The quantity of the data is on the rise as never before, while the future trends of growth are perceived only to continue increasing [2]. As the volume of the data increases in general, companies and public institutions promote transparency of their line of work by opening up their data and sharing it with the people in various formats over the Web. The large amount of data available on the Web motivates the research of new data management, storage and access techniques to be practiced on distributed datasets over the existing Web infrastructure [3].

The main goal of the Web has always been to present information in a which is understandable to humans, despite the fact that machines also use the Web to communicate through it. The Semantic Web on the other hand (Web 3.0), can be thought of as a web of data that is annotated with Semantic Web standards as RDF and OWL and is also interlinked based on the meaning [4][5]. In addition to that, a special accent is placed on the machine readability and interoperability which leads to providing endless use case scenarios for both humans and machines [6][7].

Knowing how important the quality and availability of transportation information is, in this paper we introduce a system for automated Linked Data generation and publishing, and use the Swedish Transport Administration (STA) as a case study. In order to annotate the transport data, we created the Transport Administration Ontology (TAO) and used it to make a semantic annotation of the STA's data, thus raising its quality to 5-Star Linked Open Data[2]. Furthermore, we created a web application where we demonstrated advanced use case scenarios that can be performed over the STA's data complimented with data from the LOD Cloud.

Our paper is organized as follows: in Section 2 we go over related scientific projects to the domain of Linked Data about transportation; in Section 3 we describe the automated system for gathering, transforming data and linking it to the LOD Cloud; additionally, in Section 4 we present example use case scenarios that can be performed over the linked datasets; finally, in the last section we conclude our work.

2. RELATED WORK

For the purpose of providing greater details of the problem area our research work is positioned in, we will go through several existing projects concerning Linked Open Data in relation with transportation as our topic of interest.

UK is one of the countries that is a pioneer in the incorporation and use of Linked Data concepts. They publish public data from multiple governmental agencies about multiple topics of interest and transport is one of them. Using the Semantic Web standards, UK publishes data about railway stations, airports, busses and traffic conditions which can be accessed on their website[3].

Google Transit Feed Specification (GTFS)[4] is a format for publishing data about public transport information connected with geographical locations. It is a standard proposed by Google in

Bojan Najdenov, Lund University, School of Economics and Management, Lund, Sweden (email: bojan.najdenov.947@student.lu.se)

Goran Petkovski, Faculty of Computer Science and Engineering, Skopje, Macedonia (email: petkovski.goran.1@students.finki.ukim.mk)

Milos Jovanovik, Faculty of Computer Science and Engineering, Skopje, Macedonia (email: milos.jovanovik@finki.ukim.mk)

Riste Stojanov, Faculty of Computer Science and Engineering, Skopje, Macedonia (email: riste.stojanov@finki.ukim.mk)

Dimitar Trajanov, Faculty of Computer Science and Engineering, Skopje, Macedonia (email: dimitar.trajanov@finki.ukim.mk)

[2] http://5stardata.info
[3] http://transport.data.gov.uk/
[4] https://developers.google.com/transit/gtfs/

order to help public transport agencies to publish their data and integrating with Google Maps while also providing means for developers to build applications over the data and promote interoperability. There are efforts for creating a GTFS ontology[5] and also for extending the GTFS ontology [8] to cover data that additionally has been introduced as part of the GTFS standard. The authors that extended the GTFS ontology used it to annotate transit data from the public transport agency in the city of Skopje, while also providing a SPARQL endpoint as a mean for querying the data [8].

3. CREATING LINKED DATA FROM THE SWEDISH TRANSPORT ADMINISTRATION

The Swedish Transport Administration (STA) – Trafikverket[6], is a Swedish governmental agency responsible for long-term planning and development of the Swedish national road network. They collect extensive amount of traffic data 24 hours a day, all year around and provide information to the general public either through their website or through web services which are available to software developer and researchers upon request. The information they publish includes road and traffic conditions in major cities, roads and highways within Sweden.

Currently, the data from STA accessible through their SOAP services is structured in Datex II format[7] which is derived from the XML format and perceived as standard by many governmental institutions within EU. Consequently, it can be concluded that the data from STA has 3-star data quality, according to the Star rating system of data. Even though the data published like this can be used by different applications, the possible use cases from user perspective are limited, should the data be available in 5-star quality using the Linked Open Data standards.

Therefore, we see an opportunity in transforming the public transport data from STA into 5-star quality and interlinking it with entities of the LOD Cloud. The automated workflow of the system is consisted of two main phases, as explained in the next two sub-sections, which are scheduled to be executed on hourly basis, in order to provide most-recently transport data. Explained briefly, here are the main steps which take place in the automated process:

1. Automated data gathering
    a. A script is scheduled to run and access the STA's services to obtain the XML data from the datasets of interest.
    b. The XML data is parsed and transformed into RDF/XML format using ontologies for semantic annotation.
    c. The RDF graph is loaded into Apache Jena Fuseki Server instance and is updated should new information be added.
2. Transformation into 5-Star Linked Data
    a. We run SPARQL-based procedures in order to locate instances form our local RDF Graph, locate corresponding entities in the LOD Cloud and create necessary links.

A. Automated Data Gathering

The process of automated data gathering is done using a windows batch script, which facilitates the requesting, gathering and storing the data locally, as files in XML format. Once the XML files are being stored locally, the next step in the process is transformation into RDF/XML format (4-star data according to W3C standard).

The transformation is done by parsing each XML file and adding RDF syntax to the content of each element, then every RDF/XML file is loaded into an instance of Apache Jena Fuseki Server[8]. Every time the automated process is being run, it updates the data contained in our RDF graph. The RDF graph is available via a persistent URI and it can be queried through a SPARQL endpoint.

B. Transport Administration Ontology

The main requirement for transforming the STA data in RDF format and later as Linked Data is an ontology to be present which would be used for semantic annotation of the data. Since no common ontology exist that could be used for the annotation of the whole STA data, we created the Transport Administration Ontology (TAO) and reused properties from other ontologies in the process of semantic annotation, following the Semantic Web standards.

TAO consists of classes and properties which describe every entity with all its attributes from the transport datasets we worked with: Road Condition, Road Work, Rest Area, Ferry Service and Accident Service. In this section we will describe our TAO ontology and will go through the properties which we reused from other ontologies, to be able to fulfill our goals.

The object properties that our TAO ontology has are shown in Table 1. The ontology has one object property and its correspondent inverse property, that make connections between a Situation_Record instances and the Location where they occurred.

Table 1. Object Properties of the TAO Ontology

Property	Description
has_Location	Information where the Situation instance was recorded. Inverse of location_Of_Siutation.
location_of_Situation	Used for determining situations that occurred on specified location.

Table 2 on the other hand illustrates the datatype properties that the TAO ontology introduces.

Table 2. Datatype Properties of the TAO Ontology

Property	Description
situationRecordTime	TimeStamp when the Situation_Record instance was created.
informationStatus	Current status of the Situation_Record instance .
speedLimit	Traffic speed limitation where the Accident occured.
lengthAffected	Total lenght of the affected carrigeway, measured in metres.
lanesRestricted	Number of lanes which are restricted due to the Accident event.

The TAO Ontology has been published with a persistent URI[9], and is dereferenceable via HTTP content negotiation, as the best practices suggest.

[5] http://lov.okfn.org/dataset/lov/vocabs/gtfs/
[6] http://www.trafikverket.se/
[7] http://www.datex2.eu/
[8] http://jena.apache.org/documentation/fuseki2/
[9] http://sta.linkeddata.finki.ukim.mk/ontology/tao#

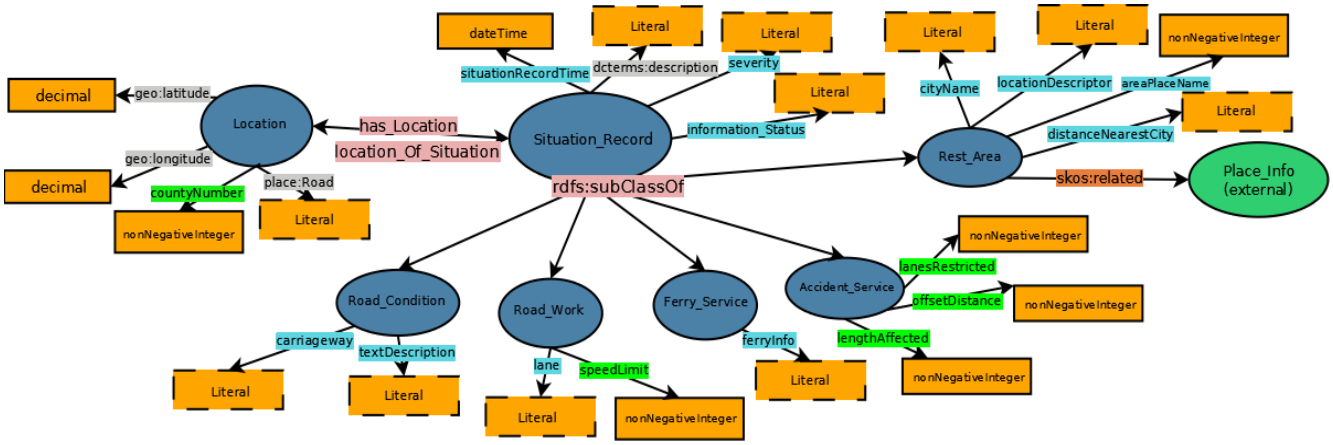

Figure 1. Diagram of the Transport Administration Ontology.

C. Transformation into 5-Star Linked Data

After the generation of the RDF graph, the next step in the process is the transformation of the data into 5-star Linked Data. The interlinking of data refers to the establishment of links between the data instances of our local dataset and other datasets available in LOD Cloud. To be able to make the connection with DBpedia, we use the skos:related property from the SKOS namespace [10] for the purpose of connecting instances of the cities found in our dataset with the related city instances described in DBpedia.

4. USE CASES

One of the main advantages of using Linked Data is the ability of accessing distributed data available on different locations on the Web, while starting from single source. This ability could provide large variety of possible use case scenarios involving transport data.

The Semantic Web technologies allow information retrieval from distributed datasets, through SPARQL federation.

In this section we will analyze two different types of scenarios, where we first query the local dataset only and afterwards we demonstrate how the information from DBpedia can be queried starting from our local dataset, providing useful information to potential users.

A. Using Road Accidents Data

In our work we address several different datasets published by STA as described previously, which we can use to demonstrate the information that could be obtained using the isolated dataset only.

Table 3. Results from the query on the local dataset.

Time Stamp	Point	Road Number	Length Affected
2015-06-10T19:35:28	12.0203247, 57.48617	Road 158	1516
2015-06-10T18:59:24	13.6646309, 55.75424	Road 1106	2060

One such scenario would be to find all Road Accidents occurred on some road, along with information about the instances. For this we can use the following SPARQL query:

```
SELECT ?Time (fn:concat(?Longitude, ",
"+?Latitude) AS ?Point) ?RoadNumber
?LengthAffected
WHERE {
   ?Accident tao:situationRecordTime ?Time;
     tao:lengthAffected ?LengthAffected;
     tao:has_Location ?Location.
   ?Location  place:Road ?RoadNumber;
              geo:longitude ?Longitude;
              geo:latitude ?Latitude. }
ORDER BY DESC(?Time)
```

As an answer to our query, we get the Time Stamp of occurrence of the road accident, geographical point that precisely shows where the accident happened, which road it happened on and the length of the affected carriageway measured in meters. The results from the query executed over our Road Accident dataset are shown in Table 3.

B. Using Additional Data from DBpedia

In this section we present a use case that is made possible only by the link we made with the LOD cloud through the skos:related property. In the query shown below, we are connecting to DBpedia as part of the LOD Cloud, in order to retrieve more information about the city which is closest to a rest area in Sweden.

```
SELECT DISTINCT ?cityName  ?RestAreaName
?abstract ?thumbnail
WHERE
{
    ?city tao:cityName ?cityName;
              skos:related ?dbpediaResource.
SERVICE <http://dbpedia.org/sparql>
  {
       ?NearestCity dbpedia-owl:nearestCity
?dbpediaResource;
              dbpedia-owl:abstract ?abstract;
              dbpedia-owl:thumbnail ?thumbnail;
              dbpprop:name ?RestAreaName.
         FILTER langMatches(lang(?abstract),
"en")
  }
}
```

This query first executes over the local RDF graph, looking for cities via the property tao:cityName. The detected instance is of a

[10] http://www.w3.org/TR/skos-reference/

contains information about multiple facts which are related with the city itself.

All of these example use case scenarios can be implemented in any applications, since the SPARQL endpoint[11] we created can be used as a REST service. The GET calls should have the following format:

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

Here, SPARQLQUERY represents the SPARQL query which is to be executed, and FORMAT represents the format of the response, such as HTML, XML, JSON, CSV, RDF/XML, N3, Turtle, JSON-LD, etc.

This could provide an opportunity for developers to access additional information, previously unavailable over the local dataset from STA website.

5. WEB APPLICATION

The main use case scenarios can be seen in the web application[12] that we built for demonstration purposes. It uses data from the Apache Fuseki instance and provides information to the user about most recent transport data from the STA site.

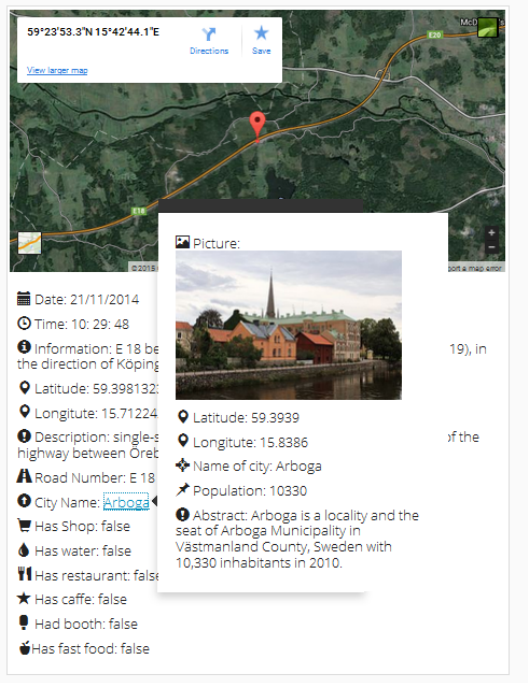

Figure 2. Details about city retrieved from DBpedia.

The web application uses our SPARQL endpoint to query the data from the local dataset and data from LOD Cloud. One simple use case would be showing information about a city (abstract, population, geolocation data, etc.) retrieved from an external data source - DBpedia. This can be done by first choosing a corresponding Rest Area instance, and then selecting the city Name. Then, the information is shown below, so the user can see all the information about the city in one window (Figure 2).

6. CONCLUSION

The Semantic Web and the Linked Data concept are considered to be the next generation web of data which is structured and interlinked by its meaning. By publishing the data with 5-star quality and contributing to the LOD Cloud, we hope that we increased the possibilities and motivated other organizations to publish the data in this way.

In this paper we described a system we developed that automatically gathers, transforms and publishes of 5-star Linked Open Data from the transport domain, while also making it accessible through a SPARQL endpoint. The initial 3-star quality from the STA is transformed into Linked Open Data using the TAO ontology. In addition to that, we demonstrated advanced use case scenarios where the information from interlinked datasets is being used thus enriching the user experience. Finally, we developed a web application as proof of concept which demonstrates the new scenarios.

The main idea behind this paper was to present the development of an automated system which brings forward new possibilities, new scenarios that can come out of the publishing of data as Linked Open Data. From a point of view of an isolated data sets, these advanced scenarios were previously unavailable. In the future, we hope to extend the ontology and develop it further so that it would follow the standards and models proposed by the INSPIRE Directive. That way, new opportunities for semantic annotation of published data which follows those standards would arise. Furthermore, we would like to extend the possibilities by interlinking the STA data with more datasets and also to motivate organizations to publish data in this way, thus increasing the value of the services they deliver.

REFERENCES

[1] T. H. Davenport, "Competing on analytics", *Harvard Business Review*, 2006, 84(1), p. 98.

[2] H. Chen, R. H. Chiang and V. C. Storey, "Business Intelligence and Analytics: From Big Data to Big Impact", *MIS Quarterly*, 36(4), 2012, pp. 1165-1188.

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

[4] T. Berners-Lee, J. Hendler and O. Lassila, "The semantic web", *Scientific American*, 2001, 284(5), pp. 28-37.

[5] N. Shadbolt, W. Hall, T.Berners-Lee, "The semantic web revisited", *Intelligent Systems*, IEEE, 21(3), 2006, pp. 96-101.

[6] T. Berners-Lee, "Semantic web road map", 1998.

[7] C. Bizer, T. Heath, K. Idehen and T. Berners-Lee, "Linked data on the web", *17th International conference on World Wide Web*, ACM, 2008, pp. 1265-1266.

[8] E. Misheva, B. Najdenov, M. Jovanovik and D. Trajanov, "Open Public Transport Data in Macedonia", 11th Conference for Informatics and Information Technology (CIIT), 2014.

[11] http://sta.linkeddata.finki.ukim.mk/sparql
[12] http://sta.linkeddata.finki.ukim.mk/