FACULTY OF ELECTRICAL ENGINEERING

Combining cycling with public transport in Prague (Checkpoint 1)

OSW - Ontologies and Semantic Web

Michal Cvach
November 18, 2018

1 DATA PIPELINE

Here, I will describe how I have transformed the data sources chosen in checkpoint 0 into RDF. Most of the changes were made using the **OntoRefine** tool in **GraphDB**, so for each dataset, I have added the OntoRefine project into the deliverable. In addition to that, **SPARQL** insert scripts for all the data are also part of the deliverable.

1.1 CYCLING ROUTES IN PRAGUE

The first data set I have chosen in the previous checkpoint was also the one most complicated to work with. The data were in **GeoJSON** format, where from the geographical point of view, each cycling route was represented by a sequence of points which, when put together, represent the shape of the route.

I have decided, that for simplicity, I will only work with the geographical data of the beginning and end of a cycling route. To transform the GeoJSON file accordingly, I have created a simple C++ program (simplify_routes.cpp in the deliverable), that does a few things with the input file. First of all, it erases all the unnecessary points in the route shape, leaving only the first and the last. Additionally, it adds some names to cycling routes, which did not have any at the beginning, because the names are later used as identifiers. And it also sets the one way route flag to 'N' (false) for routes that are not one way.

After the data were processed using the C++ program, I have loaded them into OntoRefine and then did some changes there. Most importantly, I have transposed every for cells in the

geometry column into separate columns, creating columns for BeginLatitude, BeginLongitude, EndLatitude and EndLongitude. After that, each row corresponds to one record, so the following edit operations were mostly column renames and so on.

1.2 Public transport stations in Prague by type

The second data set was a little bit easier to work with. It just needed a few adjustments in OntoRefine. Mainly, I had to again transpose some cells, this time every two cells in the coordinates column, in order to have both latitude and longitude in the same row. After that, I did some renaming and columns reordering here and there.

1.3 Parking lots in Prague

And finally the parking lots data set was almost ready from the get go, so I just did some renaming, removed some unnecessary columns, and then the data set was ready to be transformed into RDF.

2 Data set schema

In the following diagrams, schema for each of the data sets is depicted. The yellow ellipse denotes the **class**, and the cyan ellipses denote **properties**. There are some properties that could be handled a little bit differently, for example the **Longitude** and **Latitude** properties are present in each schema, so it would maybe make sense to add an additional property **Coordinates** for example, which would have subproperties Longitude and Latitude.

2.1 Cycling routes in Prague

Figure 2.1 depicts the data scheme for the cycling routes data set.

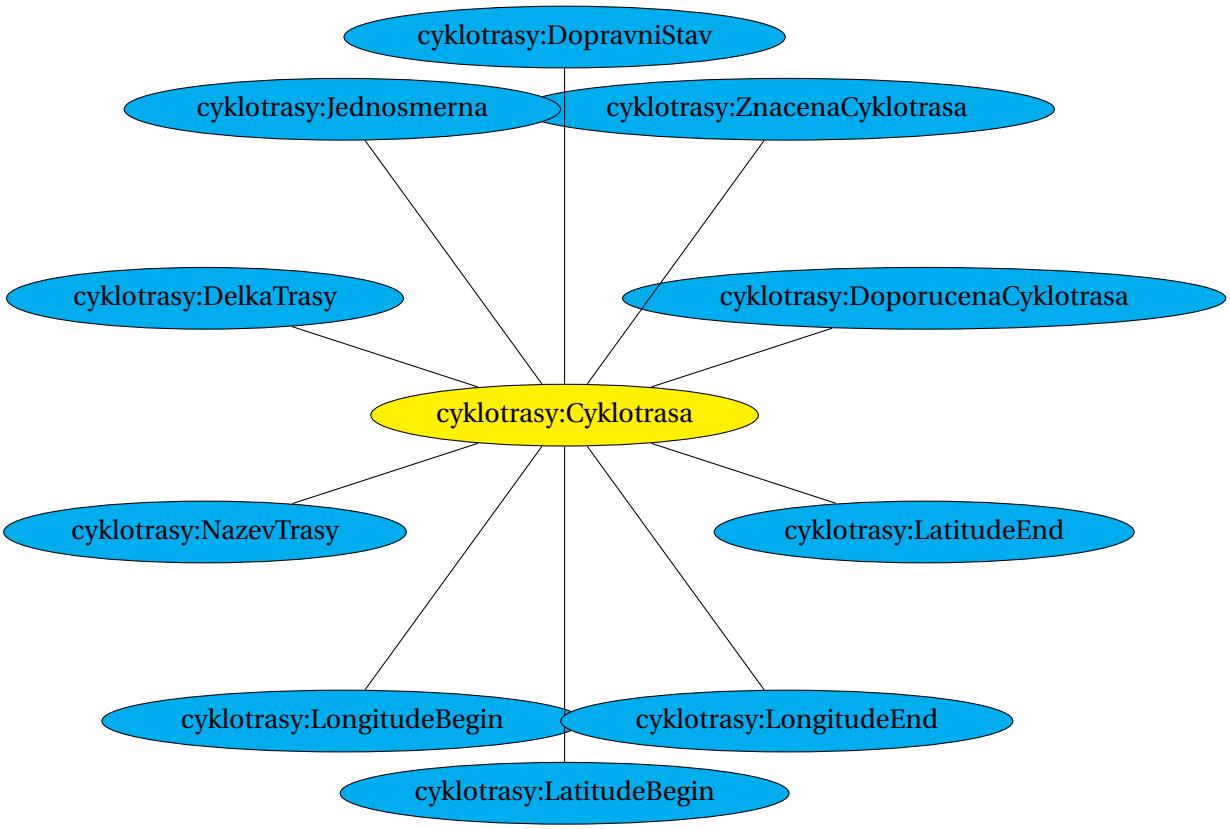

Figure 2.1: Properties for the cycling routes data

2.2 Public transport stations in Prague by type

Figure 2.2 depicts the data scheme for the public transport stations data set.

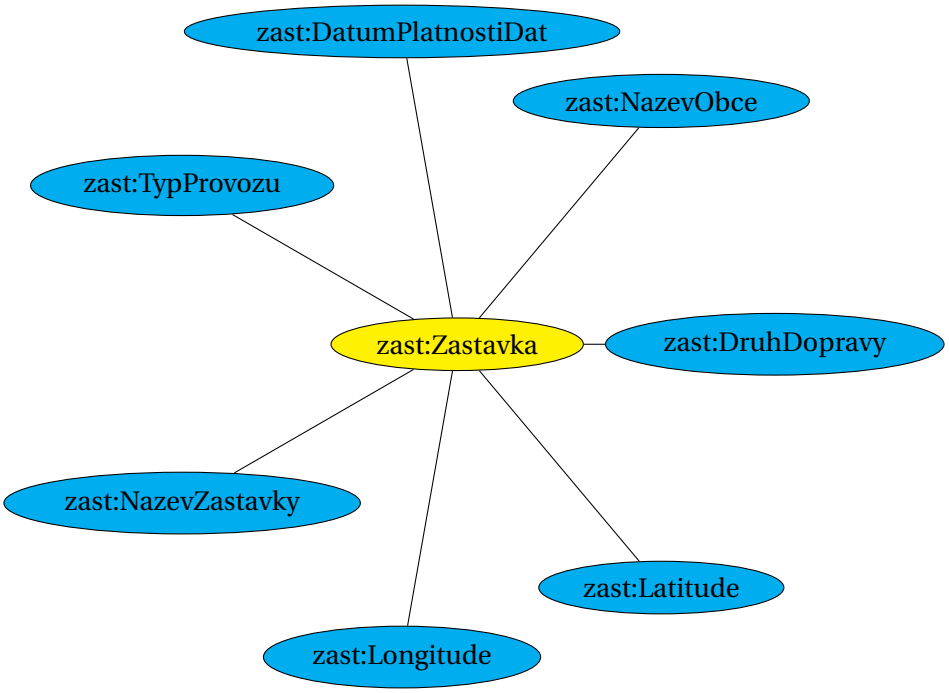

Figure 2.2: Properties for the public transport stations data

2.3 Parking lots in Prague

Figure 2.3 depicts the data scheme for the parking lots data set.

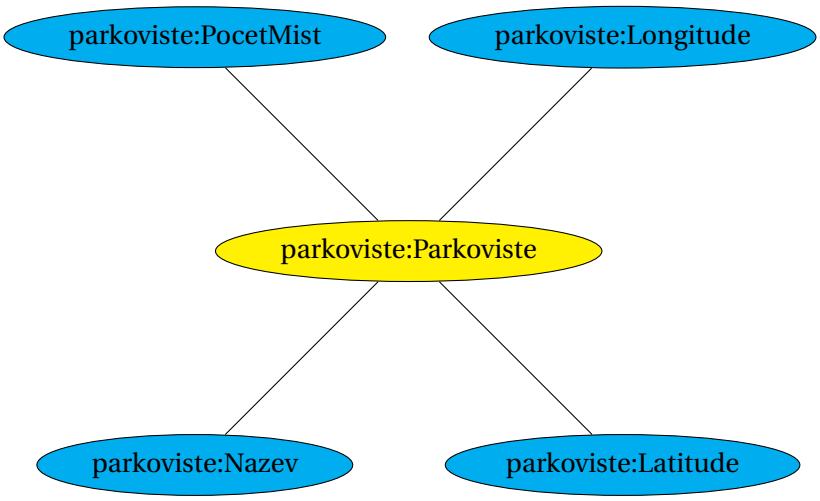

Figure 2.3: Properties for the parking lots data

3 Attachments

Other than this PDF file, the deliverable also contains the following:

- **simplify_routes.cpp** and a **Makefile.** Source code of my simple program I have used to transform the cycling routes data set before loading it into OntoRefine with a Makefile, so that it can be compiled on a computer with gcc installed just by using the command `make compile` in the command line.
- **insert_dataset1.sparql.** A SPARQL query used for generating triples from the cycling routes data set.
- **insert_dataset2.sparql.** A SPARQL query used for generating triples from the public transport stations data set.
- **insert_dataset3.sparql.** A SPARQL query used for generating triples from the parking lots data set.

  Dataset1-Cyklotrasy.openrefine.tar.gz. OntoRefine project for the cycling routes data set.

  Dataset2-Zastavky.openrefine.tar.gz. OntoRefine project for the public transport stations data set.

  Dataset3-Parkoviste.openrefine.tar.gz. OntoRefine project for the parking lots data set.