B4M33OSW: Checkpoint 1

November 29, 2017

1 Data pipeline

1.1 Prague data set

The data for Prague public transport [1] is well structured and distributed in an open format [2]. The format specification defines various files, each of which contains a different type of information. I used only two of the files:

- `stops.txt` – contains names of the stops
- `stop_times.txt` – contains arrival and departure times

Because of the format, no explicit transformation was needed. Therefore the only tool I used was OntoRefine. I uploaded each of the files into OntoRefine and created a new project for both of them. Then, I removed unnecessary columns and created RDF triples using SPARQL. Both SPARQL queries (for stops and stop times) are attached to the deliverable.

1.2 Pilsen data set

Unlike Prague, Pilsen does not provide access to structured public transport data. I had found the data available in HTML tables [3] and used a simple R script to scrape the content of the tables and to save it into a CSV file.

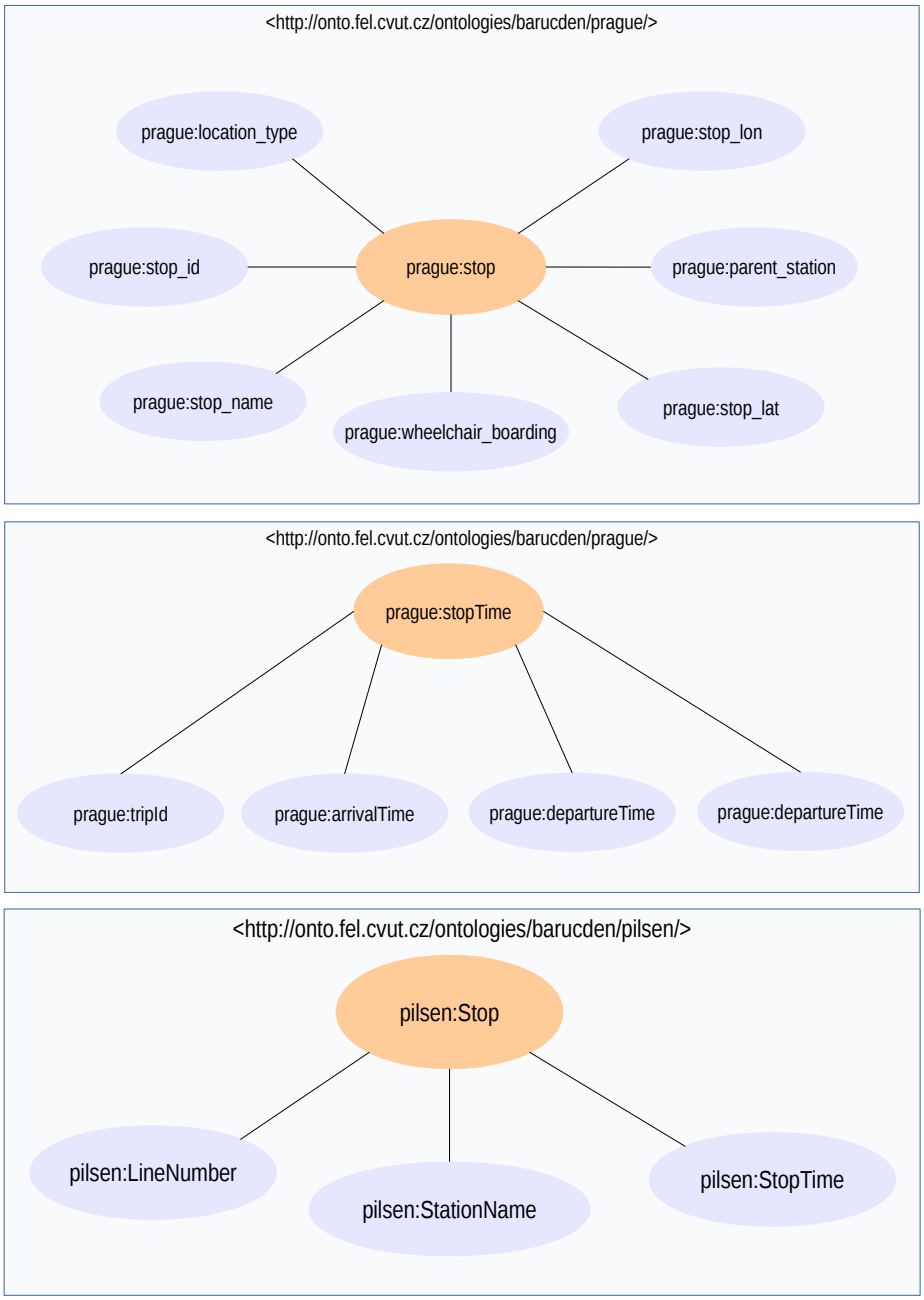
Then again I used OntoRefine to import the CSV file and to turn it into RDF triples using SPARQL. Because of the way the station names were displayed in the tables, and the R script being far from perfect, there were a few hundreds of incorrect station names in the CSV file (e.g., "`Sokolova /Z/`" instead of "`Sokolova`"). To fix those names I used the text-transform feature in OntoRefine. The R script, the SPARQL query, and an export of the transformation are attached to the deliverable.

2 Data set schema

The orange bubbles are classes, and the blue ones are properties. The data contains all the information needed to accomplish the goals specified within Checkpoint 0.

Crucial properties are:

- `pilsen:StopTime` and `pilsen:StationName` for Pilsen,
- `prague:stop_name`, `prague:arrivalTime` and `prague:departureTime` for Prague.



3 Attachments

- `pilsen_scraper.r` – the web scraper in R
- `pilsen_sparql_stops.rq` – SPAQRL query to insert station names and times for Pilsen
- `pilsen_text_transformation.txt` – the export of the text transformation from OntoRefine
- `prague_sparql_stops.rq` – the SPAQRL query to insert station names for Prague
- `prague_sparql_stop_times.rq` – the SPAQRL query to insert arrival and departure times for Prague
- `pilsen_export.ttl` – export of the graph for Pilsen
- `prague_export.ttl` – export of the graph for Prague

References

[1] http://opendata.praha.eu/

[2] https://developers.google.com/transit/gtfs/reference/

[3] http://jizdnirady.pmdp.cz/LinesList.aspx