Project idea: Integration of Hamburg District-Level Statistics with Public Transport Data

Ozan Heydt

Objective

This project aims to integrate socio-economic data for Hamburg districts from Statistikamt Nord with public transport data from GTFS sources using RDF. By transforming and linking these datasets, the project will produce a unified semantic representation that enables advanced analysis of the relationships between socio-economic indicators and transport accessibility. The output will be evaluated for quality and consistency using RDFUnit.

Scope

The project will focus on merging two distinct but complementary datasets: socio-economic district-level data and public transport data. The Hamburger Stadtteil-Profile provides detailed statistical insights into Hamburg's districts, such as population density, unemployment rates, and average income. On the other hand, the GTFS data contains rich information about transport stops, routes, and schedules for the Hamburg public transport network. These datasets will be transformed into RDF format, linked using appropriate vocabularies, and enriched with calculated metrics such as transport stops per capita and average distance to nearest transport stop.

The integration process will adhere to semantic web principles to ensure compatibility with other linked datasets and facilitate querying via SPARQL.

Methodology

The project will involve three primary steps: extraction, transformation, and linking. The statistical and transport data will be preprocessed to ensure compatibility and uniformity in structure and geospatial alignment. The RDF transformation will involve mapping the data into appropriate vocabularies, such as FOAF (Friend of a Friend) for district details and GTFS ontology for transport information. Geospatial relationships will be used to associate transport stops with their respective districts.

After integration, the RDF dataset will be validated using RDFUnit to identify and resolve inconsistencies. The validation process will ensure compliance with the chosen ontologies and the accuracy of the relationships within the data.

Data Sources

- **Hamburger Stadtteil-Profile (Statistikamt Nord)**: Provides district-level socio-economic data, including population, area, unemployment rates, and income levels. `https://region.statistik-nord.de/compare/selection/2#`
- **GTFS Data for Hamburg (GTFS.de)**: Offers information about transport stops, routes, and schedules for Hamburg's public transport system. `https://gtfs.de/de/feeds/de_nv/`
- **Geospatial Data**: Includes district boundaries and centroids, sourced from OpenStreetMap or Geoportal Hamburg, to facilitate spatial linking.

Expected Outcome

The primary deliverable will be a unified RDF dataset representing both socio-economic and transport data for Hamburg districts. The dataset will feature enriched metrics, such as transport accessibility indicators (e.g., transport stops per capita, average distance to nearest stop) linked to socio-economic attributes.

A secondary deliverable will be a comprehensive report detailing the data transformation and linking process, the ontologies used, and the results of RDFUnit evaluations. The report will include examples of SPARQL queries showcasing the utility of the dataset for exploring relationships between socio-economic factors and transport accessibility.