ECS7028P - Data Semantics

Final Mini Project – UK Company Ontology

Sami Saade

10 May 2024

I. Introduction:

The UK is the home for millions of global and local companies that play a huge role in the economic market of the world. This project attempts to build an OWL Ontology using Protégé and populating it using the GOV.UK RDF data hosted on their 'business.data.gov.uk' domain. The data is readily available for public use without any license as per the Companies Act 2006. The aim is to have a system where we can query the companies in the UK, the industries they work in, and the locations they are in. To extend the potential of this ontology, we link the data to wikidata.org where we can find information about population and more regarding the locations of the companies.

II. Creating the local RDF store:

GOV.UK does not offer a remote SPARQL Endpoint which means the data has to be stored in a local RDF store to be suitable for query using SPARQL. Though, the RDF data is not found in a complete data dump. To overcome that, the URIs are parsed individually, and the graphs are 'merged' together using RDFLIB through an addition operation provided by the library. To make the parsing more attainable though, we need to be able to select valid URIs programmatically. Thus, the full company database was downloaded from 'https://download.companieshouse.gov.uk/en_output.html' in csv from which allows us to extract the URIs using pandas and store it as a pickle file. The URIs are then iterated from the list, parsed, merging, and finally saving and serializing into a local turtle file.

III. Setting up a Federated SPARQL query:

As we are querying over a local and remote endpoint, we make use of the SPARQL SERVICE keyword. This allows us to query against two datasets and link them together. Looking at the local Company Data, we find that the town and city names are written in capital letters. We can use the UCASE SPARQL function to match the city and town labels from WikiData to the ones from our local store. We can then filter the data to find the URI of the city and make the company registered to it.

IV. The Ontology:

Class	Asserted	Inferred
Company	SIC_Code	Economic Activity
Economic Activity		Instances
Place	Population	Part of
SIC		Instances

Economic Activity holds the categories set by gov.uk. SWRL is used to infer the categories by comparing the SIC codes to the boundaries of the category SIC codes, as the SIC codes are grouped by ascending order. The company's economic activity is then inferred as indicated by its SIC.

Company has a name, registration place, and sic code. Places have a name and a population. This allows the user to ask questions like the spread of economic activity types across locations, or the population of the cities and towns a company is registered in, as of the initial implementation in this project. All properties could be extended by including more data from the gov dataset such as incorporation date, mortgages, and more and elevation, wealth, and poverty rates from wiki data for the locations.

Class	**Individuals**
Company	1433117, Rice_Co (asserted)
Town	Q24826 (Liverpool), Camberwell (Asserted)
City	London (Asserted)
SIC	Growing_of_rice (inferred)
Agriculture_Forestry_and_fishing	Growing_of_rice, Rice_Co (inferred)

Conclusion:

The ontology serves its purpose to answer geo-economic questions about the companies in the UK. Having to load a local RDF store and matching them took some time to work, but eventually, with the right formatting, the merging of the two databases was possible from the registration post town info and label town name bridge. SWRL was used to infer categories of economic activity. The ontology is successfully populated from the government dataset which is loaded locally and the wikidata dataset which is reached via a remote Federated query. Future work could be expanding the data properties to include all that is available in the gov dataset, and extend it to aspects such as visa sponsor providing companies, employee data and demographics, and more.