UNIVERSITETET I BERGEN

Det samfunnsvitenskapelige fakultet

Semantic Homicides

Semester Assignment – INFO216

Candidates: 105, 167, 112

Word count: 900

1. Introduction

1.1 Introduction to Semantic Homicides

In the United States of America more than 15 thousand homicides are committed each year (Murphy, 2017, p. 34). Records of all these cases are available online in open datasets through the Murder Accountability Project, a non-profit group that aims to uncover the low clearance rates of domestic murder cases (Murder Accountability Project, 2018). We were interested in how we could widen the usefulness of the homicide dataset through combining it with other relevant datasets in a semantic application. The group therefore set out to create a semantic homicide application that is able to connect murder data with demographic and economic data about the population in all different states in the U.S. with possibilities of visualizing query results in various ways.

1.2 Problem Specification

There already exists a vast number of statistics on violent crime for different contexts, so why make an application that extracts data from homicide records for viewing and visualizing? A semantic application that includes linked data from relevant datasets can be a useful tool because it lets the user both gather data and ready it for analyzing with visualizations, all in the same simple application. The application is also easily expandable, and it is thereby possible to enhance the scope of usefulness over time. In our project we have chosen a few datasets that we considered relevant in homicide statistics in order to show how an application like this could work and to prove its usefulness.

1.3 Data

The data used in our semantic homicides application was gathered from the following datasets:
- Supplementary Homicide Report
- 2012-2016 American Community Survey 5-year Estimates (ACS).
  - Selected Economic Characteristics
  - Race
- Annual Survey of School System Finances: Per Pupil Amounts for Current Spending of Public Elementary-Secondary School Systems by State: Fiscal Year 2014

While the dataset of Supplementary Homicide Report reaches from 2016 and all the way back to 1976, the other datasets included are only of recent matter. Earlier publications of the included datasets could with advantage be implemented, as the Annual Survey of School System Finances reaches back to 1992 and the American Community Survey (ACS) started publications in 2005. Demographic data from further back in time than 2005 is also available, however in a different format than the ACS. Historical data could certainly enhance the usefulness of our application, extending the possibilities for analyzing correlations between homicides and social phenomena.

2. Development Process

2.1 Gathering Data

While looking for data to combine with the homicide dataset, we visited a variety of opportunities. Initially we wanted to connect the data to existing semantic content on the web, like dbpedia. We succeeded with connecting the states to dbpedia, though the data dbpedia provided was neither extensive nor relevant enough for the context of a semantic homicide application. We therefore scrapped the idea of using dbpedia and tried to find more relevant datasets that could be used. At factfinder.census.gov we discovered datasets with demographic information for each state that we could combine with the homicide dataset.

2.2 Lifting Data

All the data gathered was originally available as open datasets in .csv format. In order to make use of them in a semantic application we needed to lift the data. Initially we made use of the Jena library to create RDF files from the CSVs in Java. However, the Jena CSV2RDF function caused a few issues regarding the resulting RDF file, that we to achieve solving in Java. This was part due to the auto-generated URIs and prefixes that the CSV2RDF function created based on the file's location when writing the RDF file. Because of these problems we decided to use OpenRefine (formerly Google Refine) to lift our CSV files. When lifting the data, we decided to discard data from each dataset that would not serve any purpose in the application. After deciding what data to include we made our own vocabularies "hom" for homicides and "dem" for demography, and created properties for the object types "hom:Homicide" and "dem:State". After successfully creating RDF files from the CSVs, we made a TDB database with Jena and added the RDF models into this.
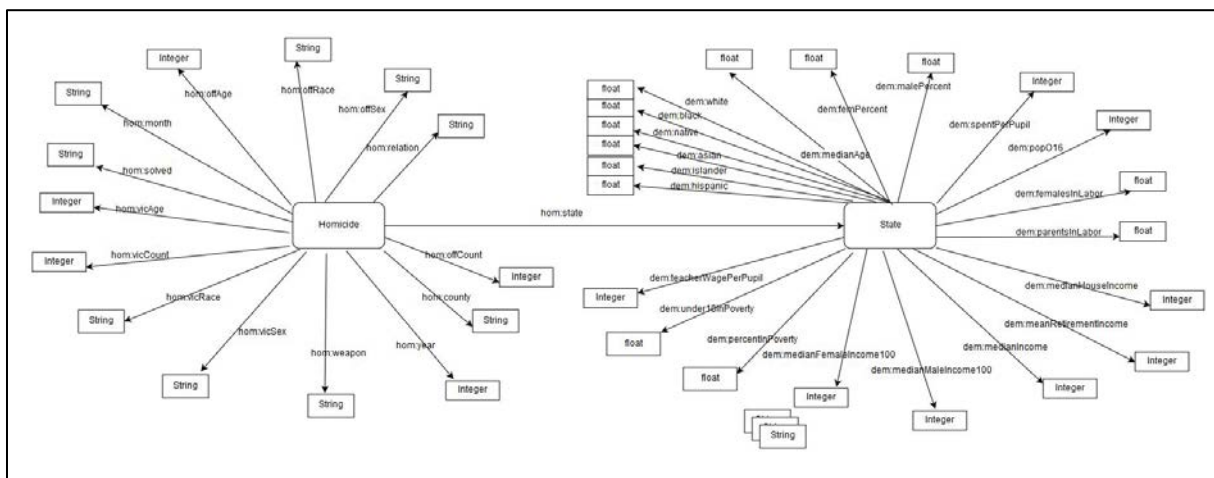

Figure 1 - Class Diagram

2.3 SPARQL Queries and Sgvizler

With the database in place, we wanted to make it possible to query it through an interface and visualize the data, demonstrating some of the opportunities for our application. An extremely simple html page was made and using Sgvizler we created some queries we found interesting presented in fitting charts and graphs. In addition, the user can make his own queries in the application and choose the preferred visualization.
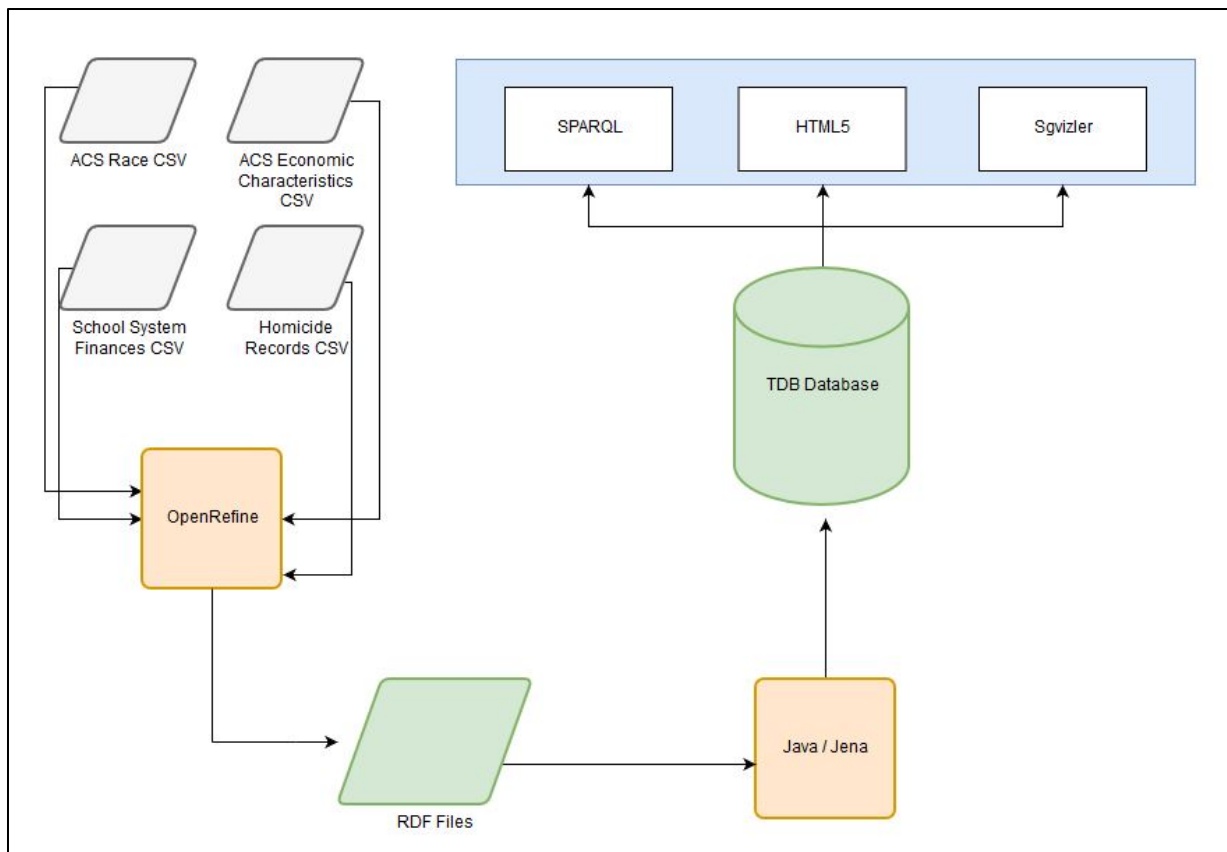
Figure 2 - Flow Chart

3. Evaluation and Further Work

3.1 Evaluation

When initiating the project, we set out to demonstrate how semantic technologies can make homicide records more useful. Our solution provides possibilities for querying in a database that includes semantic data about homicides and demography, across multiple datasets.

3.2 Further Work

With the RDF models in place, we would like to write code for a more automated way of lifting that are datasets similar to the existing ones. Expanding the database with more datasets e.g., demography datasets from further back in time, county specific and not only state specific data, in addition to law data, would help improve the usefulness of our application. Furthermore, a complete interface to replace our current basic page has to be implemented in order to complete the application.

Sources:

Murphy SL, Xu JQ, Kochanek KD, Curtin SC, Arias E (2017). Deaths: Final data for 2015. National Vital Statistics Reports, vol 66 no 6. Hyattsville, MD: National Center for Health Statistics.

Murder Accountability Project (2018). *Why We Exist* [Internet]. Available from: http://www.murderdata.org/p/about.html [22.05.2018]