Municipalities and Populations with Knowledge Graphs

Info 216

Group ID: 1776

Candidate Numbers: 167, 189

Project description

While deciding on what to create there were two general ideas. The first was to combine datasets relating to Norwegian mountains and mountain ranges with tourist locations and different mineral types. In order to create an application for hikers and campers looking for a specific type of biome.

The second idea, the one decided on moving forward with, was chosen after initial problems with implementing a reasonable solution when trying to combine the different mountain datasets, as these had locations based on different values, coordinates vs location names. Instead it was decided to go with a combination of municipalities and counties, and population numbers of said municipalities, keeping somewhat with the motif of Norway. We decided to create a program for those in need of knowing what is where.
The program is meant to be a simple and easy way for the client to find any municipality in any county, and also be able to get which counties have which municipality.
Users will also be able to get a constantly updated population count of every municipality.
While county names rarely change, population count is a figure that is constantly undergoing changes. As a result we figured that the use of semantic technologies would be beneficial to solve the issue of ever changing populations.
By using semantic technologies to retrieve data from the open cloud we are able to link our local data to the open cloud. As a result we are able to automatically and consistently keep the population numbers updated. We saw this as not only a way to solve the issue of updating the population count, but also a way to try out lifting none-semantic data.

In the program there are three different datasets.
As for tools, we decided on using python via pycharm, as it is the programming language we are the most familiar with, and would hence give us the most flexibility. Pandas was used in order to read the CSV files, so we could lift them. From rdflib we used Graph, Namespace and URIRef, to build our knowledge graph. We used prepareQuery to perform a search within our local knowledge graph, as well as SPARQLWrapper and JSON to read data from Wikidata.

When it comes to vocabularies we first considered using a norwegian one which can be found here: "http://vocab.lenka.no/geo-deling#" to describe counties and municipalities, as this was linked up to dbpedia and other big vocabulary websites. However we discovered that instead we could just use the already widely used versions on dbpedia, in addition to some namespaces from dbpedia we also used two namespaces from wikidata to better describe relations between counties and municipalities.
As for actual development, because of the corona situation, all members felt safest doing most work digitally. This was done primarily using the desktop application discord, allowing voice chat and easy screen sharing capabilities. We scheduled several digital meetings throughout the semester, where we both worked on the project.

One of the main issues faced when lifting the offline datasets, was errors created when the Norwegian letters “Æ,Ø,Å”, used in most island names were not recognized as actual symbols. At first we tried to solve this automatically, by replacing them with alternative spellings, but realised that in order to make the program as user friendly as possible for the norwegians that would use the program, it would have to take in the norwegian lettering. Therefore we manually changed the symbols ÆØÅ replaced with the respective letters.

A quite serious issue we faced was that when testing the program on a different computer than the one it was created on, was that the ask3() function which gathered population numbers from wikidata would give a “void choice” error message. While going through the code in order to find the problem, it seemingly fixed itself after about ten minutes, without any changes to the actual code.As far as we know lies the problem not in the code itself, but rather potentially in the query to Wikidata.
While it would be impossible to know beforehand, looking back we would most likely want to figure out a manageable project much earlier, in order to have more time to dedicate to the project.

Sources for local csv data:

fylker.csv is downloaded from SSB which is the norwegian statistics agency. The link to where it is downloaded from is: <https://www.ssb.no/klass/klassifikasjoner/104>.

kommuner.csv is downloaded from the SSB as well. The link to where it is downloaded from is: <https://www.ssb.no/klass/klassifikasjoner/131>.

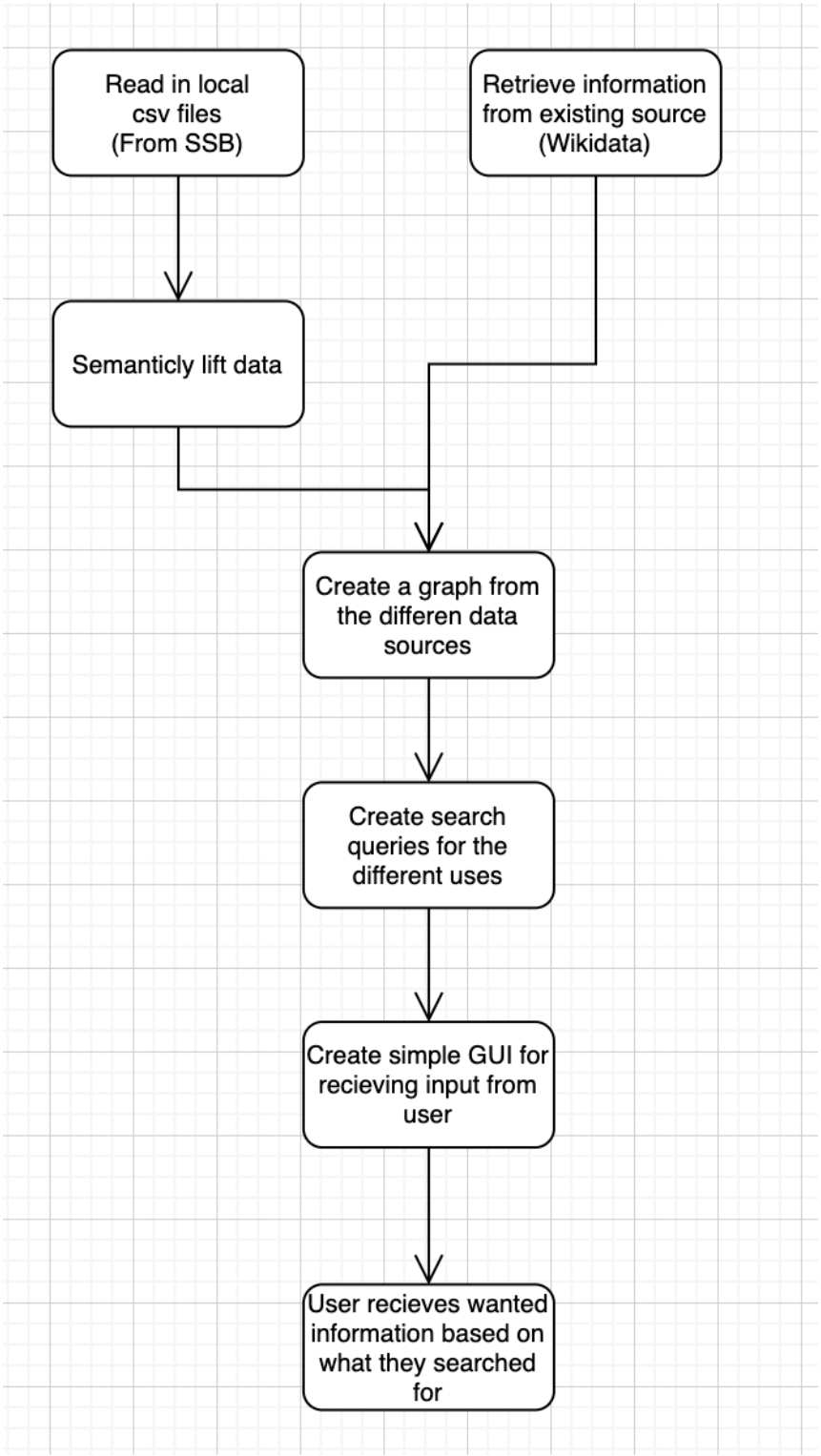

Figure 1. Simple flowchart showing program structure