Building & Mining Knowledge Graphs

(KEN4256)

Assignment 1:

Knowledge graph construction, integration and basic querying

Due date: 17 February 2020 (upload on student portal before **midnight**)

Description

For this assignment, you will convert two different datasets to a knowledge graph representation format - Resource Description Format (RDF), integrate these two datasets, import the resulting graph into an RDF triple store, and query it to answer questions about its content. You will perform the work in groups of 5 (please arrange yourself into such groups). **ONE person from each group should send an email with the names and student IDs of each member of their group to** michel.dumontier@maastrichtuniversity.nl and kody.moodley@maastrichtuniversity.nl

Recalling the lectures and labs on Monday and Wednesday, 3rd and 5th February, we learned that integrating data sources in a Knowledge Graph gives us access to more complete information about a domain that enables us to ask questions that we could not answer by querying the individual separate datasets.

WorldBank contains economic information about various countries around the world. Geonames contains geographical information about countries. Integrating these sources, and converting them to the RDF, will allow us to query them using the RDF graph querying language (SPARQL) and to answer questions that concern both economic and geographic information about countries.

You can find all necessary materials (dataset files, examples, documentation) at this GitHUb repository: https://github.com/MaastrichtU-IDS/UM_KEN4256_KnowledgeGraphs

Tasks

- Download the two datasets WorldBank and GeoNames which are in XML and CSV format [3,4].
- Use RML [1] to convert both datasets to RDF [5]. Assign types to the instances in each triple of the converted dataset (using an appropriate, publicly available shared vocabulary term). **Motivate clearly your choice for each term**. An example RML file is provided in the Github repository (mapping.ttl) with a few fields already set, edit the RML mapping file by filling in the missing information and execute it on the two data sources to convert them to RDF:
    - Make sure you convert all fields from the WorldBank XML file and GeoNames CSV file.
    - Add a class (shared vocabulary term) for 'Continent' in the Geonames dataset, and for 'Country' in the WorldBank (unless you want to represent WorldBank countries differently, motivate your choice in this case).
    - Use URIs from existing relevant vocabularies to describe the concepts and properties produced in your triples (**motivate your choice for each term**).
- Upload the converted datasets to GraphDB in a new graph with namespace: "http://kg-course/mapping".
- Link the two converted datasets.
    - The created GeoNames and WorldBank countries are not linked. What this means is that there are no relations between entities from Geonames and those from WorldBank. In order to answer some of the questions in this assignment, we need to link the two graphs.
        - **Hint:** one strategy could be to notice that both datasets have country entities
        - **Note:** some country names will not match exactly across the datasets and may need the use of an approximate instance matching tool such as LIMES (instructions to run LIMES can be found in the Git repository). A sample LIMES configuration file (limes_config.xml) is also provided in the repository.
    - Upload the triples resulting from the interlinking into GraphDB, in the graph namespace "http://kg-course/interlinking".
    - We encourage you to experiment with the settings in LIMES (specifically with the type of metrics and thresholds for matches) or to propose alternative methods to perform the interlinking (not involving LIMES). Whatever your decisions, **clearly state which choices you made and explain fully why you made them**.
- Answer the following questions using SPARQL queries (you may use the results of multiple separate queries to answer each question if you prefer, although it is possible to answer each using a single query):

1. List all countries with population less than 50,000 and order them from the smallest to the largest in terms of landmass area (square kilometres)
2. List the countries with the top 10 highest GDP values in 2017
3. List the countries with the top 10 highest **increases** in GDP between 1960 and 2017
4. For each continent, count the number of countries in that continent that are in the top 20 for highest **increases** in GDP between 1960 and 2017. Your answer should contain 2 values: 1) the continent and 2) the number of countries from the top 20 that are located in those continents (For example, Asia - 15, Europe - 5)
5. Construct the triples representing the GDP per capita for each country in 2017
6. Directly insert the triples representing the GDP per capita for each country in 2017, into your triplestore on GraphDB in the graph namespace "http://kg-course/query"
7. Bonus: feel free to propose original elaborated SPARQL queries getting new interesting insights from the dataset.

Deliverables

One student per group will deliver a written technical report (max 5 pages), which contains the following information:

- A description in your own words of the following steps of the assignment:
    - The conversion of the datasets to RDF (including a description in your words of the instructions expressed in the RML mapping file) [**minimum 80 words**]
    - The linking of the two datasets using LIMES or other methods [**minimum 80 words**]
- The number of RDF triples generated by applying the mapping file(s).
- The type and number of entities linked across the datasets.
- A list of the SPARQL queries used (clearly marking the question it answers) along with a short explanation of the rationale behind the design choices behind the query (maximum 3 lines). **Hint:** there are multiple ways to formulate the same query, explain why you chose your formulation. Also, a large query often can be seen as composed of 'sub-queries'. If you use such a query, explain what these sub-queries do and how their results help to answer the ultimate question.

And please submit the following files separately (also on Student Portal):

- All RML mapping files used (.ttl file)
- A **single** text file (.txt extension) containing all your SPARQL queries **clearly marked with the question number that it answers**
- An export of your GraphDB triplestore (all graphs produced) as N-Quads format with .nq file extension (Graphs Overview > Export Repository).

Resources

[1] - RML: http://rml.io

[2] - GraphDB (RDF store management software):
https://www.ontotext.com/products/graphdb/#button

[3] - WorldBank (financial information about countries)

Yearly GDP per country dataset (XML):
https://raw.githubusercontent.com/MaastrichtU-IDS/UM_KEN4256_KnowledgeGraphs/master/dataset-worldbank-gdp-full.xml

[4] - GeoNames (geographical information about countries)

Countries dataset (CSV):

https://raw.githubusercontent.com/MaastrichtU-IDS/UM_KEN4256_KnowledgeGraphs/master/dataset-geonames-countryInfo.csv

[5] - Resource Description Framework (RDF)

https://www.w3.org/TR/rdf11-concepts/

[6] - Course materials on Student Portal (lab and lecture slides)

Questions and comments:

Prof. Michel Dumontier: michel.dumontier@maastrichtuniversity.nl

Dr. Kody Moodley: kody.moodley@maastrichtuniversity.nl