HW1 Knowledge Technologies (M164)
Due: 23/11/23

Exercise 1 (250 marks)

The most central data source in the linked data cloud is DBpedia (http://wiki.dbpedia.org/), a big knowledge base which is essentially a "translation" of parts of Wikipedia into RDF. In this exercise you will become familiar with DBpedia by examining its contents and posing SPARQL queries. In this exercise you have to become familiar with DBpedia by browsing its web site. Pay special attention to the DBpedia ontology (https://www.dbpedia.org/resources/ontology/), which you will use to formulate your queries. Use the public SPARQL endpoint over the DBpedia data set at http://dbpedia.org/sparql to pose the following queries:

- Find all museums of Greece and present their name, hometown and the type of the museum.
- For each museum of Greece, find their hometown and present its name and number of owl:sameAs links it has.
- Find all subclasses of class Place.
- Find all the superclasses of class Place.
- Find all properties defined for the class Place together with all the properties inherited from its superclasses.
- Find all classes that are subclasses of class Thing and are found in at most 2 levels of subclass relationships away from Thing.

Exercise 2 (250 marks)

Microsoft Academic Knowledge Graph (https://makg.org/), is a large RDF data set with over eight billion triples with information about scientific publications and related entities, such as authors, institutions, journals, and fields of study. The data set is based on the Microsoft Academic Graph and licensed under the Open Data Attributions license and provides entity embeddings for all 210M represented scientific papers. First become familiar with the ontology of the data (https://makg.org/schema-linked-dataset-descriptions/) and then use the SPARQL endpoint (https://makg.org/sparql) to pose the following queries:

- Find the title and publication date of papers related to Artificial Intelligence (use keywords).
- Find the author that has the most papers with a patent.
- For papers in the area of "Computer science" find its name, the names of the authors and their affiliation. Organize the results by affiliation and author.
- For each field of study with more than 5000000 publications, retrieve the name, the date of creation and its Wikipedia article.
- For each institution present its name and rank, if a Wikipedia page for it does not exist.
- For each institution show the number of authors belong to it, the name of each author and how many publications they have.

- Give all authors that have published papers in more than two journals (you cannot use SPARQL 1.1 aggregate operators to express this query).
- Give all authors that have published more than two papers in the same conference instance (you must use SPARQL 1.1 aggregate operators to express this query).
- For each author present their name, the conference series they have published papers in and how many papers they have in each conference series.
- Find the location that has held the most conferences.

Exercise 3 (250 marks)

As we have discussed in class, http://schema.org is a major effort from the top search engine companies (Google, Bing, Yahoo and Yandex) to help web designers annotate their pages with structured information which can then be used by search engines for better indexing of these web pages.

Now that we have understood what http://schema.org/ is, let us use it to annotate the Web page of the professors in our department (https://www.di.uoa.gr/en/staff?field_staff_specialty_target_id=226). First of all read about this task on http://schema.org/docs/gs.html and familiarize yourself with the relevant technologies of Microdata, RDFa and JSON-LD. You can also see examples of using http://schema.org/ on the main web page of each of its elements (e.g., see examples of using the class Place at the bottom of the page https://schema.org/Place). Google recommends to use JSON-LD to annotate Web pages (see https://developers.google.com/search/docs/guides/intro-structured-data).

Your job in this exercise is to use the format of JSON-LD to prepare a script for the above Web page, that annotates the information regarding the professors. You should use the validator tool available at https://validator.schema.org/ to verify your code.

Exercise 4 (250 marks)

In this exercise, you will see how ontologies and KGs can be used for system diagnostics. For this exercise we focus on PEM fuel cells, which are a zero-emission, efficient and high-quality energy source that provides a future economically competitive option with respect to conventional energy sources. A drawback of these systems is that the failure of one component can cause multiple failures to the rest of the fuel cell stack, leading to a decrease in efficiency and significantly increasing maintenance and repair costs. Thus, it is crucial to notify early enough the end user of any forthcoming failures and to enable the performance of question answering.

To develop the question-answering system we need first to create a dataset with questions, queries, and answers. For this, we ask your help.

For this exercise, you are asked to download the ontology PEMFC.owl http://cgi.di.uoa.gr/~pms509/projects.htm and you will need to translate in SPARQL 1.1 ten questions sent to you privately in piazza. You will run these queries in Protégé and verify manually the results.

Bonus: Exercise 5 (100 marks)

As part of a diploma thesis in our Department, an undergrad student has developed the tool Cha2O, a "chatbot" for generating lightweight ontologies based on the competency questions given by the user. The purpose is to help non-ontology-experts to develop their own ontologies. You can download this tool from: https://cgi.di.uoa.gr/~pms509/projects.htm.

We would like you to help us evaluate this tool by doing the following:

Find a Wikipedia page and based on the content of this page generate an RDFS ontology using Protégé (https://protege.stanford.edu/).

Your ontology should contain:

a) At least 10 subclass axioms, at least 5 properties
b) Each property should have a domain and range.

Now try to do the same work using:

i) ChatGPT
ii) ChatO

Now answer the following questions:

(a) Which tool was more helpful?
(b) Did both tools generate the same ontology? What was the difference between the two?
(c) Was there any part of the initial ontology that you did not manage to recreate with any of the two ontologies and why?

Grade Chat2O giving marks from 1 to 10 (where 10 is the best mark) based on the following criteria:

(a) How helpful was the Chat2O?

(b) Would you reuse such a tool for future projects with RDFS ontologies?

(c) What level of KT-knowledge do you think is required to use Chat2O? (1 for none of it, 10 for very good knowledge of KTs)

(d) How familiar were you with the language used?

(e) Do the explanations given by the tool suffice to complete your task?

(f) To what level did the tool require from you some extra effort (e.g., to memorize information)

(g) To what level were you aware of what the tool was doing at each step?

(h) Did you often face error messages (how did you deal with them?)?

Also, please indicate any erroneous or non-expected behavior of the tool, any shortcuts that you think are missing or any other comments you feel that would be helpful for this tool.

For this homework, you will submit through e-class a zip file with the following:

1. A pdf report with all the SPARQL queries and sample results for exercises 1, 2 and 4 along with any documentation/remarks if needed. The results of the schema.org validator for your script for exercise 3. The first page should include your name and ID.

2. A txt file with the SPARQL queries for exercises 1, 2 and 4 separated by an empty line.

3. A JSON-LD file for exercise 3.

4. Three .owl files for exercise 5 one for each tool: Protégé, Chat2O, ChatGPT, and a report with your answers along with the Wikipedia page on which you based the generation of your ontology.