Queen Mary
University of London

ECS735P/U: The Semantic Web - Final Coursework

This coursework makes up 60% of the total marks for the module.

Submit your solution QM+ by the date specified in the coursework area.
You should be able to see the details of the coursework submission on your QM+ page.

The goal of this coursework is to put the ontology modelling semantic data development, Description Logic (DL) and SWRL skills into practice that you have learnt in the Semantic Web lectures and labs in a larger project.

Your task is to define, populate and query an ontology (including A-Box and a T-box) on a topic of your choice. The ontology must be able to integrate and reuse already available semantic data. At least two concepts of the ontology T-Box must be taken from external semantic data repositories. This way, the ontology will have an A-Box that can be populated with already existing data. You will use *Protégé* to design the ontology, and Python-based semantic tooling to populate the ontology with real world data.

Specifically, you should achieve the following:

Basic Task (60% coursework marks): Define your ontology in OWL2. The T-Box must be created using Protégé and should be your own work (not an existing ontology, but may import existing ontologies). Populate the knowledge base from an external semantic data repository using SPARQL 1.1. Verify that you can also query the local ontology using SPARQL.

Bonus Task 1 (20% coursework marks): As above, but your ontology should fuse information from at least two distinct external data repositories. The query to your local ontology should answer questions that cannot be answered by either remote knowledge base alone.

Bonus Task 2 (20% coursework marks): You are required to use Description Logic rules to define as many concepts as possible with the help of SWRL rules in order to compensate for the limitations of the Protégé inference engines with DL. The A-Box (individuals) must be created in a way to demonstrate the correctness and effectiveness of the logic rules defined.

You should submit a zip file with the following elements:

- **A report:** PDF document describing, in English, how you constructed the ontology: you should say where you got the data from, and you should also say what difficulties you encountered and how you solved them. The document must have also a final section explaining what source code files and models are included, and the required steps to

run the code. You are expected to submit a report explaining your assignment. If you fail to submit a report with your ontology and python files, you might receive 0% as it is the only way of proving the work is yours.

- A Protégé-OWL ontology (.owl file)
- A python script (.py) that can be used to populate the ontology from a SPARQL endpoint.
- Another python script that queries the local store to demonstrate to the user that information can be easily accessed. To test the system, the user should be able to execute any arbitrary query supported by your ontology

Marking criteria:

Basic task: Correctly designed ontology with basic taxonomy and property hierarchy (15%), correct domain and range restrictions (5%), correct and effective use of object properties (including constraints and characteristics such as functional, transitive and irreflexive), correct and effective use of data properties, logical and correct use Description Logic to define concepts (10%), Ontology population python script and SPARQL queries (20%), justification explanation and validation of the ontological modeling decision in the report (10%).

Bonus task 1: Use of appropriate, diverse data sources (5%), correct mechanism to retrieve and transform data to fit your ontology (10%), explanation of the mechanism in the report (5%). The data sources can be RDF dumps (e.g. loaded into a local database) and SPARQL end-points and may include one non-semantic dataset you convert into RDF locally for it to be queried.

Bonus task 2: Correctly working (inferencing) ontology with one of the reasoners provided with Protégé and use of SWRL rules: object and data properties and SWRL built-ins (10%). A correct A-Box with a sufficient number of individuals to use with the defined logic rules, expecting majority of the relations defined by the data properties and concepts to be inferred by the engine, not hard coded (5%), correctly commented rules and explanations in the report (5%).

Fig. 1. Example SWRL rules in Protégé using the SWRLTab plugin

This assignment is intended to be open ended and exploratory in nature. However for illustration, some examples of possible tasks could be: create and populate an ontology covering movies and cities which could be queried to find movies filmed in cities with a population less than 1M; or create and populate an ontology about companies including location, employees and profits, which could be queried to find the UK based companies with the largest profit per employee.

For convenience, we provide you some publicly available semantic web data sources that can be useful for this coursework:

Dbpedia (https://www.dbpedia.org/) provides an RDF version of the information available in the regular Wikipedia. It also provides an SPARQL endpoint for remote access https://www.dbpedia.org/resources/sparql/

You can access governmental datasets from data.gov.uk and data.gov. In some cases, you might need to download the dataset, as they don't provide a SPARQL endpoint.

Wikidata (https://www.wikidata.org/) is a free and open knowledge base that can be read and edited by both humans and machines. Wikidata acts as central storage for the **structured data** of its Wikimedia sister projects including Wikipedia, Wikivoyage, Wiktionary, Wikisource and others, with a SPARQL end point: https://query.wikidata.org/

Datasets with SPARQL end points:
https://io.datascience-paris-saclay.fr/query/List_of_datasets_with_a_SPARQL_endpoint

Over 10,000 arbitrary datasets from https://datahub.io/search (some are in RDF already).

There are links to resources and examples on QM+.

Courseworks are due on QM+ at the end of term. Please check the date specified in the coursework area of the module page. Please do not be late with the coursework otherwise a late submission penalty is applied automatically. There is a final cut-off point after one week. About six weeks should be sufficient to complete this coursework. Do NOT wait until the last week to start. **There will be no extensions.**

End of paper