Homework 4

FULYA KOCAMAN
CWID: 803023878

Due: **Check date on Canvas**. Prepare your answers as a **single PDF file**.
Group work: You may work in groups of 1-3. Include all group member names in the PDF file. Only one person in the group should submit to Canvas.

1. Consider the following table in a relational database storing the assignment of courses to classrooms in a university.

Course number (primary key)	Room	Department
CPSC-583	CS-110B	ComputerScience
CPSC-597	CS-110B	ComputerScience
CPSC-473	CS-406	ComputerScience

Convert the relational data into a Semantic Web representation. Use the following two properties:

```
http://example.org/is-located-in
http://example.org/is-offered-by
```

(a) Give the **triple** representation of the Semantic Web data. The first triple is already given (ignoring the full URI for subject and object).

PREFIX ex: http://example.org/

CPSC-583 ex:is-located-in CS-110B.

CPSC-597 ex:is-located-in CS-110B.

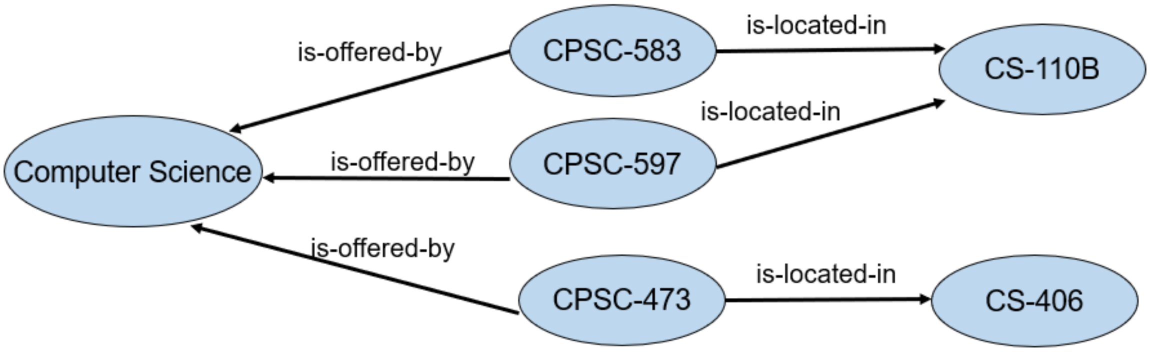
CPSC-473 ex:is-located-in CS-406.

CPSC-583 ex:is-offered-by ComputerScience.

CPSC-597 ex:is-offered-by ComputerScience.

CPSC-473 ex:is-offered-by ComputerScience.

(b) Give the **graph** representation of the Semantic Web data. The first edge is already given.



2. Read the attached paper "Data Integration through Ontology-Based Data Access to Support Integrative Data Analysis: A Case Study of Cancer Survival" by Hansi Zhang et al. (presented at the 2017 IEEE International Conference on Bioinformatics and Biomedicine).

    a. Write a short essay (approximately 500 words) that describes

        i. what is the problem the authors are trying to solve,

        ii. how Semantic Web technology provides a solution, and

        iii. what, in your opinion, are the advantages and disadvantages of their Semantic Web-based approach.

The authors of this paper propose adopting an ontology-based semantic data integration approach to improve the understanding of cancer using data coming from different sources. This semantic data integration links syntactic, schematic, and semantic heterogeneities across different data sources to create a pooled data set of integrative data analysis (IDA) that covers both the individual and the contextual level factors. In their research, they emphasize the importance of analyzing multiple factors from multiple levels, precisely both the individual and contextual level data sources for the survival of cancer patients.

For their study, they collected patients' demographic, tumor, treatment, and survival information from the 1996–2010 data, census tract-level poverty information from the 2000 U.S. census data, and also obtained 1996i2010 county-level smoking rates. All these data sources were then used to implement a relational database (MySQL). Then, they established semantic mappings between the global ontology and the data sources according to their integrative data analysis. The Ontop platform was then used to allow for semantic queries against the relational database and create an ontology-driven semantic data integration approach. They were able to integrate different data sources to form a single pooled data set and organize it into a data table. Each row represents a patient, and each column represents a risk factor. As a final step, they assembled four types of SPARQL queries to create the semantic data integration pipeline.

By following the steps of the Semantic Web technology, the authors were able to create the semantic data integration pipeline using multiple variables from multiple sources. They emphasize that once all the necessary SPARQL queries clearly specified, building the data integration. The pipeline was straightforward with the Ontop OWL API. They also conclude that the semantic data integration approach formed from linking heterogeneous data sources can automatically encode the knowledge from many data processing in the ontology without worrying about the syntactic, schematic, and semantic heterogeneities in data from different sources.

In my opinion, the advantages of their Semantic A web-based approach would be semantically integrating data and explicitly expressing the semantic relationships among variables from different sources. Modeling the semantic relationships explicitly is a very nice way to represent data processing and integration steps clearly, and it makes the data easily understandable to both humans and computers. Also, since the dependencies and constraints of the data elements are explicitly modeled in their approach, data quality and consistency checks can easily be automated. The use of an ontology-driven semantic data integration would be a great asset to use in any field because we can easily get data from a variety of sources.

On the other hand, I do not believe their Semantic web-based approach assembled with four types of SPARQL queries covers all possible cases. More cancer research on Semantic web-based approach is needed to expand their ontology. There could also be some issues with manipulating and integrating contents from heterogeneous sources such as difficulty forming standardized common vocabulary across different sources, challenges on storage and organization of the Semantic Web contents, and issues related to finding the right content in the Semantic Web.

b. Figure 2 in the paper is a fragment of the ontology they developed. Using only this information, write a SPARQL query to retrieve a list of all patients who have lung cancer and who live in a county with a high rate of smoking (ocrv:avg_smoke > 0.2). Hint: you can look at some of the example queries in Tables 4-7.

```
PREFIX : <http://www.sematicweb.org/ontologies/OCRV#>

SELECT ?patient ?country ?smoke

WHERE {

    ?patient a ocrv:patient.

    ?patient ocrv:has_disease ocrv:lung_cancer.

    ?patient ocrv:lives_in ?country.
```

```
        ?country a ocrv:county.

        ?country ocrv:avg_smoke ?smoke.

        FILTER (?smoke > 0.2).

    }
```

3. The SPARQL queries that were demonstrated in class are in the sparql_examples.txt file on Canvas. These queries can be executed on the SPARQL endpoint for the dbpedia dataset: http://dbpedia.org/snorql/. Write additional queries to list:

    a. all the systems of government in India (use property `dbo:governmentType`)

```
SELECT ?country ?gov

WHERE {

        ?country a dbo:Country.

        ?country rdfs:label "India"@en.

        ?country dbo:governmentType ?gov.

}
```

    b. all countries (names in English only) which are republics (i.e., `dbo:governmentType` is `:Republic`)

```
    SELECT ?name

    WHERE {

        ?country a dbo:Country .

        ?country rdfs:label ?name.

        FILTER (lang(?name) = "en").

        ?country dbo:governmentType :Republic.

    }
```

c. all countries (names in English only) which are republics *and* a Unitary state (i.e., `dbo:governmentType` is `:Republic` and also `:Unitary_state`)

```
SELECT ?name
WHERE {
        ?country a dbo:Country .
        ?country rdfs:label ?name.
        FILTER (lang(?name) = "en").
        ?country dbo:governmentType :Republic.
        ?country dbo:governmentType :Unitary_state.
}
```