PMS 547 Foundations of databases (and knowledge bases)
Homework I
Due on April 13, 2010

Exercise 1 (FOAF). Build a simple personal Web page using HTML and publish it on the Web. This Web page can provide only very basic information about you that machines cannot understand. Use FOAF (http://www.foaf-project.org/) to describe yourself and your life (well, may be only a part of it!) in RDF (a machine readable format) and post this information on your Web page so that it is publicly accessible. You can use FOAF-a-Matic to create a FOAF description of yourself. Link it to FOAF information of other participants of the course. Add some additional properties that are included in the FOAF Vocabulary Specification. Use the W3C RDF validation service to ensure that your RDF is valid. Do not hesitate to include a picture and geographical information! Make sure that you include the appropriate *HTML Link tag* and then a Google search can be used to help discover FOAF files across the web. You can also view your FOAF profile with FoaF Explorer and navigate through your friends' profiles as well.

In this exercise we expect the whole class to create FOAF profiles and connect with each other so that a FOAF class graph emerges. We will not give any marks if you just try to avoid this question by doing as little as you can.
[0 or 30 marks]

Exercise 2 (Querying the Semantic Web Dog Food dataset using SPARQL)
The Semantic Web Dog Food dataset available at http://data.semanticweb.org/ is a popular dataset created by Semantic Web people to describe Semantic Web events such as conferences and workshops, their participants etc. This exercise involves posing SPARQL queries against this dataset. We stress that it is part of the exercise to understand the details of this publicly available dataset and related ontologies (we will not explain it to you in every detail). Browsing these datasets (possibly using a Semantic Web browser such as Tabulator available at http://www.w3.org/2005/ajar/tab), posing SPARQL queries on them etc. will help you to understand all the details.

First, we ask you to understand the vocabularies/ontologies used in this data set by studying the FAQ at http://data.semanticweb.org/documentation/user/faq. The main ontology used in this dataset is the Semantic Web Conference Ontology (SWC, http://data.semanticweb.org/ns/swc/swc_2009-05-09.html), an ontology for modeling entities of research communities such as persons, organisations, publications and their relationships. SWC is written in OWL and is available at http://data.semanticweb.org/ns/swc/swc_2009-05-09.rdf. We have exported the RDFS class hierarchy of the SWC ontology and uploaded it at http://www.di.uoa.gr/~pms509/swc-rdfs-hierarchy-e.rdf. Load this file in Sesame and use SPARQL to express the following queries based on this RDFS class hierarchy:

- Find all subclasses of class OrganisedEvent.
- Find all the superclasses of class Presenter; exclude predefined RDFS/OWL classes.

- Find all the subclasses of class Vevent that are not a subclass of class AcademicEvent.
- Find all the classes; exclude predefined RDFS/OWL classes.
- Find all classes and their proper (direct) superclasses; exclude predefined RDFS/OWL classes.
- Find all classes that do not have superclasses; exclude predefined RDFS/OWL classes.

Now consider the dataset for the conference ISWC 2009 (http://data.semanticweb.org/conference/iswc/2009/complete). Load the dataset in Sesame and use SPARQL to express the following queries:

- Find the title and authors of each paper in the proceedings of ISWC 2009.
- Find the title and authors of each paper in the proceedings of ISWC 2009 paper that is on the topic of Linked Data.
- For each paper in the "Semantic Web in Use" track of ISWC 2009, print the title and the full names of all authors.
- For every person that had a role in ISWC 2009 (for example Program Committee chair, Tutorial chair etc.), print his/her name, the papers that he/she possibly has in the ISWC 2008 proceedings and the co-authors of these papers.
- Give the name and role of every person that had some important role in the organization of ISWC 2009 (for example Program Committee chair, Tutorial chair etc.), but has not published a paper at the conference.
- Give the names of all authors that had at least three papers in the proceedings of ISWC 2009.
- Find the authors that are co-authors of Stefan Decker in all papers that Stefan has in the proceedings of ISWC 2009.
- Find the authors that have published only papers with the string "OWL" in their title in the proceedings of ISWC 2008.
- Find the paper presentations that people attending the presentation of paper "Efficient Query Answering for OWL 2" could not attend at ISWC 2009.

Now we ask you to also consider the data set of the conference ISWC 2008 (http://data.semanticweb.org/conference/iswc/2008/complete). Use the named graph capabilities of SPARQL to express the following queries:

- Find all authors that had a paper at both ISWC 2008 and ISWC 2009.
- Find all authors that had at least three papers at ISWC 2008 and at least three papers at ISWC 2009.
- Find all authors that had at least three papers at ISWC 2008 but less than three at ISWC 2009.
- Construct an RDF graph that contains all relevant information about papers with the string "OWL" in their titles that appear in ISWC 2008 or ISWC 2009.

You can choose to use Sesame (http://www.openrdf.org/) through its Java API (http://www.openrdf.org/doc/sesame2/users/ch08.html) to load the appropriate files and pose the SPARQL queries or use a SPARQL endpoint that we will give you. Those who choose to use Sesame will get extra bonus marks.

[200 marks]

Exercise 3 (Linked Data - DBpedia)

Linked Data (http://linkeddata.org/) is a term that refers to a set of best practices for creating and connecting data sources on the Web using URIs and RDF. Tim Berners-Lee has been quoted as saying that Linked Data is "the Semantic Web done right" (http://linkeddata.org/faq). There is currently lots of data sources that have been linked together to create a Web of Linked Data (http://www4.wiwiss.fu-berlin.de/bizer/pub/lod-datasets_2009-03-05.html) and other data sources are continuously added. One of the central data sources in this web is DBpedia (http://dbpedia.org/About), a big knowledge base which is essentially a "translation" of parts of Wikipedia into RDF.

In this exercise you will become familiar with DBpedia and other Linked Data sources by examining their contents and posing SPARQL queries. More specifically, you have to do the following:

- Become familiar with DBpedia by browsing its web site (http://dbpedia.org/About). Pay special attention to the DBpedia ontology (http://wiki.dbpedia.org/Ontology?v=ril) which you will use to formulate your queries. Browse the Wikipedia knowledge captured by DBpedia starting from a resource that you know well e.g., the writer Nikos Kazantzakis (http://dbpedia.org/page/Nikos_Kazantzakis) and following links to other DBpedia resources.[1] Do the example queries given in http://wiki.dbpedia.org/OnlineAccess?v=53r .
- Use the public SPARQL endpoint over the DBpedia data set at http://dbpedia.org/sparql to pose the following queries:
    - Find all graduates of the National and Kapodistrian University of Athens that became prime ministers and optionally their pictures and the names of their wives.
    - Find all artists that have been born in a state which includes a city with more than 5 million inhabitants, in a country which is not a member of the European Union and they have represented Greece in the Eurovision contest. Output the artist name, state of birth and Eurovision song. (Συγνώμη, το κατεβάσαμε λίγο το επίπεδο εδώ!).
    - Find all books that were published in the years 1950-1960 and were subsequently made into films. Output the name of the book, the name of the author(s), the name of the film and the name of the director.
    - Find all authors of books in DBpedia that were born in Greece. Output a list of authors in descending order of how many of their books have a DBpedia entry.

      To answer this query you need to pose a SPARQL query with aggregates. The SPARQL W3C recommendation does not include aggregate functions yet (but see current work on this topic by the W3C

[1] Notice that the DBpedia information about a Wikipedia resource is not structured, might contain redundant information etc. So the purpose of this exercise is to make you familiar with this data source, to make you think about the issues involved in creating and maintaining such RDF data sources and to get you to do your best in exploring such data sources using SPARQL queries.

SPARQL working group at http://www.w3.org/TR/2009/WD-sparql11-query-20091022/). However, the Virtuoso RDF store that is used by the DBpedia public SPARQL endpoint allows you to use aggregate functions with SPARQL. See http://docs.openlinksw.com/virtuoso/rdfsparql.html#sparqlextensions for more details.

- Find all computer scientists in DBpedia that had a paper at ISWC 2009. Print the name of the computer scientist, the place he/she was born, the title of the paper and the name of the co-authors if any.

  This query involves accessing two data sets: DBpedia and the ISWC 2009 data set we discussed in Exercise 1. How can you answer this query? Discuss possible alternatives, give arguments for or against each alternative and explain your implementation choice.

[70 marks]