Guidelines for the Data Integration Project – Final Version

1. Choose a domain of interest (e.g., movies, toto, travel booking, tournaments, etc.) for which you can identify a collection of (possibly open) data sources that need to be integrated. The collection should contain data from at least three different sources.
2. Model the domain of interest by means of an ontology/mediated schema. The ontology should represent all information that is relevant for the domain of interest, and it should be rich enough (at least 10-15 classes, plus the corresponding object and data properties).
3. Represent the ontology/mediated schema using a graphical notation (e.g., as a UML class diagram).
4. Represent the ontology in Protégé.
5. Connect the data sources to a Denodo instance, and design the corresponding federated relational schema. Identify relevant constraints for the schema and represent them in Denodo.
6. Design VKG mappings to connect the ontology to the federated schema, using the Ontop plugin for Protégé.
7. Develop a Java application for your domain that makes use of Ontop as a SPARQL endpoint to query the database through the ontology, extracting information that is of interest for your domain of choice.
   As an example for the kinds of queries that could be posed via your application, you can consider the queries underlying travel booking sites, where some parameters of a request are filled in via a form (e.g., the departure city, departure and arrival date and time, etc.), and answers are retrieved using those parameters. (You have to take into account the SPARQL fragment that is supported by the current version of Ontop.)

Notes:

- The project can be developed either alone, or in a group of two. For projects developed by a group of two, only one of the two students should submit the project.
- Before starting the effective development of the project, present and discuss with the lecturer the domain and the data sources that you intend to use, and the application that you intend to develop.
- Any doubt that you have should be discussed before taking decisions that then might force you to revise your work.
- The documentation produced for the project should include a *self-contained pdf document*, which will be the basis for the discussion of the project during the oral exam. The document should contain:
  1. a header with title of the project, name(s) of the student(s) that have developed it, name of the course (i.e., Data Integration), and academic year;
  2. a description of the domain of interest;
  3. a description of the content and format of the selected data sources, and an indication of how and where these data sources can be accessed;
  4. a description of the main functionalities of the (Java) application that you have developed;
  5. the ontology/mediated schema of the domain of interest, expressed in OWL 2 QL; the ontology should be represented in the document as a UML class diagram, with (binary) associations representing object properties, and class attributes representing data properties;
  6. a (possibly graphical) representation of the relevant relations exported by Denodo, indicating attributes, keys, and foreign keys;
  7. a specification of the mappings between the mediated schema and the relational schema exported by Denodo, expressed in the Ontop mapping language; you might include in the document only a representative subset of the whole set of mappings, in order to avoid to specify many mappings that are very similar in their form.
- The documentation should include also:
  - the `.owl`, `.obda`, and `.properties` files that make up the Ontop VKG specification of the project;
  - the export of the Denodo database schema, generated (with extension `.vql`) through the Export function of Denodo;
  - the csv and/or json files and/or a dump of the SQL databases that have been connected to Denodo (provided they are not larger than 10MB - if the files are larger, a subset of the rows should be extracted);
  - some SPARQL queries that have been used either for testing or in the developed application (these queries can be saved from the SPARQL query tab of Protégé as a file with extension `.q` that is being placed in the same directory as the files that make up the Ontop VKG specification).
- All documents specified in the previous two items have to be bundled in a single ZIP file and have to be uploaded to OLE within the "Data Integration Project" assignment.
- The software part of the project (ontology, specification of the mediated schema, mappings, application) will be discussed and demoed during the oral exam.
- The deadline for submitting the project is at 23:55 two days before the day set for the project discussion.