Guidelines for the Data Curation Final Project

For the Data Curation Final Project, you should choose a domain of interest (e.g., movies, travel booking, sports, 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 two different sources.

Data Profiling Part

1. Export the datasets of interest and perform Elementary Data Analysis (EDA) by means of either existing tools and libraries or your own EDA algorithm implementations. Report the results of the EDA and comment on them in the final project documentation.
2. Design one or more relational databases to structure and store the exported datasets and show how the dependency discovery algorithms (see, UCC, FD, and IND) introduced in the course have been used to extract metadata that supported the database/s designing process.
3. The results obtained by profiling the datasets for dependency discovery, and the consequent database schema design choices, must be properly commented in the final project documentation.

Data Preparation and Integration Part

1. Consider the domain (and the data-sets within that domain) that you have chosen for the *Data Profiling* part, and model the domain 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 classes, plus the corresponding object and data properties).
2. Represent the ontology/mediated schema using the graphical notation introduced in the course (or similar).
3. Represent the ontology in Protégé.
4. Consider the relational database (or databases) that you have designed for the Data Profiling module, including the relevant constraints that you have specified and/or extracted from the data.

   Note that the choice of having multiple databases, instead of a single one, has an impact on the technological solution that you will need to deploy. In particular, having multiple databases to integrate implies the necessity to rely on a federation system (like Teiid, Denodo, or Dremio, for instance) as an intermediate layer between the sources and ontop. This is obviously not needed if you work with a single database.
5. Design VKG mappings to connect the ontology to the database (or the federated relational schema exposed by the federation system), using the *Ontop Plugin for Protégé*.
6. Develop an application (e.g., in Java) 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. During the project discussion (see later), both students should be aware of all parts of the project, and they should be able to answer autonomously (i.e., without relying on the other student) any question that concerns the project.
- Before starting the effective development of the project, present and discuss with the lecturers 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 Curation*), 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 profiling analyses that have been performed and a critical analysis of the obtained results (e.g., which techniques have been applied and for which purpose, what the results convey about the datasets at hand and the value distributions, significant attribute correlations, how the profiling results have been exploited in schema design and data cleansing, etc.);
  5. the ontology/mediated schema of the domain of interest, expressed in OWL 2 QL; the ontology should be represented in the document adopting the graphical notation presented in the course;
  6. a (possibly graphical) representation of the relevant relations in the relational schema, indicating attributes, keys, and foreign keys;
  7. a specification of the mappings between the ontology and the relational database, 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;
  8. a description of the main functionalities of the application that you have developed.
- The documentation should include also:
  - the `.owl`, `.obda`, and `.properties` files that make up the Ontop VKG specification of the project;
  - the SQL export of the relational database schema;
  - the csv and/or json files and/or a dump of the SQL database mapped to the ontology (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.
- All documents that have to be handed in have to be bundled in a single ZIP file and have to be uploaded to MS Teams within the "Data Curation Project" assignment in the "Data Preparation and Integration" Team.
- The software part of the project (profiling algorithms, ontology, specification of the mediated schema, mappings, application) has to be discussed and demoed during the oral exam.
- The deadline for submitting the project will be communicated in Teams (typically it is at 23:55 two days before the oral exam date, which is the day set for the project discussion).