Fraunhofer FIT

RWTH AACHEN UNIVERSITY

PD Dr. C. Quix, Dr. Andreas Pippow
June 22, 2015

Mid-Term Project "Scientific Data Management"

Due on July 6, 2015 SoSe 2015

General Information:

The goal of this course project is to build an integrated database for the metadata of a collection of files which have been created during some bio-medical experiments.

The solutions to the tasks should be documented in a single PDF document. Additional files with models, mappings, or queries can be also included in the solution if necessary, and can be referenced from the main document. Do not submit complete database files or virtual machine images. All files should be uploaded as a single ZIP file to the L2P system by **July 6, 2015**. The project should be done in groups of 3 to 4 students.

Task 1 [Extraction] :

The sources are given as CSV files `mergedLists_*.txt` including a single header line and data items. Use Talend Open Studio for Big Data [1] to load the data of the CSV files into a MongoDB database. The documents (or objects) in MongoDB should have a complete representation of the data in the CSV files; however, you are free to define a suitable structure of the documents.

Required documentation: Briefly describe the extraction job(s) which you have created in Talend and describe the structure of the documents in MongoDB.

Task 2 [Integration] :

The schema of the CSV files is identical to a large degree, but a few columns differ in name or coding scheme. These differences should be resolved in an integrated database. The file `synonyms.xlsx` contains a list of terms and their preferred names. Note that the list of synonyms applies to metadata (column names) as well as to individual data items (values in rows). Use Talend Open Studio for Big Data to integrate the documents loaded in the first task into an integrated database.

Required documentation: Briefly describe the integration job(s) which you have created in Talend, describe the structure of the integrated documents, and provide the mappings from the original documents to the integrated documents (screenshots are sufficient).

[1] Only this edition includes connectors for MongoDB.

Task 3 [Semantic Annotation] :

The data in the CSV files makes references to some established biological terms which are defined in various bio-medical ontologies. The file `listOfObjects.xlsx` contains a list of these terms and the corresponding 'preferred name' which has been used in our dataset. Note again that the annotations apply to metadata as well as to data items.

Extend the document structure of the integrated database and annotate the data with the references to the corresponding terms and ontologies. For those elements for which the ontology and ID is 'not defined', provide extensions for existing ontologies to define these terms.

Required documentation: Briefly describe the job(s) which you have created in Talend to do the annotation, describe the components/operators which you have used in these jobs, and describe how you change the structure of the documents to capture the annotations.

Task 4 [Queries] :

Define the following queries in MongoDB and use the semantic annotations which you have created in the previous task.

1. List the files which have been last modified in June 2015 and which deal with the cell type 'BHK570 VII high producer'.

2. List all files with their name, their measurement position, and the objective lens.

3. Count the files in which 'Gelatine' has been used as a medium.

4. List all files with location 'nucleus' and growth medium 'medium'. Determine which cell types, durations and objective lenses have been used for the resulting list. For each combination of categories you determined (e.g., cell type: BHK 570 cell; location: nucleus; duration: 20sec; growth medium: medium; objective lens: 100oi), find the corresponding background file(s) (i.e., all categories are the same as before, but location is background).

Required documentation: Provide the MongoDB queries and the number of elements in the result. The last query might require additional transformations or the storage of intermediate results; please describe your solution briefly, if it is more complex than a single MongoDB query.