Ontology-based Retrieval of Scientific Data in LIFE

Alexandr Uciteli[1,2], Toralf Kirsten[2,3]

[1]Institute for Medical Informatics, Statistics and Epidemiology, University of Leipzig
[2]LIFE Research Centre for Civilization Diseases, University of Leipzig
[3]Interdisciplinary Centre for Bioinformatics, University of Leipzig

Abstract: LIFE is an epidemiological study determining thousands of Leipzig inhabitants with a wide spectrum of interviews, questionnaires, and medical investigations. The heterogeneous data are centrally integrated into a research database and are analyzed by specific analysis projects. To semantically describe the large set of data, we have developed an ontological framework. Applicants of analysis projects and other interested people can use the LIFE Investigation Ontology (LIO) as central part of the framework to get insights, which kind of data is collected in LIFE. Moreover, we use the framework to generate queries over the collected scientific data in order to retrieve data as requested by each analysis project. A query generator transforms the ontological specifications using LIO to database queries which are implemented as project-specific database views. Since the requested data is typically complex, a manual query specification would be very time-consuming, error-prone, and is, therefore, unsuitable in this large project. We present the approach, overview LIO and show query formulation and transformation. Our approach runs in production mode for two years in LIFE.

1 Introduction

Epidemiological projects study the distribution, the causes and the consequences of health-related states and events in defined populations. The goal of such projects is to identify risk factors of (selected) diseases in order to establish and to optimize a preventive healthcare. LIFE is an epidemiological and multi-cohort study in the described context at the Leipzig Research Centre for Civilization Diseases (Univ. of Leipzig). The goal of LIFE is to determine the prevalence and causes of common civilization diseases including adiposity, depression, and dementia by examining thousands of Leipzig (Germany) inhabitants of different ages. Participants include pregnant women and children from 0-18 as well as adults in separate cohorts. All participants are determined in a possible multi-day program with a selection out of currently more than 700 assessments. To these assessments belong interviews and self-completed questionnaires to physical examinations, such as for anthropometry, EKG, MRT, and laboratory analyses of taken specimen. Data is acquired for each assessment depending on the participant's investigation program using specific input systems and prepared input forms. All collected data is integrated and, thus, harmonized in a central research database. This database consists of data tables referring to assessments (i.e., investigations). Their complexity ranges from tables with a small number of columns to tables with a very high column number. For example, the table referring to the Structured Clinical Interview (SCID) consists of more than 900 columns, i.e., questions and sub-questions of the interview input form.

The collected data are analyzed in an increasing number of analysis projects; currently, there are more than 170 projects active. Each project is initially specified by a proposal documenting the analysis goal, plan and the required data. However, there are two key aspects that are challenging. Firstly, the applicant needs to find assessments (research database tables) of interest to specify the requested data in the project proposal. This process can be very difficult and time consuming, in particular, when the scientist is looking for specific data items (columns), such as weight and height, without knowing the corresponding assessment. Secondly, current project proposals typically request data from up to 50 assessments which are then organized in project-specific views (according to the data requests). These views can be very complex. Usually, they combine data from multiple research database tables, several selection expressions and a multitude on projected columns out of the data tables. A manual specification of database queries to create such views for each analysis project would be a very error-prone and time-consuming process and is, therefore, nearly impossible. Hence, we make the following contributions.

- We developed an ontological framework. The framework utilizes the LIFE Investigation Ontology (LIO) which classifies and describes assessments, relations between them, and their items.
- We implemented ontology-based tools using LIO to generate database queries which are stored as project-specific analysis views within the central research database. The views allow scientists and us to easily access and to export the requested data of an analysis project. Both, LIO and ontology-based tools are running in production mode for two years.

The rest of the paper is organized as follows. The Section 2 describes the ontological framework and especially LIO. The Section 3 deals with ontology-based query formulation and transformation, while Section 4 describes some implementation aspects. Section 5 concludes the paper.

2 Framework

The goal of the ontological framework is to semantically describe all integrated data of biomedical investigations in LIFE using an ontology. The ontology is utilized on the one hand by scientists to search for data items or complete investigations of interest or simply to browse the ontology to get information about the captured data of the investigations. On the other hand, the ontology helps to query and retrieve data of the research database by formulating queries on a much higher level than SQL.

The ontological framework consists of three interrelated layers (Fig. 1). The integrated data layer comprises all data elements (instance data) of the central research database providing data of several source systems in an integrated, preprocessed and cleaned fashion. The metadata layer describes all instance data of the research database on a very technical level. To these metadata belongs the used table and column names, corresponding data types but also the original question or measurement text and the code list when a predefined answer set has been originally associated to the data item. This metadata is stored in a dedicated metadata repository (MDR) and is inherently interrelated with

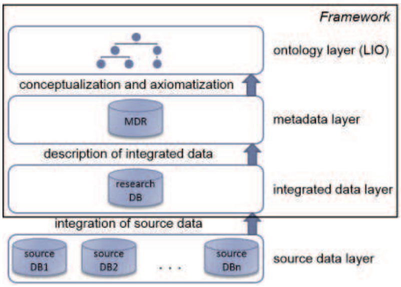

Figure 1: Framework overview

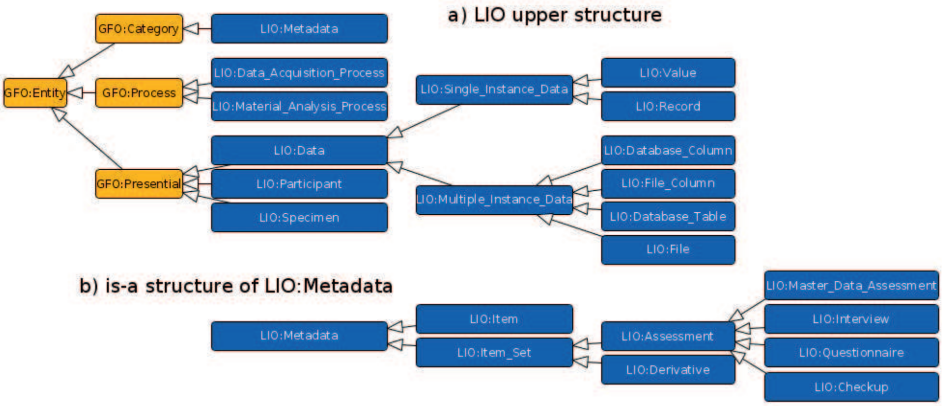

Figure 2: Selection of the LIFE Investigation Ontology

the instance data. Finally, the ontology layer is represented by the developed LIFE Investigation Ontology (LIO) [KK10] and its mapping to the collected metadata in the MDR.

LIO utilizes the General Formal Ontology (GFO) [He10] as a top-level ontology and, thus, reuses defined fundamental categories of GFO, such as *Category*, *Presential* and *Process*. Fig. 2a gives a high-level overview over LIO. Subcategories of *GFO:Presential* refer to collected scientific data, participants and specimen. Scientific data is structured on a technically level by categories within the sub-tree of *LIO:Data*. Instances of these categories are concrete data files and database tables, e.g., of the research database. They are used to locate instance data for later querying. Subcategories of *GFO:Process* refer to processes of two different types, data acquisition and material analysis processes. Instances of these categories are documented, e.g., by specific states and conditions of the examination, the examiner conducting the investigation etc. This process documentation is additionally specified to the scientific data that they possibly generate. The documentation can be used for downstream analysis of the scientific data, to evaluate the process quality and for an impact analysis on the measurement process. Finally, subcategories of *GFO:Category* are utilized to semantically classify biomedical investigations in LIFE. Fig. 2b shows an overview of main categories. Fundamentally, we differentiate between items (e.g., questions of a questionnaire) and item sets. This separation allows us to ontologically distinguish between data tables (item set) of the research database and its columns (item). Moreover, we are able to classify both, investigation forms as predefined and rather static item sets and specific project-specific analysis views which potentially include items from multiple investigation forms. The latter can be dynamically defined and provided by a user group.

All biomedical investigations together with their containing items (i.e., questions and measurements) are associated to LIO categories using the *instance_of* relationship type. Fig. 3 shows an example; the interview socio-demography consists of several questions including for country of birth, material status and graduation. Both, the interview and its items, are associated to *LIO:Interview* and *LIO:Item*, respectively. Special relationships with type *has_item* represent the internal structure of the interview. The semantic classification, i.e., the association to LIO subclasses of *LIO:Metadata*, is manually specified by the investigator in an operational software application from which it is imported into

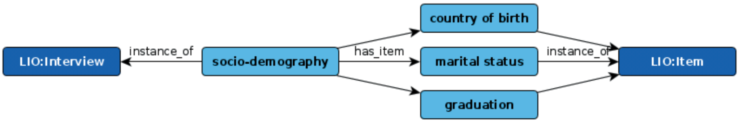

Figure 3: Utilization of item set and item categories to describe investigations

the overall metadata repository (MDR). Moreover, the MDR captures and manages the structure of each investigation form, and, thus, its items, and their representation in the central research database. By reusing both kinds of specifications in LIO, the mappings between collected metadata in the MDR and LIO categories are inherently generated. This makes it easy to describe and classify new investigations (assessments) in LIO; it necessitates only an initial manual semantic classification and import of corresponding metadata into the MDR.

3 Ontology-based Query Formulation and Transformation

Scientific data of the central research database are analyzed in specific analysis projects. Each of them is specified by a project proposal, i.e., the applicant describe the analysis goal, the analysis plan and the data she request for. In the simplest case, the data request consists in a list of assessments. This is extended, in some cases, by defined inclusion or exclusion criteria. In complex scenarios, the applicant is interested in specific items instead of all items of an assessment. In all scenarios, data can be queried per single assessment. However, it is common to request data in a joined fashion, especially, when the applicant focuses on specific items from multiple assessments. Currently, each data request is satisfied by specific analysis views which are implemented as database views within the relational research database.

We use LIO to formulate queries over the scientific data which are finally transformed and stored as project-specific analysis views in the research database. Fig. 4 sketches the query formulation and transformation process. Firstly, LIO is used to formulate queries for each analysis project. The applicant can search for assessments of interest browsing along LIO's structure and the associated instances, i.e., concrete assessments or items. She can select complete assessments as predefined item sets and specific items of an assessment. These selections are used by the query generator to create the query projection, i.e., the items for that data should be retrieved. These items are firstly sorted by the selected assessment in alphabetic order and, secondly, by their rank within each assessment, i.e., with respect to their position on the corresponding input form. Inclusion and exclusion criteria can be specified on item level. The query generator interrelates single conditions by the logical operator AND and creates the selection expression of the resulting query for each assessment.

Per default, the query generator produces one query for each selected assessment or item set of an assessment. Moreover, experienced users can create new item sets containing items from multiple assessments. These item sets result in join queries using patient identifiers and examination time points (due to recurrent visits) as join criteria. To find out which item (column) contains patient identifiers and time points, the items are specifically labelled when the assessment definitions (source schema) are imported into the MDR. Some sources allow an automatic labelling (using source-specific rules), whereas other sources need a manual intervention to fully describe their schemas and the resulting items of LIO.

The query generator takes the ontology-based specifications (selections and conditions) using LIO as input and firstly creates SQL-like queries. These queries have intermediate character and utilize keywords SELECT, FROM and WHERE with the same meaning as in SQL. In contrast to SQL, they contain ontology categories and associated instances as placeholder which are then finally replaced by conrete table and column names of the research database when SQL queries are generated. To resolve table and column names the query generator utilizes mappings between LIO and the MDR.

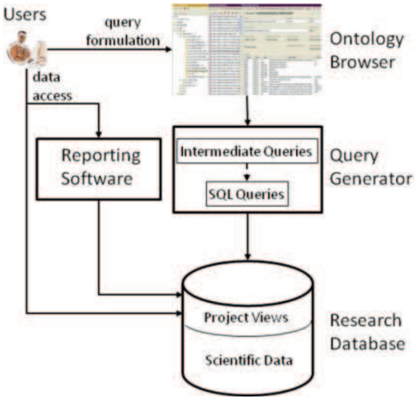

Figure 4: Query fromulation and data access

4 Implementation

LIO currently consists of 33 categories, more than 700 assessments and ca. 120 analysis results (latter two are instances in LIO) together with more than 39,000 items in total. The large and increasing number of assessments, their containing items and their correspondences (mappings) to database table and column metadata are stored in the MDR which is implemented in a relational database system. Assessments and items are loaded on demand from the MDR and are associated to LIO categories as instances. Therefore, new assessments can be easily added to LIO and without modifying the LIO's core structure or changing ontology files.

We implemented a Protégé [NFM00, Sc11] plug-in loading LIO and corresponding instances from the MDR to support an applicant when she specifies the required data for a project proposal. She can navigate along LIO's structure and, hence, is able to find and pick the items of interest for her proposal. On the other hand, the plug-in allows us to formulate and to transfer ontology-based queries into SQL-queries which are then stored as project-specific analysis views over the scientific data of the research database. These views can be access in two different ways. Firstly, the views can be used for further database-internal data processing using the database API and SQL. This is the most preferred way for persons with database skills. Secondly, the plug-in includes options to propagate views to a web-based reporting software which wraps the database views in tabular reports. These reports can be executed by an applicant. The retrieved data are then available for download to continue data processing with special analysis tools, such as SPSS, R etc.

There are other approaches which are highly related to our ontology-based framework. i2b2 [Mu10] is a framework for analyzing data in clinical context. In contrast to our approach, it utilizes a separate data management and, thus, necessitates additional data load and transformation processes. Moreover, the goal of i2b2 is primarily to find relevant patients and not to retrieve scientific data. Like LIO, the Search Ontology [Uc14] is used to formulate queries over data. Its focus is on queries for search engines, while our approach focuses on structured data in a relational database. Similar to LIO, the Ontology of Biomedical Investigations (OBI) [Br10] classifies and describes biomedical investigations. In contrast to OBI, LIO utilizes a core structure which is dynamically extended by assessments fully described in a dedicated metadata repository. Hence, our framework is able to generate queries over data of the research database and prevents from describing each investigation in detail by using OBI.

5 Conclusion

We introduced an ontology-based framework to query large and heterogeneous sets of scientific data. The framework consists of the developed LIFE Investigation Ontology (LIO) on the top level which semantically describes scientific data of the central research database (base level). Both levels are interrelated by (technical) metadata which are managed in a metadata repository. LIO gets insight which data are available within the research database, on the one hand, and is used, on the other hand, to formulate queries over the collected scientific data. The ontology-based queries are transformed into database queries which are stored as analysis-specific database views. The queries include per default items of a single assessment. Moreover, join queries merging items from multiple assessments are also supported. Together, ontology-based querying simplifies the data querying for end users and frees IT-people from implementing rather complex SQL queries. In future, we will extend LIO and the query generator to overcome current limitations, e.g., according to the specification and transformation of query conditions.

Acknowledgment: This work was supported by the LIFE project. The research project is funded by financial means of the European Union and of the Free State of Saxony. LIFE is the largest scientific project of the Saxon excellence initiative.

References

[Br10] Brinkman, R. R. et al.: Modeling biomedical experimental processes with OBI. In Journal of biomedical semantics, 2010, 1 Suppl 1; pp. S7.

[He10] Herre, H.: General Formal Ontology (GFO): A Foundational Ontology for Conceptual Modelling. In (Poli, R.; Healy, M.; Kameas, A. Eds.): Theory and Applications of Ontology: Computer Applications. Springer Netherlands, Dordrecht, 2010; pp. 297–345.

[KK10] Kirsten, T.; Kiel, A.: Ontology-based Registration of Entities for Data Integration in Large Biomedical Research Projects. In (Fähnrich, K.-P.; Franczyk, B. Eds.): Proceedings of the annual meeting of the GI. Köllen Druck+Verlag GmbH, Bonn, 2010; pp. 711–720.

[Mu10] Murphy, S. N. et al.: Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). In Journal of the American Medical Informatics Association JAMIA, 2010, 17; pp. 124–130.

[NFM00] Noy, N. F.; Fergerson, R. W.; Musen, M. A.: The Knowledge Model of Protégé-2000: Combining Interoperability and Flexibility. In (Goos, G. et al. Eds.): Knowledge Engineering and Knowledge Management Methods, Models, and Tools. Springer Berlin Heidelberg, Berlin, Heidelberg, 2000; pp. 17–32.

[Sc11] Schalkoff, R. J.: Protégé, OO-Based Ontologies, CLIPS, and COOL: Intelligent systems: Principles, paradigms, and pragmatics. Jones and Bartlett Publishers, Sudbury, Mass., 2011; pp. 266–272.

[Uc14] Uciteli, A. et al.: Search Ontology, a new approach towards Semantic Search. In (Plödereder, E. et al. Eds.): FoRESEE: Future Search Engines 2014 - 44. annual meeting of the GI, Stuttgart - GI Edition Proceedings P-232. Köllen, Bonn, 2014; pp. 667–672.