Graphical Query Construction over Scientific Data Sets using Semantic Technologies

Bartosz Balis\(^1\), Marian Bubak\(^{1,2}\), Tomasz Grabiec\(^2\)

\(^1\)AGH University of Science and Technology, Department of Computer Science, Krakow, Poland
\(^2\)AGH University of Science and Technology, ACC CYFRONET AGH, Krakow, Poland

I. INTRODUCTION

Semantic Web in which a well-defined meaning is associated with information, is among major visions of the Future Internet. Also in e-Science, which relies on integration of diverse heterogeneous resources (data, services, etc.), semantic technologies play an increasingly important role. RDF (Resource Description Framework) has been proven as a suitable model for representing scientific data, with OWL (Web Ontology Language) enhancing RDF data with well-defined semantics. Available literature mentions advantages of using RDF / OWL to describe and represent scientific data over the alternative relational model \([2]\). An extensive catalogue of publicly available RDF data sets, published by W3C, contains a substantial amount of scientific data, such as the GeoSpecies database describing biological species (over 2M triples) or the protein database Uniprot (over 300M triples).

One of barriers for effective utilization of RDF/OWL-represented scientific data is the lack of methods and tools facilitating information retrieval, on the one hand enabling formulation of advanced queries, on the other hand oriented towards domain specialists knowing little about query languages such as SPARQL. Existing tools \([3,4,5]\) have numerous limitations: they are tied to a specific domain, offer limited query capabilities, or require expert-level knowledge of the underlying query language.

We propose a graphical approach for formulating semantic queries over data sets represented in RDF and described by ontologies in OWL. The approach consists in using ontologies as a query language familiar to the domain specialist who, using a high-usability graphical user interface, constructs advanced queries which are translated to the underlying data model and executed. Based on this idea, we have implemented the QUaTRO2 tool and applied it to query the UniProt database.

II. QUaTRO2: SEMANTIC QUERY CONSTRUCTION DESIGN

Our previous experience with graphical querying concerned querying provenance data \([1]\). QUaTRO2 was driven by significantly different goals and does not reuse any algorithms, design or code from the previous tool. Among main design goals in creating QUaTRO2 were (1) **end-user orientation**: the tool should be designed for domain experts, without IT-related experience; (2) **domain-independence**: the tool should work with any RDF OWL-described data set. (3) **high query expressiveness**: user-orientation should not significantly limit the query expressiveness.

A sample query based on the domain model of the UniProt protein database is shown in Fig. 1. The goal of the query is to return all genes encoding proteins which at the same time cause human diseases, along with description of these diseases. The query is basically a graph of domain concepts (ontology classes) connected by relationships: **Protein encoded by Gene which exists in organism belonging to Taxon ‘Human’**; in addition, the protein has annotation about a related disease. The result of the query can be any subset of attributes of all classes involved in the query. In this particular case, the query chooses **uri** from class **Protein** (the URI of the RDF resource representing the protein), **core#prefLabel** from class **Gene** (preferred name of the Gene), and **comment** from class **DiseaseAnnotation** (description of the disease).

III. QUaTRO2 GUI AND EXAMPLE QUERIES

In view of our design goals, the usability of the Graphical User Interface of the query construction tool is as much important as the underlying algorithms and design. We have validated both these aspects on the case of UniProt, a large-scale protein RDF database. Fig. 2 shows a simple query constructed in QUaTRO2 GUI returning organisms which are human parasites. Fig. 3 presents a more complex query which returns citations about proteins involved in apoptosis. Both figures demonstrate some features of the QUaTRO2 interface, such as selecting classes and attributes, forming the query graph, and displaying query results.

IV. CONCLUSION

Graphical construction of queries over RDF scientific data sets, using ontologies as a query language, can provide a powerful solution enabling scientists to formulate advanced queries without knowing an RDF query language. Future work involves quantitative and qualitative study of QUaTRO2 regarding query expressiveness and performance, and extending of graphical query capabilities.

The QUaTRO2 service for the UniProt database is available online at: [http://149.156.9.71:8080/quatro/](http://149.156.9.71:8080/quatro/).

ACKNOWLEDGMENT

This work is supported by the EU FP7 Project UrbanFlood, grant N 248767.

---

1 Linked Data Sets, [http://www.w3.org/wiki/DataSetRDFDumps](http://www.w3.org/wiki/DataSetRDFDumps).
Fig. 1. Example of a complex semantic query based on UniProt ontology, and its internal graph representation.

Fig. 2. QUaTRO2 graphical interface. Query: human parasites – Taxon (organism) whose host is another Taxon commonly named ‘Human’. Example GUI feature: selecting a class.

Fig. 3. Query: citations about proteins involved in apoptosis – Proteins classified with a Concept ‘apoptosis’ that have any associated Journal citations. GUI features: (i) Forking query graph (and operator). (ii) Selecting class attributes. (iii) Displaying query results.

REFERENCES


