BIOINFORMATICS APPLICATIONS NOTE

Vol. 28 no. 3 2012, pages 448–450
doi:10.1093/bioinformatics/btr662

Databases and ontologies

Advance Access publication November 29, 2011

An infrastructure for ontology-based information systems in biomedicine: RICORDO case study

Sarala M. Wimalaratne[1,*], Pierre Grenon[1], Robert Hoehndorf[2], Georgios V. Gkoutos[2] and Bernard de Bono[1,3]

[1]European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge CB10 1SD, [2]Department of Genetics, University of Cambridge, Cambridge CB2 3EH, UK and [3]Auckland Bioengineering Institute, University of Auckland, Symonds Street, Auckland 1010, New Zealand

Associate Editor: Martin Bishop

ABSTRACT

Summary: The article presents an infrastructure for supporting the semantic interoperability of biomedical resources based on the management (storing and inference-based querying) of their ontology-based annotations. This infrastructure consists of: (i) a repository to store and query ontology-based annotations; (ii) a knowledge base server with an inference engine to support the storage of and reasoning over ontologies used in the annotation of resources; (iii) a set of applications and services allowing interaction with the integrated repository and knowledge base. The infrastructure is being prototyped and developed and evaluated by the RICORDO project in support of the knowledge management of biomedical resources, including physiology and pharmacology models and associated clinical data.

Availability and implementation: The RICORDO toolkit and its source code are freely available from http://ricordo.eu/relevant-resources.

Contact: sarala@ebi.ac.uk

Received on September 28, 2011; revised on November 23, 2011; accepted on November 24, 2011

1 INTRODUCTION

Improvement in computer technology makes it possible to store large volumes of biomedical resources (e.g. mathematical models of physiological processes and related data). The biomedical community is becoming increasingly aware of the importance of annotating this data in order to enable querying and retrieval. As a result, communities are engaged in working together to create annotation standards [e.g. MIRIAM (Le Novère *et al*., 2005)] and biomedical ontologies [e.g. OBO (Smith *et al*., 2007)] to provide a consistent method of sharing heterogeneous resources. These initiatives have improved the prospect of semantic interoperability of resource annotation based on biology. Nevertheless, reaching this interoperability goal remains a challenge.

While simple retrieval based on direct matching of terms used in annotation is certainly straightforward, search using complex descriptions of biological entities denoted by those terms involves possibly demanding reasoning over ontologies. Reasoning is the process by which statements are automatically inferred based on a set of axioms. Automated reasoning enables flexible retrieval of stated and inferred knowledge, as well as consistency checking. However, with an increasing number of classes and relations in an ontology, such a task can become increasingly more complex and requires extensive computational power. The difficulty is further compounded when integrating knowledge across multiple communities (de Bono *et al*., 2011).

In this article, we propose an infrastructure for real-time reasoning over very large ontologies to express complex ontology concepts and use these concepts to retrieve relevant resource metadata. Our use case, the RICORDO project (de Bono *et al*., 2011), focuses on biomedical resources, and related ontologies, relevant to a number of communities including the physiology modeling community (VPH; http://www.vph-noe.eu/), the pharmacology modeling community (DDMORE; http://www.ddmore.eu) and the medical education community (e.g. http://www.meducator.net). These communities mainly deal with large collections of anatomical, physiological and pathological resources including computational models, such as models encoded in CellML (Lloyd *et al*., 2004) and SBML (Hucka *et al*., 2004), and clinical databases.

2 IMPLEMENTATION

The central module in the infrastructure is a store of ontology-based annotations of resources that are machine processable. The infrastructure comprises components for managing this central module. This management includes (i) store maintenance and (ii) querying that is achieved through intermediate reasoning over ontologies used in annotating resources (Fig. 1).

The centralized store consists of metadata statements that link identified resources and their components to named terms in biomedical ontologies. The Resource Description Framework (RDF; http://www.w3.org/RDF/) is used to record these statements. Storing metadata in RDF also supports complex querying via SPARQL (http://www.w3.org/TR/rdf-sparql-query/), a query language for RDF. A Virtuoso (Erling and Mikhailov, 2007) server is used to store metadata in RDF while the Virtuoso API, coupled with Jena API (http://jena.sourceforge.net/), is used to handle RDF triples and SPARQL queries.

The OWLlink server (http://owllink-owlapi.sourceforge.net/) stores OWL-EL versions of ontologies (http://www.w3.org/TR/owl2-profiles/). OWL-EL is used as it allows automated reasoning over large ontologies (Hoehndorf *et al*., 2011). In addition, OWLlink supports OWL reasoning using the Pellet reasoner (http://clarkparsia.com/pellet/). Once the ontologies are loaded into

*To whom correspondence should be addressed.

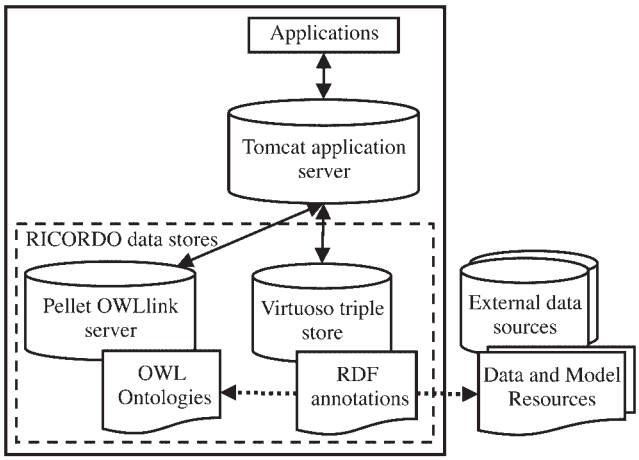

Fig. 1. The current implementation of the infrastructure (enclosed by a black margin). The applications are deployed in a tomcat application server. The applications interact with (i) the OWL knowledge base which is deployed in a Pellet OWLlink server, and (ii) the virtuoso RDF repository. The application does not require access to original data and model resources, but has to point to a relevant set of IDs, thus resources are depicted outside the box.

a knowledge base created by the Pellet OWLlink server, it is possible to query over the ontologies using OWLlink API requests and responses.

The infrastructure includes a web application for user interaction with resources, as well as web services that provide programmatic access to the OWL reasoner server. On the client side, the web application is implemented in Google Web Toolkit (http://code.google.com/webtoolkit/) and hosted on a Tomcat (http://tomcat.apache.org/) server, with query services running on the server side to interact with the resources. REST (Fielding, 2000)/JSON (http://www.json.org/) based web services provide external access to the OWL reasoner server, thus enabling the integration of reasoning over large ontologies into third-party applications.

The web application supports a number of functionalities, including: (i) the annotation of resources, using a URI for the resource to be annotated and a URI for an ontology term to be associated with that resource, to create a related RDF statement that is stored; (ii) the definition of complex terms based on class expressions in OWL-EL and using terms and relations from biomedical ontologies. We call such defined terms 'composites' (de Bono *et al.*, 2011). Defining a composite assigns a unique identifier to this term for the storage and subsequent re-use of its definition; (iii) the querying of resources, achieved in two steps. The first generates a list of ontology terms by querying the OWL knowledge base. The RDF store is then queried for resources annotated with terms from this list.

In our current configuration, the knowledge base includes biomedical ontologies relevant to data and model resources. This set of core ontologies (CORDO) includes FMA (Rosse and Mejino, 2007), PATO (Gkoutos *et al.*, 2005), GO (Ashburner *et al.*, 2000), Cell Type (Bard *et al.*, 2005), ChEBI (De Matos *et al.*, 2010), HPO and its class definitions (Gkoutos *et al.*, 2009; Robinson and Mundlos, 2010), as well as composite terms developed within RICORDO. The RDF store maps ontology terms according to the MIRIAM URN scheme (Le Novère *et al.*, 2005), and MIRIAM web services (Laibe and Le Novère, 2007) are applied to resolve MIRIAM URNs.

Table 1. Query times in order of which they were executed

Query	Query time (ms)
FMA_7088	528
part-of some FMA_7088	34 468
inheres-in some (part-of some FMA_7088)	2035
PATO_0000918 and inheres-in some (part-of some FMA_7088)	101

An example of a search is to query the RDF store for resources related to volumes of some part of the heart. Finding the relevant ontology terms is achieved by querying the knowledge base, using the Manchester OWL Syntax (Horridge *et al.*, 2006). In this use case, *PATO_0000918 and inheres-in some (part_of some FMA_7088)* represents the class of volumes of a part of the heart. This class definition is used to generate a list of subclasses and equivalent classes from the knowledge base via the OWLlink interface. This list is then passed to the RDF query engine.

The application implements a set of templates that allow the formulation of queries. Each template has a particular form to specify the query terms and relations. An example of a query template is *<relation> some <term>* where *relation* and *term* refer to an ontology property and a class, respectively. Selecting a particular template generates fields to capture the terms and relations entering into the description of an OWL class. Auto completion of ontology terms and relations in these fields is supported by the Ontology Lookup Service (Côté *et al.*, 2008).

Performance of the overall prototype is largely influenced by performance of the OWL reasoning module, thus an initial evaluation of reasoning over the CORDO ontologies was performed. This was carried out on a server with a dual CPU Intel Xeon 2.4 GHz with 24 GB memory. The start up of the KB involves loading the ontologies (90 s) and classification of the ontologies (9.4 min). Query times depend on the complexity of the queries as well as caching. Thus, the results of the query 'PATO_0000918 and inheres-in some (part-of some FMA_7088)' can be retrieved in 0.1 s (Table 1).

3 DISCUSSION

The prototypical implementation of the infrastructure in the RICORDO context allows searching data and model resources using ontologies. The web applications allow the retrieval and annotation of resources with both terms from reference ontologies and composites of those terms. Web services provide programmatic access to the ontology resources. Future work will include distribution of the RDF store and OWL reasoning.

The metadata is stored independently of the annotated resources. This independent storage supports efficient metadata management and preserves both the structural integrity and confidentiality of these resources. Furthermore, the separation of data and ontologies, as well as the subsequent separation of querying of metadata and reasoning over ontologies, allows the storage of large amounts of metadata without affecting the performance of ontological reasoning. The architecture can be employed to support ontology-based metadata management in any area of application. The solution

is general and reusable by multiple biomedical communities in integrating and sharing metadata.

Funding: European Commission, grant agreement number (248502) (RICORDO) and 223920 (VPH NoE) within the 7th Framework Programme.

Conflict of Interest: none declared.

REFERENCES

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Bard,J. *et al.* (2005) An ontology for cell types. *Genome Biol.*, **6**, R21.

de Bono,B. *et al.* (2011) The RICORDO approach to semantic interoperability for biomedical data and models: strategy, standards and solutions. *BMC Res. Notes*, **4**, 313.

Côté,R.G. *et al.* (2008) The Ontology Lookup Service: more data and better tools for controlled vocabulary queries. *Nucleic Acids Res.*, **36**, W372–W376.

Erling,O. and Mikhailov,I. (2007) RDF support in the virtuoso DBMS. *Network. Know. Network. Media*, **221**, 59–68.

Fielding,R.T. (2000) Architectural Styles and the Design of Network-based Software Architectures. PhD Thesis. Avaliable at http://www.ics.uci .edu/~fielding/pubs/dissertation/top.htm.

Gkoutos,G.V. *et al.* (2005) Using ontologies to describe mouse phenotypes. *Genome Biol.*, **6**, R8.

Gkoutos,G.V. *et al.* (2009) Entity/quality-based logical definitions for the human skeletal phenome using PATO. *Conf. Proc. IEEE Eng. Med. Biol. Soc.*, **2009**, 7069–7072.

Hoehndorf,R. *et al.* (2011) A common layer of interoperability for biomedical ontologies based on OWL EL. *Bioinformatics*, **27**, 1001–8.

Horridge,M. *et al.* (2006) The Manchester OWL Syntax. *Syntax*, **216**, 10–11.

Hucka,M. *et al.* (2004) Evolving a lingua franca and associated software infrastructure for computational systems biology: the Systems Biology Markup Language (SBML) project. *Syst. Biol.*, **1**, 41–53.

Laibe,C. and Le Novère,N. (2007) MIRIAM resources: tools to generate and resolve robust cross-references in Systems Biology. *BMC Syst. Biol.*, **1**, 58.

Lloyd,C.M. *et al.* (2004) CellML: its future, present and past. *Progr. Biophys. Mol. Biol.*, **85**, 433–450.

De Matos,P. *et al.* (2010) ChEBI: a chemistry ontology and database. *J. Cheminformat.*, **2**, P6.

Le Novère,N. *et al.* (2005) Minimum information requested in the annotation of biochemical models (MIRIAM). *Nat. Biotechnol.*, **23**, 1509–1515.

Robinson,P. N. and Mundlos,S. (2010) The human phenotype ontology. *Clin. Genet.*, **77**, 525–534.

Rosse,C. and Mejino,J.L.V. (2007) The foundational model of anatomy ontology. *Esophagus*, **2007**, 59–117.

Smith,B. *et al.* (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.*, **25**, 1251–1255.