BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–2

Brain: Biomedical Knowledge Manipulation

Samuel Croset [1,*], Robert Hoehndorf [2], John Overington [1] and Dietrich Rebholz-Schuhmann [1]

[1]EMBL European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD UK
[2]Department of Genetics, University of Cambridge, Downing Street, Cambridge, CB2 3EH, UK

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Summary: Brain is a software library facilitating the creation and manipulation of ontologies and knowledge-bases represented in the Web Ontology Language (OWL).

Availability and Implementation: The Java source code and the library are freely available at https://github.com/loopasam/Brain and on the Maven Central repository (GroupId: uk.ac.ebi.brain). The documentation is available at https://github.com/loopasam/Brain/wiki.

Contact: croset@ebi.ac.uk

Supplementary information: Supplementary data available at the journal's web site.

1 MOTIVATION

Relational databases currently hold most of the available structured biomedical information. The content of these repositories is often extracted from scientific literature by manual curation with the help of text-mining tools. The transformation from raw text into structured data is a key step, since the curated information can then be classified, managed and queried more easily. Databases facilitate the re-use of previous work in a computer-friendly manner and support large-scale mining of biomedical knowledge. In order to leverage further the existing information, the current trend is towards coordinated data integration and database interoperability, with large projects such as ELIXIR (Crosswell and Thornton, 2012) leading the way. The underlying driver for these efforts is that more complex biomedical challenges, such as finding new treatments for diseases, could be better addressed by combining and integrating the content of a range of specialized repositories. Knowledge-bases, a concept from computer science (see Krötzsch *et al.*, 2012 for an introduction), could be a solution to improve the interoperability and the value of the data; yet at the time of writing, no framework is available for the biomedical domain. Brain - the software library presented in this manuscript addresses this matter and provides a simplified interface dedicated to handle and query biomedical knowledge-bases.

*to whom correspondence should be addressed

2 KNOWLEDGE-BASES

Traditional relational databases are often an obstacle in realizing large-scale data integration, mostly because of their lack of support for interoperability: The schema structuring the data is usually repository-specific which limits the combination of the native content with external data in an efficient, flexible and meaningful fashion. In order to address this issue, a series of community standards forming the *semantic web* have been developed. These standards provide the means to better model the domain knowledge and to facilitate data exchange and interoperability at a larger scale. One of these standards, the Resources Description Framework (RDF) enables the exposition of the underlying structure as part of the data themselves (Manola and Miller, 2004). This representation relies on building blocks composed as *subject - relation - object* triples which serves to describe the data as well as their types. More complicated data structures, such as sub-class or transitive relationships can be further expressed via another standard, the Web Ontology Language (OWL). OWL derives from description logic and has been designed to capture the knowledge of a domain of interest in the form of a structured vocabulary (W3C OWL Working Group, 2009). This feature makes it particularly interesting from the perspective of the life sciences, since a number of ontologies and classification schemes have been developed from the origin of the discipline. OWL is often expressed in combination with RDF in order to reveal the underlying schema of the data, but can also be used as such, as an implementation of description logic for computers. Knowledge-bases and ontologies can therefore be built without being necessarily expressed as RDF triples while still preserving all the advantages in regards to data integration and interoperability. Brain aims at facilitating the construction and manipulation of such knowledge-bases or ontologies, rather than being oriented towards the consumption of RDF data. We will present first the biomedical motivation for the particular subset of OWL supported by Brain - so called OWL 2EL, then we will discuss the main features implemented by the library.

3 SCALABILITY AND COMPLEXITY

Knowledge representation in the biomedical domain differs from other disciplines, because of the diversity and abundance of information that can potentially be interconnected. Brain appreciates this aspect and focuses on a particular profile of OWL,

called EL which consists of a subset of the constructs available in the original language (Motik *et al.*, 2009). This profile is designed to be *tractable*, meaning that the constructs available have a polynomial complexity, and are therefore easier to compute than the full version of OWL. These constructs are called *axioms* and are the fundamental notion behind OWL knowledge-bases. Axioms assert the facts and relations present in the knowledge-base and are computable by a program named *reasoner*. Based on the logical structure of the axioms, the reasoner is capable of deriving new facts from the asserted ones as well as retrieving some implicit information, enabling powerful query mechanisms surpassing whose available in the Structured Query Language (SQL). Brain primarily supports the OWL 2EL profile for its computational properties, and is suitable for real-life biomedical applications, where millions of axioms could be potentially extracted from complex repositories such as ChEMBL (Gaulton *et al.*, 2012). Moreover, the EL profile is expressive enough to cover a good portion of biomedical knowledge: Most of biological ontologies such as the Gene Ontology (GO - Ashburner *et al.*, 2000) or the Chemical Entities of Biological Interest (ChEBI - De Matos *et al.*, 2010) ontology are already included in this profile, and any relational model can be easily converted and represented using OWL 2EL - opening doors to large-scale meaningful data integration. Brain builds on the top of Elk, a fast reasoner dedicated to EL ontologies (Kazakov *et al.*, 2011). Elk shows good performances at handling large datasets and offers the possibility to run some reasoning tasks in parallel; therefore clusters or multicore architecture can scale the speed of reasoning as more data are added to the knowledge-base. Brain wraps and simplifies the interaction with Elk while still leaving the possibility to fine tune the configuration for advanced users.

4 PROGRAMMATIC FEATURES

One purpose of Brain is to ease the manipulation of knowledge-bases as well as increasing the rate at which they can be developed and validated. In practice, the implementation of OWL ontologies and knowledge-bases can be done in either a programmatic way via the OWL-API (Horridge and Bechhofer, 2011) or with the help of a visual tool such as Protege (Stanford Center for Biomedical Informatics Research, 2012). The graphical interface of Protege allows the user to focus on the generation of axioms rather than on details of the underlying programmatic implementation: OWL expressions can be entered with the user-friendly Manchester syntax (Horridge *et al.*, 2006) and the name of the classes are displayed in a convenient way. More complex applications, requiring a deeper control over the ontology, can take advantage the OWL-API but require proficiency in Java and can be daunting for casual users. OWL-API deals with all aspects of ontology generation for semantic web purposes, nonetheless the level of granularity can be cumbersome for some biomedical applications. Brain aims at filling the gap between the OWL-API and graphical interfaces: It is implemented as a facade, providing a series of convenience methods for common use-cases encountered in the biomedical domain and leveraging the access to the OWL-API. The full list of currently supported constructs is available in the supplementary material. Brain relies on the intuitive and explicit Manchester syntax to formulate OWL class expressions, just as in the Protege editor. The interaction with the OWL-API centers around strings handling rather than Java objects, making Brain suitable to parse and answer requests in the context of a web service or an OWL end-point for instance. Using strings as input also speeds the production and flexibility of the code written to move to an OWL representation from a relational or flat-file database for example. The library supports the loading and referencing of external ontologies in order to integrate and reason over data from different sources. An important feature of Brain is the query mechanism, which is similar to that implemented in Protege. Powerful questions can be formulated over the knowledge-base using the Manchester syntax, abstracting away complex interaction with the Java object provided by the OWL-API. An example of question answering over the GO using Brain is compared against a traditional SQL query in the supplementary material.

5 CONCLUSION

Brain is an Open-source Java library designed to build and query biomedical knowledge-bases or OWL ontologies. The library is centered on the EL profile, and designed to be suitable and scalable for biomedical knowledge representation. The convenience methods provided by Brain should simplify the development of biomedical knowledge-bases and allow developers to increase their productivity while effectively dealing with data integration challenges.

ACKNOWLEDGEMENT

Funding: This work was supported by funding from the EMBL member states. Samuel Croset is a member of Darwin College, University of Cambridge.

REFERENCES

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics*, **25**(1), 25–29.

Crosswell, L. C. and Thornton, J. M. (2012). ELIXIR: a distributed infrastructure for European biological data. *Trends Biotechnol.*, **30**, 241—-242.

De Matos, P., Alcántara, R., Dekker, A., Ennis, M., Hastings, J., Haug, K., Spiteri, I., Turner, S., and Steinbeck, C. (2010). Chemical Entities of Biological Interest: an update. *Nucleic Acids Research*, **38**(Database issue), D249–D254.

Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., and Overington, J. P. (2012). ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, **40**(Database issue), D1100–7.

Horridge and Bechhofer (2011). The OWL API: A Java API for OWL Ontologies. *Semantic Web Journal*, pages 11–21.

Horridge, M., Drummond, N., Goodwin, J., Rector, A., Stevens, R., and Wang, H. H. (2006). The Manchester OWL Syntax. *Syntax*, **216**, 10–11.

Kazakov, Markus Krötzsch, and František Simančík (2011). Concurrent Classification of EL Ontologies. *Proceedings of the 10th International Semantic Web Conference (ISWC'11)*, **7032**.

Krötzsch, M., Siman, F., and Horrocks, I. (2012). A Description Logic Primer. *Language*, **abs/1201.4**(January), 1–16.

Manola and Miller (2004). RDF Primer.

Motik, B., Grau, B. C., Horrocks, I., Wu, Z., Fokoue, A., and Lutz, C. (2009). OWL 2 Web Ontology Language Profiles. *Language*, **2009**(October), 1–53.

Stanford Center for Biomedical Informatics Research (2012). Protégé Project.

W3C OWL Working Group (2009). OWL 2 Web Ontology Language Document Overview.