

Brain: Biomedical Knowledge Manipulation

Samuel Croset^{1,*}, JPO² and DRS²

¹Department of XXXXXXXX, Address XXXX etc.

²Department of XXXXXXXX, Address XXXX etc.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXXX

ABSTRACT

Summary: This section should summarize the purpose/novel features of the program in one or two sentences.

Availability and Implementation: This section should state software availability if the paper focuses mainly on software development or on the implementation of an algorithm. Examples are: 'Freely available on the web at <http://www.example.org>. Website implemented in Perl, MySQL and Apache, with all major browsers supported'; or 'Source code and binaries freely available for download at URL, implemented in C++ and supported on linux and MS Windows'. The complete address (URL) should be given. If the manuscript describes new software tools or the implementation of novel algorithms the software must be freely available to non-commercial users. Authors must also ensure that the software is available for a full TWO YEARS following publication. The editors of Bioinformatics encourage authors to make their source code available and, if possible, to provide access through an open source license see www.opensource.org for examples.

Contact: croset@ebi.ac.uk

Supplementary information: Links to additional figures/data available on a web site, or reference to online-only Supplementary data available at the journal's web site.

Relational databases 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 most important, as 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 the biomedical knowledge to scale-up. In order to leverage further more the existing information, the current trend is at data integration and interoperability, with large projects such as ELIXIR leading the way. The underlying idea assumes that increasingly complex biomedical challenges such as finding new treatments for diseases could be addressed by combining the content of independent repositories via the Internet. Traditional relational databases are however an obstacle to realize this vision, mostly because of their lack of support for interoperability: The schema structuring the data is indeed very repository-specific which limits the combination of the native content with external data in an efficient and meaningful fashion. In order to address this issue, a series a standard forming the semantic web have been

developed. One of them, the Resources Description Framework (RDF) enables the exposition of the underlying structure as part of the data themselves. This representation relies on triples as building block, composed as *subject - relation - object* and where the types of the data are identified with a special relation called *rdf:type*. More complicated data structures, such as sub-classes or transitive relations can be further expressed via another standard, the Web Ontology Language (OWL). OWL derives from Description logic and is used to capture the knowledge of a domain of interest in the form of a structured vocabulary. This feature makes it particularly interesting from the point of view of life science, as a lot of ontologies and classification have been developed since the origin of the discipline. OWL is often expressed in combination with RDF data in order to reveal the underlying schema, but it can also be used as such as an implemetation of a part of description logic. 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, the library presented in this manuscript aims at facilitating the contruction and manipulation of such knowledge bases (referred in the manuscript also as 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. The will be discussed the main features of Brain in regards to this profile.

OWL 2EL is a profile of OWL as it features only a subset of the constructs available in the original language. This profile is designed to be *tractable*, meaning that the constructs available have a polynomial complexity which is easier to compute than the full version of OWL. These constructs are called *axioms* and are the fundamental block behind OWL knowledge bases. Axioms assert the facts and relations present in the knowledge base and are understood by a computer program named *reasoner*. Based on the logical structure of the axioms, the reasoner is indeed capable of deriving new facts from the asserted ones as well as retrieving some implicit information, enabling powerful query mecanisms surpassing the Structured Query Language (SQL) expressivity. A reasoner can also check the consistency of an ontology and report back the axioms violating a particular profile. Because of these computational properties, OWL 2EL is suitable for real-life bomedical applications, where millions of axioms are potentially present. Moreover, the profile is expressive enough for the biomedical knowledge: Most of the ontologies such as the Gene Ontology or Chebi are already respecting this profile and any relational model can be easily converted and represented

*to whom correspondence should be addressed

using OWL 2EL. OWL knowledge representation also separates *individuals* known as Assertion Box (ABox) from *classes* and concepts, the Terminological Box (TBox). In life science, most of the entries of databases are to be considered as classes rather than instances, as they describe the generic version of a molecular entity, such as a protein which has a lot of materialization in practice, namely the actual proteins. This representation differs a lot from the one used in the relational model or with RDF, where database entries are considered as instance records rather than as actual biological objects. While working this OWL this distinction is important to make, as it influences the inferences made by the reasoner. Fortunately OWL 2EL focuses on this aspect and enables the handling of knowledge bases with a large number of classes.

- scalability - parallelism - interoperability

In practice, the implementation of OWL ontologies and knowledge bases can be done either in a programmatic way via the OWL-API or with the help of a graphical tool such as Protege.

how does one implement a knowledge base? need to use the OWL-API and a specific reasoner or Protege ok, what's the problem with that? extremely comprehensive solution for the API but heavy. Protege is a good GUI but not suitable for programmatic access. Need for a lightweight solution specific for OWL 2 EL, in order to build web service and OWL end-points. ok, what is the solution? Brain, a facade for the owl api and the ELk reasoner. what are the main features of brain? brain deals with one ontology at a time only focusing on the biology, less the computer science. brain deals only with unique names non redundant names for entites string based interaction to build classes and class expression using the intuitive manchester syntax error-handling driven convenience method for import/export of stuff, management of prefixes list of the construct and implementation as table (copy pasta from website). scalable

because of Elk. creration and iomport of external ontologies (learn) consistency checking

Why is brain interesting for the community tool to build KB scalable solution in order to preapre the next stage of data integration with KB reasoner driven and semantic web.

ACKNOWLEDGEMENT

These should be included at the end of the text and not in footnotes. Please ensure you acknowledge all sources of funding, see funding section below. Details of all funding sources for the work in question should be given in a separate section entitled 'Funding'. This should appear before the 'Acknowledgements' section.

Funding: The following rules should be followed: The sentence should begin: This work was supported by

The full official funding agency name should be given, i.e. National Institutes of Health, not NIH (full RIN-approved list of UK funding agencies) Grant numbers should be given in brackets as follows: grant number xxxx Multiple grant numbers should be separated by a comma as follows: grant numbers xxxx, yyyy Agencies should be separated by a semi-colon (plus and before the last funding agency) Where individuals need to be specified for certain sources of funding the following text should be added after the relevant agency or grant number to author initials. Oxford Journals will deposit all NIH-funded articles in PubMed Central. See Depositing articles in repositories information for authors for details. Authors must ensure that manuscripts are clearly indicated as NIH-funded using the guidelines above.

REFERENCES