BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–2

Brain: Biomedical Knowledge Manipulation

Samuel Croset [1,*], JPO [2] and DRS [2]

[1]Department of XXXXXXX, Address XXXX etc.
[2]Department of XXXXXXXX, Address XXXX etc.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

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, large projects such as ELIXIR leading the way. The underlying idea assumes that increasingly complex biomedical questions could be answered by combining the content of discrete repositories, with application such as catalysing the developement of new treatments. Tradiotional 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 repository-specific which impairs the combination with external resources. Moreover, the relational representation does not provide native means to abstract the formulation of complex queries.

*to whom correspondence should be addressed

Life-science repositories are sometimes seen as isolated island of information focusing traditionally on one thematic or biological concept. Interoperability between various resources is limited to hyperlinks or references in practice, impairing large-scale data integration and analysis. In order to benefit

Life science databases have particular features making them unique in regards to other domains. First, they essentially serve as read-only resources where the researchers come to browse or download the content from. The majority of the time, no transactions or updates are hapenning as opposed to services such as Amazon or Ebay. Second, biomedical databases are highly interconnected via the Internet. It is really easy for a user to move from one repository to an other via the hyperlinks present on web interfaces. This distribution of workload and information is likely to increase in the future, driven by large projects such as ELIXIR for instance. Third, biomedical repositories have to be able to answer increasingly complex questions which are traditionally formulated with the Structured Query Language (SQL). As databases are well interlinked, one would expect to be able to gather quickly results coming from different sources. In practice this is however not the case because of the discrepencies in structures and schemas, the technical difficulties to run federated queries and simply because relational databases were not design to accomplish this task in the first place. Finally, the biological information is not necessarily best represented as relational tables. In particular hierarchical structures such as taxonomies or ontologies such as the Gene Ontology are known to be challenging to represent in relational databases.

Because of these particular features, the biomedical domain is particularly attractive for semantic web technologies. First coined by Berners-Lee, the semantic web is a set of standards promoting the interoperability and integration of information at the scope of the World Wide Web. Data are represented as triples using the Resource Description Framework (RDF) which explicitly expose the relations and types between entities, in contrast to relational databases where it is part of the schema. RDF triples are kept in a triplestore which can be queried using a language very similar to SQL called SPARQL. Triplestores are usually much slower than relational databases and provide very little added value to the original content alone in practice. We argue here that the main benefit from semantic web in the scope of life science comes from the Web Ontology Language (OWL). OWL grew from mathematics and description logic as a mean to formally represent some knowledge. It has then been adapted for the semantic web in order to express the ontologies supposed to represent the possible types of RDF triples. Despite being best known for its role in semantic web, OWL could also

be used by itself to build knowledgebases, namely a collection of OWL statements. Knowledgebases provide some clear advantages over relational databases and triplestores.

expressivity of OWL –¿ problems –¿ OWL 2 EL to the rescue

accurate representation as class relations can be combined taxonomies represented reasoner table of definition for knowledge bases Abox tbox leverage of resources new definitions from simple concepts

The availability of structured data enables the formulation of queries with the Structured Query Language (SQL) in order to retrieve the desired information over relational databases. For instance, it is possible to quickly find the cellular location of a protein just by looking at its entry on Uniprot and without having to read several scientific articles on the topic. The query language used and the structure of the database are the factors limiting the complexity of the questions that can be formulated over the biomedical information. For example, relational databases are powerful to store values about instance data such as the sequence or the name of a gene product, but are daunting to use to represent hierarchical structures (ref) such as the Gene Ontology. Therefore queries resolving around hierarchies are going to be complicated to formulate using SQL and will not necessarly deliver a complete answer. For instance, simple queries such as retrieving the list of all the mammals from the NCBI taxonomy involves retrieving first the direct descendants of the mammals and for each of them recursively retrieving their descendants. This type of construct is not trivial to implement and is not yet supported natively by open-source solution like MySQL or PostgreSQL. Biomedical databases are not designed for interoperability. Combining the information present in two separate repositories can indeed be a painful task as the underlying schemas supporting the databases often differ.

The future of biomedical databases (ELIXIR) is leverage of their content via interoperability with other resources Knowledge bases Complicated queries

knowledge versus information

Ontologies and databases in biology are somehow used differently than in other fields.

- The future of biomedical databases (ELIXIR) is leverage of their content via interoperability with other resources. OWL provide the means to do this, databases don't. - Features: Creation and storage of OWL knowledge-bases. Import of external knowledge-bases/ontologies. Simplification of interaction in regards to OWL-API. Fast classification time (Elk reasoner). Support complex queries via inference. Fast and suitable to be used in production. - Evaluation: MySQL build of Go is compared versus OWL build of GO (identical content - different representation). A series of biomedical questions will be answered in SQL and OWL respectively on the MySQL and on the OWL ontology. Comparison of performances –¿ Brain is fast and scalable (thread friendly implementation).

Problems: - Need to be able to formulate complex queries –¿ hidden/implicit knowledge - inference. - Need for interoperability among resources –¿ semantic web technologies. - Need to be able to leverage resources –¿ benefit from integration - Relational databases are struggling with biomedical hierarchical representations. Instance versus classes - Databases in biology are used as read-only resources - Databases don't support inference natively but it could be overcome by a pre-processing step (similar to a classification for a knowledge base).

Solution: - Semantic Web technologies could be slow –¿ OWL 2 EL. - Advantages of triple -stores –¿ none, they still are relational flat data. - Simplying interaction on OWL-API, tracktable problem. Focusing on biology, not computer science. - More and more ontologies are represented in OWL.

OWL 2 EL building blocks: - Individual, properties, classes - Axioms - Reasoner

- relation with description logic - tracktability argument + definition - OWL –¿ query and representation language - table comparison query + times

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