Filipe Santana

De: ISMB 2016 <ismb2016@easychair.org>
Enviado em: sexta-feira, 19 de fevereiro de 2016 18:43
Para: Filipe Santana Da Silva
Assunto: ISMB 2016 notification for paper 143

Categorias: filipe.santana.silva@gmail.com

Dear Authors,

Thank you for your submission to the ISMB 2016 Proceedings Track. We regret that we are unable to accept your paper for publication at ISMB 2016. There were many strong submissions this year and only a small fraction could be accepted.

The organizers received 187 submissions for an anticipated total of about 40 presentation slots. Each paper was reviewed by several members of the appropriate program committee, overseen by the relevant Area Chairs and Theme Chairs. Papers were individually reviewed and subsequently open for discussion by their reviewers to encourage a robust debate and resolution of discrepancies between reviewers. Recommendations put forward by the Area Chairs were reviewed and discussed by the Theme Chairs, with moderation by the Conference Chairs, resulting in the final selection of conditionally accepted papers. Attached you will find the reviews of your manuscript.

Although your paper was not accepted for conference presentation, your work may be appropriate for poster presentation in one of the poster sessions. To participate, you must submit your abstract as a poster no later than Thursday, March 10, 2016 to this site:
https://www.iscb.org/ismb2016-submission/ismb2016-call-for-posters

We hope you will be able to participate at ISMB 2016 in Orlando, United States, July 8-12.

Conference registration opens March 28 and registration details are available at
http://www.iscb.org/ismbeccb2015-registration.

Sincerely,

Pierre Baldi and Teresa Przytycka
Conference Chairs

---------- REVIEW 1 ----------
PAPER: 143
TITLE: Re-interpreting biomedical database content as ontologies for enhanced querying
AUTHORS: Filipe Santana Da Silva, Fred Freitas and Stefan Schulz

---------- REVIEW ----------
The manuscript "Re-interpreting biomedical database content as ontologies for enhanced querying" describes an approach to represent the content of some biological databases as an OWL ontologies. The authors perform an ontological analysis of the database content based on the BioTop Ontology, and implement a prototype OWL ontology for which they report (time) performance results and demonstrate that it is possible to answer a set of previously defined competency questions.

Comments:

1. The link to the results (http://www.cin.ufpe.br/integrativo, in the
abstract) does not work; the link later in the paper also does not work (http://www.cin.ufpe.br/~integrativo). The correct link seems to be http://integrativo.github.io/, but this was hard to find. I base my review on this link.

2. I cannot reproduce answering the competency questions with the ontology provided. For example, for question 6, the class "Ruminantia"
is required; there is no such class in integrative_new.owl.

3. The main issue, in my opinion, with the presented manuscript is the lack of a motivation for the work, and consequently a lack of an evaluation of the developed model: why is this work done, and how does it provide advantages over, for example, an RDF/SPARQL-based representation (which is pursued by most major databases now, including UniProt, TREMBL, ChEMBL, etc.). All the competency questions provided in the manuscript can already be answered using SPARQL queries. Why do we need an OWL-based representation? A consequence of this is that the model is not evaluated, only the performance of reasoning. But the model decisions the authors make stand without justification, and are sometimes rather counterintuitive. For example, the authors merge several database records from multiple databases, and then claim that each entry in their merged file represents a process. Why is this the case? And what would be gained by systematically defining complex processes based on all these patterns, instead of using a combinatorial approach (as currently used, and directly available through SPARQL endpoints) in which these classes are combined at query time? Additionally, the choice of the subset of data the authors focus on does not allow generalizing; question 1, for example, for biological processes related to Hcy, does not actually use "Hcy" in the DL query, but simply asks for all processes that are "included in" the mouse; the choice of the subset is what makes the answer correct, but this cannot be extended.

Finally, parts of this work have already been done previously. In "Logical Gene Ontology Annotations (GOAL): exploring gene ontology annotations with OWL" (2012), proteins (in the mouse) have been integrated with GO in an OWL ontology, and a DL reasoner is used to query the resulting database (and would scale well enough to allow a web interface to be based on it). Since the model (based on BioTop and different ontological analyses) is not evaluated (or motivated) in the manuscript, it is difficult to judge how this work differs except that it also incorporates a few more databases (but then, only a small subset thereof).

In summary, I think this is excellent and exciting work that should be pursued further, but it may be a bit premature to publish this at ISMB in its current form.

Robert Hoehndorf

---------- REVIEW 2 ----------
PAPER: 143
TITLE: Re-interpreting biomedical database content as ontologies for enhanced querying
AUTHORS: Filipe Santana Da Silva, Fred Freitas and Stefan Schulz

---------- REVIEW ----------
This paper is presenting a new approach to query biomedical databases (e.g. Uniprot), in order to enable a formalized interpretation of the results: database content and ontology are automatically integrated together to form a large set of OWL axioms. As a test, axioms are generated out of a a subset of 46 Uniprot records and 6 queries are applied to these axioms.

The article should be revised by an english expert. The reading is complicate due to numerous mistypes (words missing, spelling mistakes, etc.) and sentences are often extremely long. Some abbreviations are never defined (e.g. BFO, RO, ALC)
- page 2: « Ontology-level content in identified in databases, and axiomatized under an upper level ontology. »
- page 2: « TBoxed »
- page 3: « They explore database content concerning phenotypes, proteins, molecules and biological processes from several organisms, using ontologies and databases described in Section 3). »
- page 3: « and responsible the current structure for the UniProt database »

- page 4: « The ontological content is evaluated by six competency questions (CQs),as introduced above. »
- page 4: « with the classes referred to in the selected database content as signature, »
- page 4: « To avoid biased judgements, this process is assessed by means of DL classification and retrieval of content from the axioms generated. are rendered as DL queries and submitted to the final ontology. »
- page 4: « Giant panda, Bovine, white-tufted marmoset, dog, zebrafish, chicken, human, West Indian ocean coelacanth, African elephant, mouse, European domestic ferret, Nile tilapia, rabbit, chimpanzee, Sumatran orangutan, rat, Tasmanian devil, pig, Japanese pufferfish. Western clawed frog »
- page 5: « as introducing a series of defined »
- page 5: « each biological process class referred to by this DB record. »
- page 5: « The left six columns contain UniProt entries; the two right columns correspond to Ensembl in entried. »
- page 5: « Each Bp referred to by in a DB record may happen within the same cellular structure in several organisms O »
- page 6: « each of the newly defined classed corresponds to at least one fact described in literature »

Abstract:
URL mentioned in the abstract is not available (http://www.cin.ufpe.br/integrativo).

Background:
Ontology grounding (pages 2-3): I would appreciate more explanation of figure 1 or an example. How is the relationships between the referents and their types analyzed? What is use to check consistency?
The CQs on HCY should be mentioned in the method section (evaluation methodology) rather than in the background section.

Methods:
21 species: but only 20 specified
Preparation of the Hcy-related records: how do you proceeded to filter your 212,156 records to 46 records? Automatic process? Manual cleaning?
In the abstract, it is claimed that the resource is fully automatically created, « guided by manual work ». In the methods, several elements seems to indicate that is it mainly manual processes, partially « semi-automated »: Excel spreadsheet, in-depth ontological analysis done by the author, agreement reaching, manually-dissected into code fragments.
The 6 queries formulated by the author: it would be interesting to see the full picture of queries of the databases by real users.

Results:
Table 2: in legend you mention « Genes G » but G does not appear in the table.
« We translate the content from table 1 as an example table 2 for interpretation » => what does it mean?
Section 5.3 is very confusing. Examples would help.

Discussion and conclusion are clear.

---------- REVIEW 3 ----------
PAPER: 143
TITLE: Re-interpreting biomedical database content as ontologies for enhanced querying
AUTHORS: Filipe Santana Da Silva, Fred Freitas and Stefan Schulz

---------- REVIEW ----------
This paper introduces the use of ontologies for structuring biological databases, and explores the performance impact of scaling on classification and querying.

The explanations of the background of ontologies and description logics are concise and clear, and in general I found the work interesting and important for grounding databases in ontologies. I particularly liked the framework of Competency Questions for guiding the testing of the work.

There are a few things that would increase the overall impact of the work and address its novelty:
- discussion of efforts such as Bio2RDF [1] and KaBOB [2] -- how is this effort distinct?
- similarly, the bottom of p8 refers to "many solutions" -- citations, and perhaps some elaborated discussion of more specifics, would be helpful
- where BTL2 is introduced, clarify why BTL2 is preferable to the existing BFO and RO ontologies
- ODBA is mentioned in several places; please explain what this is and expand on the limitations and differences in approach

I also didn't quite follow the scaling strategy in 4.3 -- is it simple duplication? How does that impact the querying/classification? Why is it a challenge? "Tangledness" is mentioned ... What does that mean/impact?

[1] http://link.springer.com/chapter/10.1007/978-3-642-38288-8_14
[2] http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-015-0559-3

---------- REVIEW 4 ----------
PAPER: 143
TITLE: Re-interpreting biomedical database content as ontologies for enhanced querying
AUTHORS: Filipe Santana Da Silva, Fred Freitas and Stefan Schulz

---------- REVIEW ----------
This paper argues for a formal-ontological approach to encoding and reasoning with the content of biological databases and the ontologies they are annotated to. An example is presented.

Pros:
- generally interesting
- reasonably well written
- relatively non-technical (as much as possible at least)
- proof of concept presented

Cons:
- complex method for axiomatizing the databases
- lot of manual work involved
- no discussion of whether the complexity of the axiomatization process is a limitation to wide adoption of this approach by biologists
- the authors fail to demonstrate that the same results could not have been obtained when using a simpler approach (e.g., Linked Data)

Specific comments
- The non trivial results seem to come mostly from exploiting subsumption relations in ontologies in conjunction with the annotation databases. This is interesting, but not really novel. See for example:
Sahoo SS, Zeng K, Bodenreider O, Sheth A. From "glycosyltransferase" to "congenital muscular dystrophy": integrating knowledge from NCBI Entrez Gene and the Gene Ontology. Stud Health Technol Inform. 2007;129(Pt 2):1260-4. PubMed
PMID: 17911917; PubMed Central PMCID: PMC2562001.

Minor comments
- OBDA: ??? (expand)
- "all adenine molecules are nucleotides" -- sounds like a shortcut for "all adenine molecules are *purine bases in* nucleotides", which still sounds dubious as some adenine molecules may not be associated with a pentose or phosphate group, and thus may not form a nucleotide.
- "Tangledness of these experimental ontologies is guaranteed by random assignment of i in the axioms. The resulting ontologies are submitted to a DL reasoner to verify classific" -- please explain for this audience

Grammar and typos
- "TBoxed are constituted" -> TBoxe*s*
- "and other 31 biological processes" -> "and 31 other biological processes"
- "Organisms from different species that include for which the same proteins under the same conditions are described may not include similar processes" -- cannot parse; please rephrase.

---------- REVIEW 5 ----------
PAPER: 143
TITLE: Re-interpreting biomedical database content as ontologies for enhanced querying
AUTHORS: Filipe Santana Da Silva, Fred Freitas and Stefan Schulz

---------- REVIEW ----------
The manuscript presents a method to give explicit semantics to database content, to translate the content as ontologies, to enable reasoning for consistency checking, and to improve query experience.

While it is an interesting idea to translate database content as subclasses of existing ontology classes, it seems the work is at a preliminary or experimental status.

First, the scale of experiment (903 classes out of 46 records) is too small compared to the scale of actual databases. Although the authors present extended experiments with simulated data increased by a factor of 30, still the scale is too small to think about serious application fully exploiting available public databases.
I am wondering if the experimental result can be generalized to entire UniProt and Ensemble, in terms not only of reasoning performance but also of the complexity of manual interpretation work.

There are other (lighter-weight) approaches to give explicit semantics to existing databases, e.g., D2RQ, Aber-OWL. It would be helpful if such approaches are discussed in comparison with the presented approach.