Ontological interpretation of biomedical database annotations

Filipe Santana da Silva[1,*], Ludger Jansen [2], Fred Freitas[1] and Stefan Schulz[3]

[1] Centro de Informática (CIn), Universidade Federal de Pernambuco (UFPE), Recife, Brazil
[2] Institut für Philosophie, Universität Rostock, Germany
[3] Institut für Medizinische Informatik, Statistik und Dokumentation, Medzinische Universität Graz, Austria

ABSTRACT

Motivation: In general, the meaning of biological database records is not sufficiently specified from an ontological point of view. We explore the options for an ontology-based integration and interpretation of database content of individuals, defined classes, dispositions and a combination of these.

Results: Four interpretation models are created, interpreting annotations in database records as referring to (i) individuals, (ii) defined classes, (iii) disposition universals, and (iv) a combination of these. Evaluation is done by using competency questions to test the retrieval capacities.

Availability: Interpretation models and sample data are available at http://www.cin.ufpe.br/~integrativo.

*** Contact:** fss3@cin.ufpe.br

1 INTRODUCTION

Biological databases (BIO-DBs) are used to store summarized results of laboratory experiments. Apart from numeric and textual entries, they include semantic annotations. E.g., the Unified Protein Resource (UniProt) (The UniProt Consortium, 2015) includes annotations from the Protein Ontology (PR) (Natale et al., 2014) and the Gene Ontology (GO) (The Gene Ontology Consortium, 2014). While these ontologies, in isolation, obey formal principles and convey precise meaning, the meaning of the database record as a whole remains vague and depends on implicit background assumptions. What it means when, e.g., in an annotation the UniProt protein term *Methionine synthase* is linked to the GO process term *Methylation*, is left to the user. Hence, on the one hand, we have rich and well-curated BIO-DBs with highly structured tabular content, but limited ontological explicitness. On the other hand, large bio-ontologies provide formal descriptions of their content, enabling logic-based reasoning. In order to use these features with BIO-BDs, we want to make explicit what annotations exactly refer to and to express this in a formal, computer-processable way.

It has already been argued that there are benefits for content retrieval, regarding correctness, completeness, and user-friendliness given a seamless integration between BIO-DBs and ontologies, and that such systems could accommodate large amounts of data from BIO-DBs (Hoehndorf et al., 2011; Santana et al., 2011). It is, however, still an open question (1) how implicit knowledge about the entities and relationships described in the structure of a BIO-DB be represented, (2) whether the content denoted by BIO-DBs (i.e. the domain entities represented by the data elements and the way how the former are connected) is fit to be represented, and, if this is the case, (3) how it can be translated into axioms using appropriate representational patterns, and finally, once database structure and content are expressed by formal-ontological means, (4) how the existing bio-ontologies can be plugged into this structure. Addressing these questions, we demonstrate that there are feasible ways to express implicit and explicit database content by formal-ontological means and combine it with existing domain ontologies. We show how annotation terms used in a typical BIO-DB entry can be interpreted as referring to entities from different ontological categories. Each of these interpretations requires different means like the introduction of individuals, the addition of new axioms to existing classes or the introduction of additional defined classes. The resulting OWL models are tested under three aspects: (i) database content retrieval, using ontologies as query vocabulary for data integration; (ii) information completeness; and (iii) reasoning behaviour in Description Logics (DL).

2 METHODS

For the analyses, we selected a typical example from biomedical databases, generated by joining data from UniProt and Ensembl (Cunningham et al., 2014). Records in BIO-DBs are mainly composed of (i) one protein term (e.g., CBS); (ii) one taxon term (e.g., *Rattus norvegicus*); (iii) one to many terms from GO for biological processes (e.g., *Methylation*); (iv) one to many terms from GO for cellular components (e.g., *Cytoplasm*); (v) zero to many phenotype terms (e.g., *Endocrine pancreas increased size*); and (vi) one to many small molecules (e.g. *Homocysteine*). We implement four different interpretive strategies (IND, SUBC, DISP and HYB) in OWL using the editor Protégé v.5 and the reasoner FACT++ (Tsarkov & Horrocks, 2006) to check for consistency and taxonomic subsumption. We used *BioTopLite2* (BTL2) as an upper-level ontology with highly constrained classes and a small set of relations (Schulz & Boeker, 2013). To test each interpretation model, we created four competency questions (CQs), first in natural language, and then translated into DL queries.

3 RESULTS

3.1 Individuals as the referents of annotations (IND)

The first interpretation rests on the fact that a database entry is about the outcome of a concrete experiment. Accordingly, the annotations that feature in such an entry can be interpreted as referring to the individual molecules, objects and processes that belonged to that particular experiment. Thus, the entry "Cystathionine gamma-lyase" denotes a molecule or a collection of molecules of the class '*Cystathionine gamma-lyase*'. BIO-DB content is therefore represented as a set of Abox-level class-membership assertions and relationships.

3.2 Subclasses as the referents of annotations (SUBC)

Second, database content can be interpreted by means of a number of maximally fine-grained defined classes, introduced by means of equivalence axioms for each universal entity which the annotations refer to. For instance, the annotations of a record combining the protein term *Methionine synthase* and the species term *Rattus norvegicus* are represented by a customized defined class combining the information about a subclass of *Methionine synthase*, defined as *Methionine synthase* that is part of an organism of the type *Rattus*

norvegicus. Using OWL-EL expressiveness, we can formalize this as follows:

'*Methionine synthase_in_Rattus Norvegicus*' *equivalentTo*
Methionine_Synthase and ('**is part of**' some '*Rattus norvegicus*')

3.3 Dispositions as the referents of annotations (DISP)

Real world entities are often described scientifically in terms of dispositions, i.e. tendencies to behave in a certain way under certain circumstances. Biomedical observations yield statistical results indicating that participants of an experiment (a protein *Methionine synthase*) have dispositions to bear certain capabilities (Jansen, 2007), like being able to perform a *Methylation* process. Interpreting database entries as statements about dispositions means that we represent the database content regarding a disposition of organisms of a certain species, e.g., that all instances of *Homo sapiens* have the disposition to develop a pathological condition *P*. For this purpose, we use *General Class Inclusion* (GCI) axioms that allow for subclass assertions between two complex class expressions, e.g.:

'*Endochondral ossification'*
and ('**is included in**' some '*Bos taurus*')
subClassOf '**has participant**' some '*Cysthationine beta-synthase*'

The output of DISP is an ontology file representing the classes referred to by the annotations together with a small set of GCIs, using DL-*SHI* expressiveness.

3.4 Hybrid interpretation (HYB)

To avoid the complexity of GCI expressions, we combine SUBC with DISP. HYB uses subclass statements like SUBC, enriched by axioms about dispositions like in DISP. This combination reduces the amount subclasses to be created. Disposition axioms are limited to material objects like proteins and organisms, asserting that they are capable of participating in specific biological processes. The HYB output needs DL-*SHI* expressiveness.

3.5 Fitness test

The four ontology models were tested for consistency and the following queries were used for retrieval evaluation: (Q1) Which biological processes have proteins of the kind $Prot_1$ as participants? (Q2) In which cellular locations is $Prot_2$ active in organisms of the type Org_1? (Q3) Which proteins are involved in processes of the type BProc in organisms of the type Org_1? (Q4) Which organisms are able to exhibit a specific phenotype $Phen_1$? – These queries were translated into DL, which enabled the retrieval of content in interpretations IND, SUBC and HYB. The model HYB was the only one able to retrieve content for Q4. As DISP expresses everything in GCIs, retrieval is not supported at all.

4 DISCUSSION

We proposed four interpretation strategies: IND, SUBC, DISP and HYB. Of these, only IND is completely based on single individuals (Abox entities). Ceusters et al. (2014) use a similar approach for applying relations between individuals in electronic health records.

SUBC is based on generating customized definitions of classes. This approach is not far from the work of Hoehndorf et al. (Hoehndorf et al., 2011). However, in SUBC an annotation does not refer directly to the class matching to the annotation term, but to a defined subclass of it. This requires a non-standard interpretation of DL queries, targeting the existence of subclasses. On the downside, SUBC involves an excessive number of subclasses. However, this does not have severe consequences on performance because of the good scaling behaviour of OWL-EL ontologies. This has also been confirmed by our preliminary experiments.

DISP alone is not helpful for most of the queries. It provides a more compact representation, but it is also incomplete because not all knowledge embedded within a database record can be sensibly expressed by dispositions. The combination of SUBC and DISP in HYB has finally the huge advantage that it enables querying whether certain biological entities are capable of participating certain processes, assuming that we agree that parts of the underlying knowledge in BIO-DBs is about dispositions.

5 CONCLUSION

We proposed four ontological representations of structure and content of biological databases. The solutions we presented targeted aspects of ontology-based database retrieval, expressiveness and content retrieval based on DL reasoning. Only part of database content is really of ontological nature in a strict sense, i.e., expressible by axioms that hold universally for all instances of a class. We addressed this limitation by three ways. Firstly, we interpreted the denoted entities as (prototypical) individuals, which requires representation and reasoning on an Abox level. Secondly, we expressed contingent database content by creating defined subclasses for which then universally valid statements could be made. Thirdly, we interpreted part of the database content as reporting dispositions, which was, however, not very helpful for the answering of our queries, in contrast with the second modelling approach, when DL reasoning was used to check for the existence of subclasses.

Funding: This work was funded by *Conselho Nacional de Aperfeiçoamento de Pessoal de Nível Superior* (CAPES) 3914/2014-03 and *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq) 140698/2012-4.

REFERENCES

Ceusters, W., et al. (2014). Clinical Data Wrangling using Ontological Realism and Referent Tracking. In W. R. Hogan, et al. (Eds.), *ICBO 2014* (pp. 27–32).

Cunningham, F., et al.(2014). Ensembl 2015. *Nucleic Acids Research*, *43*(D1), D662–D669.

Hoehndorf, R., et al. (2011). Integrating systems biology models and biomedical ontologies. *BMC Systems Biology*, *5*, 124.

Jansen, L. (2007). Tendencies and other Realizables in Medical Information Sciences. *The Monist*, *90*(4), 1–23.

Natale, D.A., et al. (2014). Protein Ontology: A controlled structured network of protein entities. *Nucleic Acids Research*, *42*(D1), D415-D421

Santana, F., et al. (2011). Ontology patterns for tabular representations of biomedical knowledge on neglected tropical diseases. *Bioinformatics*, *27*(13), i349–i356.

Schulz, S., & Boeker, M. (2013). BioTopLite: An Upper Level Ontology for the Life Sciences. Evolution, Design and Application. In M. Horbach (Ed.), *Informatik* (pp. 1889–1899). GI.

The Gene Ontology Consortium. (2014). Gene Ontology Consortium: going forward. *Nucleic Acids Research*, *43*(D1), D1049-D1056.

The UniProt Consortium. (2015). UniProt: a hub for protein information. *Nucleic Acids Research*, *43*(D1), D204–D212.

Tsarkov, D., & Horrocks, I. (2006). FaCT ++ Description Logic Reasoner : System Description. In *LNCS* (pp. 292–297). Springer: Berlin/Heidelberg.