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 database records like in biological databases is not sufficiently specified from an ontological point of view. This paper explores the options for an ontology-based integration and interpretation of database content. of individuals, subclasses, dispositions and a combination of these.

Results: We create and evaluate four interpretation models, making use of Four database structure interpretation patterns are created, making use of (i) individuals, (ii) defined subclasses, (iii) disposition universals, and (iv) a combination of these. Evaluation is done by competency questions testing the retrieval capacities.

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

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

1 INTRODUCTION

Biological databases (BIO-DBs) store summarized results from scientific investigations. 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).

Whereas the ontologies, in isolation, obey formal principles and convey precise meaning, the meaning of the complete database records remains vague and depend on shared background assumptions. What it means when, e.g., the protein *Methionine synthase* is linked to the GO process *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, in an increasing number of large bio-ontologies logic-based axioms provide ontologically unambiguous formal descriptions of their content, enabling formal reasoning.

It has already been argued that there are benefits for content retrieval, regarding correctness, completeness, and user-friendliness given a seamless integration between BIO-BDs 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 ontologically, (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 ontologically, and (3) if this is the case, how it can be translated into axioms using appropriated representational patterns, and, (4) once database structure and content are expressed by ontological means, how the existing bio-ontologies can be plugged into this structure. Addressing these questions, we hypothesise that there are feasible ways to express implicit and explicit database content by formal-ontological means and combine it with existing domain ontologies. We demonstrate how entities referenced by a typical extract (as annotations) from BIO-DBs can be interpreted under several ontological viewpoints, *viz.* regarding the introduction of individuals, the addition of new axioms to existing classes and 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) Description Logics (DL) reasoning behaviour.

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). DB records are mainly composed by (i) one protein (e.g. CBS); (ii) one taxon (e.g. *Rattus norvegicus*); (iii) one to many GO biological processes (e.g., *Methylation*); (iv) one to many GO cellular components (e.g., *Cytoplasm*); (v) zero to many phenotypes (e.g., *Endocrine pancreas increased size*); and, (vi) one to many small molecules (e.g. *Homocysteine*). We create for ontologies (individuals, subclasses, disposition and hybrid) written 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 Interpreting annotations as individuals (IND)

The first interpretation rests on the fact that database content is derived from reports of experimental assays. Hence, they deal with observations about individual entities. Data entries are always individuals, e.g. the annotation “Cystathionine gamma-lyase” is an information entity. But in this interpretation, these data individuals denote individual substances, biological objects and processes in a concrete experiment. Thus, the entry “Cystathionine gamma-lyase” denotes a member of the class ‘*Cystathionine gamma-lyase*'. DB content is therefore represented as a set of ABox-level class membership assertions and relationships.

3.2 Interpreting annotations as subclasses (SUBC)

Here, database content is interpreted as a set of maximally fine-grained equivalence class axioms for each entity to which the annotations refer to (e.g. classes). For instance, an annotation about the protein *Methionine synthase* in rats is here represented by asserting a link between the annotation individual and a subclass of *Methionine synthase*, defined as *Methionine synthase* part of an organism of the type *Rattus norvegicus*. The output of SUBC is a set of defined classes, such as:

'*Methionine synthase_in_Rattus Norvegicus*' *equivalentTo*
Methionine_Synthase
and (**'is part of'** some '*Rattus norvegicus*')
within the OWL-EL expressiveness.

Axioms like this are created to highlight the denotation among individuals from ABOX and their correspondences in TBOX.

3.3 Interpreting annotations with dispositions (DISP)

Real world entities are often described scientifically in terms of dispositions, i.e. tendencies of something to act in a certain manner 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 databases with dispositions means that we represent the database content as an indication of 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 with classes referred by the annotations, and a small set of GCIs (DL-*SHI* expressiveness).

3.4 Hybrid interpretation (HYB)

To avoid the complexity of GCI expressions, we combine SUBC with DISP. HYB includes subclass statements like in SUBC, enriched by axioms on 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 ontology models were tested for consistency and the following queries were used for retrieval evaluation: (Q1) Which biological processes have proteins of the kind Prot1 as participants? (Q2) In which cellular locations is Prot2 active in organisms of the type Org1? (Q3) Which proteins are involved in processes of the type BProc in organisms of the type Org1? (Q4) Which organisms are able to exhibit a specific pheno-type Phen1?

These queries were translated into DL, which enabled the retrieval of content in interpretations ABOX, SUBC and HYB. The model HYB was the only one able to retrieve content for Q4. As DISP expresses everything in GGIs, retrieval is not enabled.

4 DISCUSSION

We proposed four interpretation strategies: ABOX, SUBC, DISP and HYB. ABOX 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 definitions of subclasses and classes. This approach is not far from the work of Hoehndorf et al. (Hoehndorf et al., 2011). However, in SUBC we distinguish that annotations refer to subclasses, and not directly to the class referred by the annotation. This requires a non-standard interpretation of DL queries, which target on 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 could also be confirmed by 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 DB record can be sensibly expressed by dispositions. The advantage of combining SUBC and DISP in HYB is the ease of querying whether biological entities are capable of participating certain processes, assuming that we agree that parts of the underlying knowledge in BIO-DBs is dispositional.

5 CONCLUSION

We proposed ontological representations of structure and content of biological databases. The solutions we presented covered 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 all denoted entities as (prototypical) individuals, which requires representation and reasoning on an ABox level. Secondly, we expressed contingent DB content by creating subclasses for which then universally valid statements could be made. Thirdly, we interpreted part of the database content as 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*(1), 124.

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

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

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),

The UniProt Consortium. (2015). UniProt: a hub for protein information. *Nucleic Acids Research*, *43*(Database), D204–12.

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