RESEARCH

Re-interpreting biomedical database content as ontologies for enhanced querying

Filipe Santana da Silva[1*†], Robert Hoehndorf[2†], Fred Freitas[1†] and Stefan Schulz[3†]

*Correspondence: fss3@cin.ufpe.br
[1]Centro de Informática, Universidade Federal de Pernambuco, Av. Jornalista Anibal Fernandes, 50.740-560 Recife, Brazil
Full list of author information is available at the end of the article
†Equal contributor

Abstract

Motivation: Although content from biomedical domain ontologies constitutes a major source for biological databases, current querying strategies mainly follow the syntax of the underlying database scheme and use ontologies mainly as domain vocabularies. We address this shortcoming by ontology grounding, an approach that makes hidden ontological assumptions of databases explicit and exploits the representational richness of the related domain ontologies. The result is an ontological resource, in which database and domain ontology content is seamlessly integrated as a large set of OWL axioms, underpinned by formal-ontological principles. This resource is generated fully automatically, guided by a set of manually crafted, database-specific design patterns.

Results: We applied ontology grounding to a database extract (Uniprot, Ensembl) on amino acid metabolism. Ontology completeness and correctness were tested using pre-formulated competence questions as description logics (DL) queries, encompassing proteins, biological processes, molecular activities and dysfunctional phenotypes. The expected results could be reproduced by the description logics reasoner. Benchmark tests showed moderate scalability. Real-world use of this method still depends on large-scale content filtering and modularization prior to ontology creation.

Availability: http://www.cin.ufpe.br/~integrativo

Keywords: sample; article; author

Introduction

Biomedical research increasingly depends on large databases, such as UniProt [1] or Ensembl [2]. The formulation of database queries and the correct interpretation of query results requires knowledge of the database structure as well as implicit background knowledge about the way data is produced and recorded. Such background knowledge may vary between users, thus biasing uniform data interpretation.

There have been several attempts to enhance data retrieval by formal ontologies like ontology-based data access (OBDA) [3] or the combination of ontologies with machine learning [4]. However, such approaches require in-depth manual interpretation of the finally retrieved content; and the representation of data entities as individuals (ABox elements in description logics) results in high processing cost [5].

Accordingly, the current situation regarding the (semi-)automated support for database content retrieval and interpretation is characterized by a concurrent and continuous evolution of high-quality structured knowledge resources, whereas less progress can be seen regarding their interoperability and ontological underpinning. Superficial use of ontologies - just as vocabularies - restricts their applicability. We

hypothesize that with a formally explicit view on biomedical data as provided by principled ontologies, users are better served to integrate, retrieve, validate and interpret these data, due to automated classification and consistency checking provided by Description Logics (DL) representation and reasoning [6].

We investigate this hypothesis by proposing a model for the seamless integration of the content of biological databases and ontologies, guided by formal-ontological principles [7], which enforce univocal interpretation of database content in order to improve (semi-)automated data interpretation and understanding. Ontology-level content in identified in databases, and axiomatized under an upper level ontology. Thus it delivers a homogeneous representation of data, linked to (modules of) existing ontologies. The proposed approach might be applied to enhance database curation, as new statements extracted from publications and recorded in databases can be tested for adequacy. Supported by description logics (DL) based reasoning, it might enable validation using DL Query as an expressive and simple query language based on DL syntax and semantics.

We will substantiate this claim by investigating a subset of biomedical ontologies and databases. We (i) propose an ontology-based framework that makes explicit both database content and the domain entities denoted thereby; (ii) relate this framework to current workflows in which life science data and knowledge are acquired and processed; (iii) implement an example ontology from real data as an exemplar for database-ontology integration; (iv) validate it by demonstrating how querying is done by using DL Query; and (vi) experimentally assess correctness and scalability.

The biological use case is addressed by Competency Questions (CQs) [8] as DL queries on the metabolism of small molecules in model organisms, encompassing physiological processes as well as phenotypical changes in dysfunctional metabolism. The examples make use of UniProt, NCBI Taxonomy [9], Ensembl[2], SNOMED CT [10], GO [11], ChEBI [12] and PR [13], organized under the upper domain ontology BioTopLite (BTL2) [14]. Our approach capitalizes on previous work about the formalization of tabular representations in scientific literature [15] and models of structured clinical information [16].

Background

Formal ontologies

Formal ontologies are centred on classes of individual entities. Together with a set of binary relations and operators in a logic-based language such as a DL dialect, they constitute formal axioms. Axioms are limited to statements about what is universally true for all members of a class or a class-like expression, e.g. that all nucleic acid molecules contain nucleotides, or that all adenine molecules are nucleotides. The construction of (formal) ontologies should obey principled criteria [17] and good practice guidelines [14]. Important principles are (i) naming conventions [18]; (ii) mutually disjoint upper-level classes like *Process* or *Quality* as a fundamental ordering framework; and, (iii) a limited set of canonical relations [19].

We distinguish *ontology content* properly from both the notions of *knowledge* [20] and *data*. Knowledge, in a broader sense [21], emcompasses assertions about what is frequently associated or what is true by default. This is beyond the scope of formal ontologies. Data, in contrast, are first of all information entities in a database.

Like words in a text, data elements constitute denoting entities. In databases, the interpretation of data elements is determined by the use cases embedded in the underlying schema. The distinction between what is a data element and what is denoted by it is often neglected and leads to use-mention confusions. Formal ontologies must enforce this distinction: all data items are instances of a specific class like *information object* in BTL2, or *generically dependent continuant* in BFO, whereas the things data items denote are manifold: individuals, classes, or even nothing [22].

Interpreting databases with ontologies

Databases and less-principled domain ontologies leave the real nature of the entities as well as and the circumstances of denotation underspecified, assuming that this is intuitively known by the users and interpreted accordingly within the expected context of use. An example is "Human", which may denote an individual person, the class *Homo sapiens*, the quality of an object belonging to the taxon *Homo sapiens* [23], or a population of humans. In a similar vein, "Animal" could be interpreted as including the class *Homo sapiens*, in the context of Biology or excluding it, e.g. in the context of Law. Such ambiguities underline the need to make hidden assumptions explicit, which is the main rationale for the ontological grounding mechanism as described in the following. Content retrieval applications that use existing domain ontologies as vocabularies, like the ones derived from OBDA might benefit from this. As ODBA tools are unable to retrieve and generate new ontology content, the interpretation goes beyond what the user has specified as task-specific mappings between databases and ontologies. This requires that databases do not only use domain ontologies as standardized vocabularies, but that the meaning of their entire structure and content is described by ontology axioms and assertions. This is what we propose.

Description Logics and OWL2

Description Logics (DLs) are representation languages used to formalise ontology content[1]. DL classifiers, like Fact++ [24], find new subclass links, identify equivalent classes, and assure satisfiability, by spotting contradictory axioms. An important distinction of ontology content is between ABox and TBox. TBoxed are constituted by class level axioms (e.g., "all chimps are primates"), whereas ABoxes contain assertions on individuals (e.g., "Washoo is a chimp").

The Semantic Web Standard OWL2 [25] uses the DL language $SROIQ$ [26], with limited expressiveness but complete and finite reasoning. OWL2 supports classes, binary relations (called object properties), and individuals, together with related axioms and assertions. For instance, the OWL2 class *Drosophila melanogaster* has all individual fruit flies as members. Because all individual fruit flies are also members of the class *Organism*, the class *Drosophila melanogaster* forms a subclass of the class *Organism*. Such class statements are constructed by the combination of operators specified in OWL2; *viz.* 'and' for conjunctions, 'or' for disjunctions, 'some' for existential restrictions, and 'only' for value restrictions, under the Manchester Syntax for OWL2 [27], which we will use in this paper.

[1]For DL syntax and semantics, Cf. [6].

Application background

For the use case, we use data and ontologies related to the metabolism of homocysteine (Hcy). Hcy is an amino acid, which plays a key role in vitamin and cofactor metabolism, neuronal metabolism, and in the biological oxidation of enzymes [28]. High levels of Hcy are reported to play a role in the pathogenesis of atherosclerosis [29], and of hepatic steatosis in hepatitis C infected subjects [30]. Many organisms host Hcy-related bioprocesses, e.g. *Mus musculus*, *Homo sapiens*, *Gallus gallus*, *Schizosaccharomyces pombe*, and *Oryza sativa*.

Ontology grounding

Ontology grounding means the detailed description of database structure and content by formal ontology mechanisms. This process requires in-depth domain knowledge, insight into the way databases are populated, as well as ontology engineering skills, based on the understanding of upper level ontology principles and description logics. Aware that a straightforward, automatized "ontologization" of a database schema is not sufficient, the ontology engineer has to critically assess the pros and cons of competing modelling strategies, in a way that correctly accounts for the underlying domain segment, and which provides enough expressiveness to address typical queries (Fig. 1).

This grounding process starts with the identification of typical domain-related queries and their definition as Competency Questions (CQs) [8]. Then databases that are needed to answer these CQs are selected, as well as the ontologies that are referred to by these databases. Ontology annotations in databases as well as database structure and content are identified and linked to the ontology entities they denote, which can be TBox or ABox entities.

Top-level classes and relations of the domain ontologies are mapped to an upper level ontology. During this mapping process, consistency is iteratively checked by a DL reasoner. Repetitive structures are identified in database records according to the ontology organization and described as generalizable ontology design patterns. In this process, the interdependencies and relationships between the referents and/or their types are analysed, based on domain knowledge, and the need for newly defined subclasses is assessed.

General patterns are applied to the overall databases or to database subsets under consideration and are formatively evaluated by test queries that were formulated beforehand. When the content generated is logically consistent and the answers to the test queries are complete, the process reaches the end, and the output corresponds to an ontologized interpretation of the database. Otherwise, more iterations and evaluations are required in order to meet consistency, domain validity and user expectations.

For assessment and demonstration of the ontology grounding approach we have formulated the following CQs on Hcy metabolism. They explore database content concerning phenotypes, proteins, molecules and biological processes from several organisms, using ontologies and databases described in Section **??**).

1 Which kinds of biological processes related to Hcy can be found in mice?
2 Which are the proteins that exhibit *methyltransferase activity*?
3 Which are the kinds of biological processes in which proteins of the type *cystathionine gama lyase* participate, exhibiting *carbon-sulfur lyase activity*?

4 Which dysfunctional biological processes entail a risk of *Myocardial infarction*?
5 Which kinds of organisms are capable of performing '*cysteine biosynthetic process*'?
6 Which proteins found in ruminants have the capability of methionine biosynthesis?

Resources

0.1 Biomedical Ontologies

- The **Gene Ontology** (GO) [11] was created in 1998 to address biomedical information integration through standardization of terms for the annotation of DNA sequences and their respective characteristics. GO has become a crucial resource for functional genomics, as an ongoing collaborative effort that delivers a controlled vocabulary underpinned by an ontology language. GO provides class hierarchies under *Cellular component*, *Biological process*, and *Molecular function* (ontologically better described as molecular activities or processes), together with relations between them.
- **Chemical Entities of Biological Interest** (ChEBI) [12] describes low-molecular-weight chemical entities for understanding and intervening in biological functioning. Each ChEBI entry denotes a chemical structure in a graphical form, together with ontological axioms. The ontology is subdivided into *Molecular structure* and *Biological role*. Whereas the former represents the structure of small molecules and their constituents, the latter is used to classify molecules depending on their disposition of participating in biological processes.
- The **Protein Ontology** (PR) [13] is held by the Protein Information Resource (PIR), integrating several databases and responsible the current structure for the UniProt database. It represents modified forms, isoforms and protein complexes from living organisms and provides relations between them.
- **SNOMED CT** [10] is a large clinical terminology for human and veterinary medicine, containing formal definitions, which can be expressed as an OWL-EL ontology. SNOMED CT covers clinical findings and disorders, body parts, devices, drugs, substances, organisms and clinical procedures, among others.
- **BioTopLite 2** (BTL2)[31] is a lightweight and redesigned version of BioTop, created in 2006 as an upper-domain ontological layer to enable the representation of general aspects of biology and medicine. BTL2 offers highly constrained classes, using a small set of relations. Classes like *Organism*, *Mono molecular entity*, and *Body part* facilitate the alignment with other ontologies like GO, PR and ChEBI. BTL2 can be aligned with most of BFO and RO. Available biomedical ontologies compliant with these two sources can easily be integrated with BTL2.

0.2 Biological Databases

- The **Universal Protein Resource** (UniProt) [1] was created in order to enable a quick understanding of the field of proteomics. It provides a comprehensive, open-access resource of protein sequences and functional information.

UniProt is mainly composed by a Knowledge Base (UniProtKB), subdivided in SwissProt (manually curated) and TrEMBL (generated and maintained by automated tools). Other parts are databases for sequences, closely related protein sequences, protein information from fully sequenced organisms, and metagenomics. Data from literature and available in UniProt are organized and stored according protein and gene names, function, catalytic activity, co-factors, pathway information, sub-cellular location, among others. UniProt embeds NCBI Taxonomy identifiers directly throughout its structure, as well as GO annotations [32], together with mappings to several biological databases including Ensembl.

- The **Ensembl** [2] project was launched in 1999 in order to automatically annotate genomes and to integrate this data with other biological data sources, thus creating a freely available online source. Ensembl processes and summarizes large-scale genomic data for chordates and model organisms. Its content is related to the annotation of gene and transcript locations, gene sequence evolution, genome evolution, sequence and structural variants and regulatory elements.
- The **NCBI Taxonomy** [9] was derived from a project on the taxonomy of biological organisms that aimed at extracting sequences not available in dedicated databases from genomic literature. This coincided with the collection of data about taxonomic classifications. The goal of NCBI Taxonomy is to combine existent, distributed organism taxonomies into a single one that is included in NCBI GenBank.

Methods

In the following, data acquisition and database conversion are described. Content and related files, such as spreadsheets, scripts, and ontology files can be downloaded from the project website (http://www.cin.ufpe.br/~integrativo).

0.3 Sampling

Data related to 21 species[2], together with processes and by-products related to Hcy metabolism are retrieved from the UniProt and Ensembl websites[3]. The ontologies GO, ChEBI and BTL2 are downloaded in OWL2 format[4].

For the creation of a subset from UniProt and Ensembl, UniProt data are filtered by the string "homocysteine", thus retrieving all Hcy-related data from UniProt/SwissProt+Trembl. From the obtained 212,156 records the ones with GO annotations, specified gene names and proteins as described by [28] are selected. From the resulting 1,716 records fragments, isoforms, or homologue

[2] 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

[3] UniProt: Release 2015_04, Ensembl Release 79, NCBI Taxonomy 2015AA.

[4] GO Revision 25527, ChEBI Release 127, BTL2 Release 8th march 2015, PR release 22nd may 2015, SNOMED CT July 31st 2014.

entries are excluded. The resulting set includes the proteins Methionine synthase (MS), Methylenetetrahydrofolate reductase (MTHFR), Cystathionine beta-synthase (CBS), and Gamma-cystathionase (CSE). After removing records without Ensembl IDs, a final sample with 46 Hcy-related records is made available as a Microsoft Excel spreadsheet with the following tabular structure:

- One Protein (e.g. *CBS)*;
- One Taxon (e.g. *Rattus norvegicus*);
- One to many GO biological processes (e.g., *Blood vessel remodeling*)
- One to many GO molecular functions (e.g., *CBS activity*)
- One to many GO cellular components (e.g., *Cytoplasm*)
- Zero to many phenotypes (e.g., *Endocrine pancreas increased size*)

0.4 Ontology mappings

Root classes from GO, PR, ChEBI, NCBI Taxonomy, and SNOMED CT are included as subclasses of BTL2 nodes and tested for logical consistency (Fig. 2). Mapping consistency is assured with the DL reasoner HermiT 1.3.8. For performance optimization, modules of external ontologies (GO, ChEBI, PR and SNOMED CT) are created, with the classes referred to in the selected database content as signature, using the Protégé plug-in Ontology Modularity [33].

All database objects are subjected to in-depth ontological analysis done by the authors. The content of a representative sample record is entirely modelled in Protégé as an OWL ontology, and tested for consistency and adequacy under BTL2. Once agreement is reached, this sample ontology is saved in OWL/XML format and manually dissected into code fragments, each of which is numbered and placed into a spreadsheet cell. The code fragments are analysed to identify variable elements, i.e. class names specific to the underlying sample, encompassing protein, cellular component, biological function and other classes. All these names are then replaced by placeholders, thus yielding an ontology pattern specific to the data structure. The target ontology is then generated by iteratively filling the placeholders with the ontology class names from the database. This is done using a customized VBA script. This script has to account for the creation of named subclasses as well as for iterations over content of multi-valued fields.

0.5 Evaluation methodology

The ontological content is evaluated by six competency questions (CQs),as introduced above. Formulated in English by the first author, a biologist, they are shaped according to how domain experts would query a biological database, and not how ontology engineers would interpret it, in order to be neutral regarding the internal structure of the ontology. The translation of CQs into DL queries relies on the correct identification of query components that denote relations, referents, and the way how domain entities are related to one another (cf. section 0.5). 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.

Scalability is evaluated by artificially increasing the size of the ontology by a scaling factor $f \in 1$, 3, 10, and 30. This is done programmatically, creating new

classes suffixed by $i \in 1, 2, \ldots, f$. 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 classification time. In addition, satisfiability time is measured, defined as the time it takes for validating a CQ against the derived model from the test ontology. All tests are performed in Intel core i7-4510U laptop, with 8Gbytes of RAM, running Windows 10 (x64) and Java 8 (update 66, x64).

Results

Basic ontological assumptions

The thorough inspection of the database (DB) content and its discussion among the authors yielded the following ontology-based interpretation: "Each DB record can be understood as introducing a series of defined subclasses of each biological process class referred to by this DB record. Each of these subclasses is defined by having one or more proteins of a certain type as participants, together with the small molecule Hcy. These proteins occur within one or more cell components. If dysfunctional, these processes lead to the risk of developing the pathological phenotypes mentioned in the source".

Thus, the content of the biological database extract under scrutiny is entirely expressed at class level. The underlying assumption is that all these defined process subclasses are non-empty, as otherwise there would not have been any experimental evidence manifested as a curated database entry. Additionally, we assume that no wrong data occur (data instances that do not have any referent in reality). This interpretation allows us to refrain from reasoning about individuals, thus avoiding scaling problems.

Ontological Grounding

In the following, the ontological grounding steps of the selected content is described. Table 14) shows a subset of the table created as a view from UniProt and Ensembl (Table 14). We translate the content from table 14 as an example table 15 for interpretation. As a first task, we determine the ontological representation of data in Table 15:

- There exist biological processes of the type *Bp* in organisms of the type *O* that have the protein *P* and the small molecule *M* as participants;
- In each *Bp*, the protein *P* is capable of performing one or more molecular functions (processes) *Mf* ;
- *Bp* processes occur in one or more types of cellular components *C*;
- There exist biological processes of the type *Bp* that are dysfunctional and therefore bear the risks of causing one or more pathological phenotypes of the type *Ph*;
- All organisms of the type *O* have dispositions to be realized by *Bp* processes;
- All types of protein *P* in *O* are able to perform *Mf* processes;
- Proteins of class *P* are not organism-specific. However, as the DB records refer to organism-specific proteins we introduce subclasses *P_sensu_O* for each DB record (Protein *P* from Organism type *O*).
- Each *Bp* referred to by in a DB record may happen within the same cellular structure in several organisms *O*, including organism-specific proteins *P* and

molecules *M*. However, each DB record denotes an exclusive occurrence of *Bp*. In this sense, each DB record is interpreted as denoting specific subclasses of *Bp*, identified as *Bp_in_O_with_P_and_M*, generated as a combination of biological process, organism, protein and small molecule;

- The database structure leaves open in which cellular component C a given *Bp* subclass is located, when there is more than one entry in the cellular component field. For this reason, we generate union classes of the type C_1 or C_2 or ... or C_n to which the process locations can be safely assigned.
- The DB structure is not explicit enough to connect a *Bp* subclass to a specific *Mf* process. Therefore in the definition of each *Bp* subclass the *Mf* processes are attached to the protein agent of that *Bp* subclass as possible realisations of the related disposition class ('**is realized by**' only *Mf*);
- If there are phenotype entries *Ph*, a new class of the type *Dysfunctional_Bp_in_O_with_P_and_M* is generated for every *Bp_in_O_with_P_and_M*, and all phenotypes are referred to as being the realizations of risks.

With this strategy, process subclasses appear in query results, which may appear confusing for the user, who only expects the parent classes from which they are derived, i.e. biological process classes from GO, as they exist as entries in database records. This is the reason why the query results need to be post-processed in a way that only the original superclasses are filtered out. E.g., if the class *BpX_in_O_with_P_and_M* is retrieved by the query, only its superclass *BpX* is displayed.

Ontology patterns

This analysis allows us to identify ontology patterns. First, we present the definitions of *P*(table 1), *Bp* (table 2) and *C* (table 3).

Table 1 shows how organism-specific proteins types *P* are introduced as defined subclasses of pr:*Protein*.

Table 1 Defined subclasses of proteins *P*

P_sensu_O equivalentTo *P* and ('**is included in**' some *O*)
P subclassOf pr:*Protein*
P_sensu_O subclassOf *P*

The composed name denotes a species-specific protein class: *P_sensu_O* is a sublass of *P*.

Table 2 Defined subclasses of biological process *Bp*

Bp subclassOf go:'*Biological process*'
Bp_in_O_with_P_and_M subclassOf *Bp*
Dysfunctional_Bp_in_O_with_P_and_M subclassOf *Bp_in_O_with_P_and_M*

Biological processes *Bp* are subclasses of the GO class *Biological process*, and the mention of a specific biological process in a database entry in the context of a specific organism, a specific protein, and a specific small molecule determines the creation of the class *Bp_in_O_with_P_and_M* as subclass of *Bp*.

Cellular components of any type *C* (within a single DB record) are defined as subclasses of the GO class *Cellular component* (Table 3).

Table 3 Cellular component C union classes

$C_1\_or\_C_2\_or\_\ldots\_or\_C_n$ subclassOf go:'*Cellular component*'
$C_1\_or\_C_2\_or\_\ldots\_or\_C_n$ equivalentTo C_1 or C_2 or ... or C_n

When a record refers to more than one cellular component class, union classes type C_1 or C_2 or ... or C_n are created under the GO class *Cellular component*. This is due to the fact that the DB structure is not explicit enough to connect specific cellular components to specific process subclasses.

Table 4 shows the axioms for *Bp_in_O_with_P_and_M* classes.

Table 4 *Bp_in_O_with_P_and_M*

Bp_in_O_with_P_and_M equivalentTo *Bp*
and ('**has participant**' some *M*)
and ('**has participant**' some
(*P* and ('**is bearer of**' some (btl2:*Function* and
('**is realization of**' only *Mf*)))
and ('**is included in**' some (C_1 or C_2 or ... or C_n))
and ('**is included in**' some *O*)

Axioms for *Bp_in_O_with_P_and_M* (Table 4) describe that a biological process from a single record has one or more small molecules as participants; the process is included in the combination of one or more cellular components within organism of a specific species; and the protein from the record is a participant in the process, and has the function of performing molecular processes of certain types.

Some *Bp_in_O_with_P_and_M* processes are dysfunctional and therefore entail the risk of pathological phenotypes (represented as SNOMED CT findings, ontologically subclasses of btl2:*situation*).

Table 5 Dysfunctional phenotypes of *Bp_in_O_with_P_and_M*

Dysfunctional_Bp_in_O_with_P_and_M equivalentTo
Bp_in_O_with_P_and_M
and ('**is bearer of**' some '*Dysfunctional Quality*')
Dysfunctional_Bp_in_O_with_P_and_M subClassOf
Bp_in_O_with_P_and_M
and ('**is realization of**' only (*Risk* and (**causes** some
Ph)))

Dysfunctional_Bp_in_O_with_P_and_M are processes defined by the quality of being dysfunctional. They are realizations of the risk (a sort of disposition) of causing the dysfunctional phenotype stated in the DB.

Table 6 presents the axioms required to represent *P_sensu_O*, i.e. the species-specific class of proteins *P*.

Table 6 Subclasses created for the organism specific protein (*P_sensu_O*) classes in database records

P_sensu_O equivalentTo *P* and ('**is included in**' some *O*)
P_sensu_O subClassOf *P* and ('**is bearer of**' some (*Function*
and
('**has realization**' only *Mf*)))

Definitions that follow the pattern from Table 6 describe organism-specific protein molecule classes. In addition, the *Mf* processes specified in the DB records are added.

The last axiom required is about organisms as bearers of dispositions related to performing biological processes (Table 7) .

Table 7 Axioms generated for organisms *O* in database records

O subClassOf btl2:*Organism* and ('**is bearer of**' some (*Disposition* and '**has realization**' only *Bp*)))

Table 7 attaches dispositions realized by specific biological processes to organisms.

Evaluating the content generated

The analysis of the content of database entries has resulted in a set of OWL T-Box axioms for each database record as specified above. We recall two basic assumptions made, *viz.* (i) non-emptiness of classes: i.e. each of the newly defined classed corresponds to at least one fact described in literature, and (ii) the veracity of database entries, i.e. each information is considered a statement of truth. Given these boundary conditions, evaluation of the generated TBoxes will address the aspects: (i) logical satisfiability when importing all constraints from the upper-level-ontology BTL2; (ii) adequacy (correctness and completeness) of entailments against CQs; and, (iii) computational performance.

Evaluation of Competency Questions (CQs)

In the following, each competency question is translated into a DL query. The result is analysed and discussed.

CQ1: Which kinds of biological processes related to Hcy can be found in mice? (Table 8) This query is intended to retrieve all biological process classes that takes place in organisms.

Table 8 Competency Question CQ1 in DL

'*Biological process*' and '**is included in**' some '*Mus musculus*'

34 subclasses (including ancestors) are retrieved. After filtering out the system-defined subclasses we obtain: *Amino acid betaine catabolic process*, *Blood vessel remodeling*, *Cellular response to hypoxia*, and other 31 biological processes. The results are expected as they correspond to the content of the database.

CQ2: Which are the proteins that exhibit methyltransferase activity? (Table 9) This query is meant to retrieve classes of proteins that are able to perform certain *Mf* processes. It is important that some proteins are capable to act in a specific way, like polymorphisms in gene MS leads to methionine synthase deficiency, which leads to higher Hcy levels together with dysfunctional phenotypes in humans and mice.

Table 9 Competency Question CQ2 in DL

Protein and ('**is bearer of**' some *Function* and ('**has realization**' only '*Methyltransferase activity*'))

38 subclasses (including ancestors) are retrieved. After filtering out the system-defined subclasses we obtain: *Betaine homocysteine S methyltransferase 1*, *Cystathionine beta synthase*, *Cystathionine gamma lyase*, *Cystathionine gamma lyase*, *Methionine synthase*, and *Methylenetetrahydrofolate reductase*. These five protein classes, which have the ablility to perform the process *Methyltransferase activity* correspond to the database content.

CQ3: Which are the kinds of biological processes in which proteins of the type cystationine gamma lyase participate, exhibiting carbon-sulfur lyase activity? (Table 10) This query is related to the identification of biological processes (e.g. biochemical reactions) that involve a specific protein (enzyme), which should play a role in this reaction. The relevance of this query is related to the capability of retrieving specific biological processes by means of reaction-specific proteins .

Table 10 Competency Question CQ3 in DL

'*Biological process*' and '**has participant**' some ('*Cystationine gamma lyase*' and ('**is bearer of**' some (*Function* and ('**has realization**' only '*Carbon-sulfur lyase activity*')))))

34 subclasses (including ancestors) are retrieved. After filtering out system-defined subclasses, the following eleven remain: *Cellular nitrogen compound metabolic process*, *Cysteine biosynthetic process*, *Endoplasmic reticulum unfolded protein response*, *Hydrogen sulfide biosynthetic process*, *Negative regulation of apoptotic process*, *Negative regulation of apoptotic signaling pathway*, *Positive regulation of I kappaB kinase NF kappaB signalling*, *Protein homotetramerization*, *Protein pyridoxal phosphate linkage via peptidyl N pyridoxal phosphate L lysine*, *protein sulfhydration*, and *small molecule metabolic process*.

CQ4: Which dysfunctional biological processes entail a risk of Myocardial infarction? (Table 11) This query retrieves biological processes that are associated with the risk of developing a dysfunctional phenotype. This query is relevant as it helps identify pathological phenomena associated with a specific biological process.

Table 11 Competency Question CQ4

'*Biological process*' and ('**is realization of**' only (*Risk* and (**causes** some '*Myocardial infarction*')))

Ten classes (including ancestors) are retrieved: *Dysfunctional homocysteine metabolic process*, *Dysfunctional methionine biosynthetic process*, *Dysfunctional one carbon metabolic metabolic*, *Dysfunctional response to amino acid*, *Dysfunctional response to drug*, *Dysfunctional response to hypoxia*, *Dysfunctional response to interleukin 1*, *Dysfunctional response to vitamin B2*, *Dysfunctional S adenosylmethionine metabolic process*, and *Dysfunctional tetrahydrofolate metabolic process*. None of these processes is related in the database with the entry *Myocardial infarctation*, but they are correctly retrieved because they are related to *Acute myocardial infarction of anterior wall*, which is a subclass of *Myocardial infarction* in SNOMED CT.

CQ5: Which kinds of organisms are capable of performing cysteine biosynthesis? (Table 12) This query retrieves organisms that are capable of performing specific biological processes. This query is relevant because not all biological processes for organisms are fully described. Organisms from different species that include for which the same proteins under the same conditions are described may not include similar processes

Table 12 Competency Question CQ5

Organism and (‘**is bearer of**’ some (*Disposition* and (‘**has realization**’ only ‘*Cysteine biosynthetic process*’)))

The following organism classes were retrieved: *Ailuropoda melanoleuca*, *Bos taurus*, *Callithrix jacchus*, *Danio rerio*, *Gallus gallus*, *Homo sapiens*, *Latimeria chalumnae*, *Loxodonta africana*, *Mus musculus*, *Mustela putorius furo*, *Oreochromis niloticus*, *Oryctolagus cuniculus*, *Pan troglodytes*, *Rattus norvegicus*, *Sarcophilus harrisii* and *Takifugu rubripes*.

CQ6: Which proteins found in ruminants have the capability of methionine biosynthesis? (Table 0.5) The aim of this query is to retrieve specific proteins that are related to the process *methionine biosynthetic process*, when performed by a organism of the suborder *Ruminantia*. In other words, we are able to identify specific proteins by means of organisms and biological processes among the content embedded in databases.

Table 13 Competency Question CQ6

Protein and (‘**is included in**’ some (*Ruminantia* and (‘**is bearer of**’ some (*Disposition* and (‘**has realization**’ only ‘*Methionine biosynthetic process*’))))

Five subclasses (including ancestors) are retrieved, of which the following are displayed after filtering: *Betaine homocysteine S methyltransferase 1*, *Cystathionine beta synthase*, *Cystathionine gamma lyase*, *Methionine synthase*, and *Methilenetetrahydrofolate reductase*. Although the entry *Ruminantia* is not found in the database source, the result is correct, because all entries refer to *Bos taurus*, which is a subclass of *Ruminantia* in the NCBI taxonomy.

Computational performance

In order to evaluate the computational costs of our approach, we compare the generated and mapped ontologies with an upscaled, artificially generated ontology having the same configuration.

Table 16 presents the reasoning performance of these OWL files with the queries CQ1-CQ6. It shows that under DL expressivity *ALC* classification time increases proportionally with the number of classes and axioms, without sacrificing satisfiability of queries and the retrieval of classes.

With the increase of expressivity from *ALC* to *SRI* (inherited from BTL2), reasoning time amounts to approximately two hours for classification and consistency checking. Thus, CQ satisfiability time also increases, but keeps performance in a reasonable time, i.e. not more than half a second.

Discussion

The focus of this work is the formal interpretation of biological database content with the goal to enable content retrieval using powerful ontology-based queries. We demonstrated the ability of OWL ontologies using description logics reasoning to retrieve statements of interest from biological databases.

We implemented and tested a use case based on database extracts describing amino acid metabolism. The interpretation is based on a specific ontological interpretation of representing biological database content: From each database record numerous defined OWL classes are generated, covering proteins, processes, molecular activities and dysfunctional phenotypes. Each class is assumed to contain at least one individual entity, corresponding to the outcome of a specific biological experiment described in the database. These individuals, however, are not explicitly represented in the ontology, as little as the data individuals in the database, which denote the domain entities. Thus, all content is exclusively represented in a Tbox, which warrants finite and complete reasoning, in contrast to the costs of rich TBoxes together with populated ABoxes [34].

The output, an "ontologized" database representation, enables queries on the experimental results summarized in database records. The fact that typical queries often target possibilities "*Are members of the class A able to do B*?" is addressed by two mechanisms. Firstly, by the inclusion of dispositions and functions as first-class entities in our ontology (as provided by the upper-level ontology BTL2), and secondly by the above mentioned addition of specific subclasses, assuming that these subclasses are populated. An example: the fact that our generation approach creates the class *Cellular nitrogen compound metabolic process in Homo sapiens with Betaine homocysteine S methyltransferase 1 and Homocysteine*, which is created as a defined subclass of the GO class 0034641 (*Cellular nitrogen compound metabolic process*), means that this process has been described at least once for Human with Homocystein and the related enzyme. In this sense, the question "Are members of the class A able to do B?" boils down to the question "Does A have a subclass A' all members of which actually do B?"

Interpreting database content under an ontological perspective has been a topic of research interest [4, 35, 36]. The DL-Learner system [4] is grounded on the requirement for schema acquisition methods, addressed by class learning techniques. DL-Learner is designed to find logical explanations for individuals. It is limited by the fact that positive and negative examples must be provided, and individuals must be included directly in the ontology. In the biomedical domain, [35] highlighted the importance of data interpretation from biomedical databases. Interpretation strategies are embedded in KEGG by a set of mapping operations among internal modules, allowing the identification of organism-specific pathways. QueryGen [37] presented a system that, given user keywords, proposes formal queries to retrieve data from repositories.

The usage of DL reasoning may be considered a costly approach. However, our experiments demonstrate a reasonable performance in medium-sized scenarios. New DL calculus methods are under development, which may decrease reasoning time [38]. An improvement of our work could be the use of more expressive power to retrieve generalizable content by means of DL Query. Retrieval using OBDA-based approaches, like SPARQL endpoints fairly support reasoning that goes beyond what is available in current relational queries [39]. Our approach allows evaluating databases from the ontological level, e.g. computing class-subclass relations, consistency checking and subsumption. This reduces the need to manually filter/interpret data, without compromising the capability to be queried with SPARQL endpoints.

For instance, to retrieve a protein that has methylation capability, with relational or SPARQL (without ontological treatment) queries, the user must create joins and filters to gather content from different sources. With DL query, the user only needs to define how the process behaves and leave the querying and computing to the machine.

A certain degree of engineering complexity, *viz.* the manual creation of ontology patterns is necessary when aiming at the production of a precise, ontology-based picture of what a database really represents, and how the informal database–ontology links are to be interpreted. Our solution has the advantage that it completely refrains from representing the denoting entities (i.e. the data instances), setting the full focus on the denotations, which – in the presented use case – are exclusively classes, which limits the task to pure TBox reasoning.

Many solutions for life science data processing have mainly focused on network, pathway, and sequence analysis, and more recently on the functional analysis of data. All these approaches have mainly been limited to syntax, and they refrain from bringing implicit domain assumptions into the scope of representation and reasoning. Our approach incorporates what is underneath the surface of data structures, representing it by basic ontology axioms, which are generated out of patterns that formalize how entities like processes, enzymes, molecules, phenotypes are stuck together in biological organisms and their substructures. Several limitations of our work need to be highlighted:

- Scaling problems could be demonstrated when increasing the size of data. Representing the whole content of databases in OWL is therefore not realistic. A mitigation strategy is to narrow down the database content of interest before creating the ontology. In our use case we did so by filtering database content by a single chemical entity.
- Computational issues can be addressed by decreasing the expressiveness of the representational language. The ontology patterns presented in this paper used disjunctions and value restrictions, both of which are not supported by the computationally ideal OWL EL profile. It has to be tested whether the substitution of axioms with disjunctions by simple subclass axioms and the substitution of value restrictions (used for the targets of dispositions and risks) by existential restrictions would yield the same reasoning results (even aware of the ontological improperness of the latter, cf. [40]).
- The proposed solution is highly prolific regarding the introduction of new entities as defined subclasses. This leads to a ratio of about 1:36 comparing database records with ontology axioms. This ratio increases with the number of fields considered, and with the average cardinality in multi valued fields. An alternative strategy, expressing everything as dispositions, would yield less axioms, but longer and less user friendly axioms, such as "Organism O has the disposition of performing a process P with the participants Q, R, and S in the cellular components C_1 and C_2", or "Organism O has the disposition of developing a pathological phenotype T as the effect dysfunction of P with...".
- Although DL Queries are relatively simple to create, users need to get familiar with description logics syntax and semantics, as well as they need to have basic notions of formal-ontological thinking. In addition, many queries require a

two-step processing, i.e. the list of classes as the result of the DL query proper has to be filtered afterwards.

- From an epistemic point of view, our approach makes a rough simplification by interpreting every database entry as statement of a scientific law. This ignores the mechanism of database population (e.g. by expert curation vs. text mining, with the latter being less trustworthy), fallibility and fraud in research, as well as the amount of experimental evidence to support the veracity of a given class-level assertion.

Further investigations are required to address the impact that database updates may generate. Modification in the schema level requires adaptations of the interpretation procedure, such as table joins or the obsolescence of certain content. In this case, the interpretation procedure must be recreated to include updates. We are currently developing a system to support the interpretation procedure, which minimizes the deep ontological understanding required by the current approach and addresses the inherent awareness of data interpretation. A fully automated approach may obfuscate some ontological assumptions that cannot be logically represented, and it may be epistemologically controversial.

Conclusion

Biological databases represent large amounts of experimental data and connect them with content from domain ontologies. Querying these databases generally follows the syntax of the underlying database scheme, in which ontologies represent little more than a simple domain vocabulary. Two important aspects are ignored, viz. (i) making the underlying ontological assumptions of the database scheme explicit (i.e. how the entities are related with each other) and (ii) exploiting the representational richness of the related ontologies. Both aspects are addressed by this work, in which the interpretation of biological database content is supported by an ontology grounding framework under the upper-level ontology BTL2, importing parts of the ontologies GO, PR, SNOMED CT and ChEBI.

The content of biological databases is automatically converted into OWL axioms, guided by a set of ontology patterns, which were manually crafted after a thorough scrutiny of the implicit meaning of the database structures of Uniprot, Ensembl, and NCBI Taxonomy. The ontology creation process obeyed the principles of philosophically founded and formally accurate ontology design and resulted in a large ontology, which uniquely represented TBox entities, i.e. no individuals.

This output ontology was then tested with six competency questions, formulated as description logics queries with subsequent filtering. The results corresponded to the expectations. Performance benchmarks were done with programmatically enlarged test ontologies, which showed limitations when TBoxes were increased by a factor of ¿ 30, partly due to hardware limitations.

The methodology appears suitable to be integrated within a larger query environment for biological knowledge. Performance bottlenecks may be addressed by content pre-filtering and the extraction of modules from the imported bio-ontologies.

Competing interests
The authors declare that they have no competing interests.

Author's contributions

All authors contributed equally to this work.

Acknowledgements

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.

Author details

[1]Centro de Informática, Universidade Federal de Pernambuco, Av. Jornalista Anibal Fernandes, 50.740-560 Recife, Brazil. [2]Computational Bioscience Research Center, King Abdullah University of Science and Technology, 23955-6900 Thuwal, Kingdom of Saudi Arabia. [3]Institut für Medizinische Informatik, Statistik und Dokumentation, Medzinische Universität Graz, Auenbruggerplatz 2, 8036 Graz, Austria.

References

1. UniProt Consortium: Activities at the universal protein resource (UniProt). Nucleic Acids Research **42**(D1), 191–198 (2013)
2. Cunningham, F., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., Gil, L., Giron, C.G., Gordon, L., Hourlier, T., Hunt, S.E., Janacek, S.H., Johnson, N., Juettemann, T., Kahari, A.K., Keenan, S., Martin, F.J., Maurel, T., McLaren, W., Murphy, D.N., Nag, R., Overduin, B., Parker, A., Patricio, M., Perry, E., Pignatelli, M., Riat, H.S., Sheppard, D., Taylor, K., Thormann, A., Vullo, A., Wilder, S.P., Zadissa, A., Aken, B.L., Birney, E., Harrow, J., Kinsella, R., Muffato, M., Ruffier, M., Searle, S.M.J., Spudich, G., Trevanion, S.J., Yates, A., Zerbino, D.R., Flicek, P.: Ensembl 2015. Nucleic Acids Research **43**(D1), 662–669 (2014)
3. Poggi, A., Lembo, D., Calvanese, D., Giacomo, G.D., Lenzerini, M., Rosati, R.: Linking data to ontologies. In: Journal on Data Semantics X vol. 4900, pp. 133–173. Springer, ??? (2008)
4. Lehmann, J.: DL-Learner: Learning Concepts in Description Logics. J. Mach. Learn. Res. **10**, 2639–2642 (2009)
5. Hustadt, U., Motik, B., Sattler, U.: Data Complexity of Reasoning in Very Expressive Description Logics. In: IJCAI-05, pp. 466–471. Morgan Kaufmann Publishers Inc., Edinburg, Scotland (2005)
6. Baader, F., McGuinness, D.L., Nardi, D., Patel-Schneider, P., Calvanese, D., McGuinness, D.L., Nardi, D., Patel-Schneider, P.: The Description Logics Handbook: Theory, Implementation, and Applications, 2nd edn., p. 601. Cambridge University Press, Cambridge (2007)
7. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.-A., Scheuermann, R.H., Shah, N., Whetzel, P.L., Lewis, S.: The OBO foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol **25**(11), 1251–1255 (2007)
8. Gruninger, M., Fox, M.S.: The Role of Competency Questions in Enterprise Engineering. In: IFIP WG5.7 Work., Trondheim, Norway, pp. 1–17 (1994)
9. NCBI Resource Coordinators: Database resources of the national center for biotechnology information. Nucleic Acids Research **43**(D1), 6–17 (2014)
10. IHTSDO: International Health Terminology Standards Development Organisation: SNOMED CT® Technical Imple- mentation Guide. http://www.snomed.org/tig.pdf (2015)
11. The Gene Ontology Consortium: Gene ontology consortium: going forward. Nucleic Acids Research **43**(D1), 1049–1056 (2014)
12. Hastings, J., de Matos, P., Dekker, A., Ennis, M., Harsha, B., Kale, N., Muthukrishnan, V., Owen, G., Turner, S., Williams, M., Steinbeck, C.: The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. Nucleic Acids Research **41**(D1), 456–463 (2012)
13. Natale, D.A., Arighi, C.N., Blake, J.A., Bult, C.J., Christie, K.R., Cowart, J., D'Eustachio, P., Diehl, A.D., Drabkin, H.J., Helfer, O., Huang, H., Masci, A.M., Ren, J., Roberts, N.V., Ross, K., Ruttenberg, A., Shamovsky, V., Smith, B., Yerramalla, M.S., Zhang, J., AlJanahi, A., Çelen, I., Gan, C., Lv, M., Schuster-Lezell, E., Wu, C.H.: Protein ontology: a controlled structured network of protein entities. Nucleic Acids Research **42**(D1), 415–421 (2013)
14. Schulz, S., Grewe, N., Röhl, J., Schober, D., Boeker, M., Jansen, L.: Guideline on Developing Good Ontologies in the Biomedical Domain with Description Logics. Technical Report December, Universität Rostock, Rostock (2012)
15. Santana, F., Schober, D., Medeiros, Z., Freitas, F., Schulz, S.: Ontology patterns for tabular representations of biomedical knowledge on neglected tropical diseases. Bioinformatics **27**(13), 349–356 (2011)
16. Martinez-Costa, C., Cornet, R., Karlsson, D., Schulz, S., Kalra, D.: Semantic enrichment of clinical models towards semantic interoperability. the heart failure summary use case. Journal of the American Medical Informatics Association, 565–576 (2015)
17. Spear, A.D.: Ontology for the Twenty First Century : An Introduction with Recommendations, p. 132. IFOMIS, Saarbrücken, Germany (2006). http://www.ifomis.org/bfo/manual
18. Schober, D., Smith, B., Lewis, S.E., Kusnierczyk, W., Lomax, J., Mungall, C., Taylor, C.F., Rocca-Serra, P., Sansone, S.-A.: Survey-based naming conventions for use in OBO foundry ontology development. BMC Bioinformatics **10**(1), 125 (2009)
19. Smith, B., Ceusters, W., Klagges, B., Köhler, J., Kumar, A., Lomax, J., Mungall, C., Neuhaus, F., Rector, A.L., Rosse, C.: Relations in biomedical ontologies. Genome Biol **6**(5), 46 (2005)
20. Schulz, S., Jansen, L.: Formal Ontologies in Biomedical Knowledge Representation. IMIA Yearb. 2013 **8**(Evidence-based Health Informatics), 132–146 (2013)
21. Rector, A.: Barriers, approaches and research priorities for integrating biomedical ontologies. Technical report, EU Semantic Health Support Action (2008). http://www.semantichealth.org/DELIVERABLES/SemanticHEALTH_D6_1.pdf

22. Schulz, S., Brochhausen, M., Hoehndorf, R.: Higgs Bosons, Mars Missions, and Unicorn Delusions: How to Deal with Terms of Dubious Reference in Scientific Ontologies. In: ICBO 2011, Buffalo, pp. 183–189 (2011)
23. Schulz, S., Stenzhorn, H., Boeker, M.: The ontology of biological taxa. Bioinformatics **24**(13), 313–321 (2008)
24. Tsarkov, D., Horrocks, I.: FaCT++ description logic reasoner: System description. In: Automated Reasoning vol. 4130, pp. 292–297. Springer, Seattle (2006)
25. W3C: OWL 2 Web Ontology Language Document Overview. http://www.w3.org/TR/owl2-overview/ (2012)
26. Horrocks, I., Kutz, O., Sattler, U.: The Even More Irresistible SROIQ. In: KR 2006, pp. 1–36 (2006)
27. Horridge, M., Patel-Schneider, P.F.: OWL 2 Web Ontology Language: Manchester Syntax. http://www.w3.org/TR/owl2-manchester-syntax/
28. Selhub, J.: HOMOCYSTEINE METABOLISM. Annu. Rev. Nutr. **19**(1), 217–246 (1999)
29. Muniz, M.T.C., Siqueira, E.R.F., Fonseca, R.A., D?Almeida, V., Hotta, J.K., dos Santos, J.E., do S.M. Cavalcanti, M., Sampaio, C.A.M.: Avaliação da relação entre o polimorfismo C677T no gene para MTHFR e a concentração plasmática de homocisteína na doença arterial coronariana. FapUNIFESP (SciELO) (2006)
30. Siqueira, E.R., Oliveira, C.P., Muniz, M.T., Silva, F., Pereira, L.M., Carrilho, F.J.: Methylenetetrahydrofolate reductase (MTHFR) c677T polymorphism and high plasma homocysteine in chronic hepatitis c (CHC) infected patients from the northeast of Brazil. Nutrition Journal **10**(1), 86 (2011)
31. Schulz, S., Boeker, M.: BioTopLite: An Upper Level Ontology for the Life Sciences. Evolution, Design and Application. In: Furbach, U., Staab, S. (eds.) Inform. 2013. IOS Press, Koblenz (2013)
32. Huntley, R.P., Sawford, T., Mutowo-Meullenet, P., Shypitsyna, A., Bonilla, C., Martin, M.J., O'Donovan, C.: The GOA database: Gene ontology annotation updates for 2015. Nucleic Acids Research **43**(D1), 1057–1063 (2014)
33. Jiang, G., Solbrig, H.R., Chalmers, R.J.G., Spackman, K., Rector, A.L., Chute, C.G.: A Case Study of ICD-11 Anatomy Value Set Extraction from SNOMED CT. In: ICBO 2011, Buffalo, pp. 133–138 (2011)
34. Motik, B., Sattler, U.: A comparison of reasoning techniques for querying large description logic ABoxes. In: Logic for Programming, Artificial Intelligence, and Reasoning, pp. 227–241. Springer, Phnom Penh, Cambodia (2006)
35. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., Tanabe, M.: KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Research **40**(D1), 109–114 (2011)
36. Carnielli, C.M., Winck, F.V., Leme, A.F.P.: Functional annotation and biological interpretation of proteomics data. Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics **1854**(1), 46–54 (2015)
37. Bobed, C., Mena, E.: QueryGen: Semantic interpretation of keyword queries over heterogeneous information systems. Information Sciences **329**, 412–433 (2016)
38. Freitas, F.: A Connection Method for the Description Logic ALC. In: Rosati, R., Rudolph, S., Zakharyaschev, M. (eds.) DL 2011. CEUR-WS, Barcelona (2011)
39. Angles, R., Gutierrez, C.: The expressive power of SPARQL. LNCS **5318**, 114–129 (2008)
40. Schulz, S., Martínez-Costa, C., Karlsson, D., Cornet, R., Brochhausen, M., Rector, A.: An Ontological Analysis of Reference in Health Record Statements. Front. Artif. Intell. Appl. **267**(FOIS'14), 289–302 (2014)

Figures

./PIC/process

Figure 1 Ontology grounding process.

/PIC/HierarquiaURI

Figure 2 Alignment of GO, ChEBI, SNOMED CT and PR under BTL2

Tables

Table 14 Uniprot and Ensembl table view

Entry	Protein	Organism	GO (bp)	GO (mf)	GO (cc)	Ensembl ID	Ensembl Phenoty
F1MEW4	CBS	*Bos taurus*	blood vessel remodeling; ...	cystathionine β-synthase activity ...	cytoplasm ...	ENSBTAT00000000184; ...	No phenotype ass
Q99707	MS	*Homo sapiens*	cobalamin metabolic process; ...	cobalamin binding; ...	cytoplasm ...	ENST00000366577; ENST00000535889	Neural tube Megaloblastic ...

The left six columns contain UniProt entries; the two right columns correspond to Ensembl in entried. GO (bp) , GO (mf) and GO (cc) r rows from UniProt that include annotations for GO classes '*Biological process*', '*Molecular function*' (understood as activity) and *component*'. IDs from UniProt and Ensembl are used for mapping purposes.

Table 15 Template table

#	P	O	Bp	Mf	C	Ph	M
k	P_k	O_k	$Bp_1 \dots n$	$Mf_1 \dots n$	$C_1 \dots n$	$Ph1 \dots n$	$M_1 \dots n$
...	...	...	...	...	...	...	...
l	P_l	O_l	$Bp_1 \dots n$	$Mf_1 \dots n$	$C_1 \dots n$	$Ph1 \dots n$	$M_1 \dots n$
...	...	...	...	...	...	...	...
m	P_m	O_m	$Bp_1 \dots n$	$Mf_1 \dots n$	$C_1 \dots n$	$Ph1 \dots n$	$M_1 \dots n$

The symbol # represents record IDs; *P* proteins; *G* genes; *O* organisms; *Bp* biological processes; *Mf* molecular function; *C* cellular co
Ph phenotype; and *M* the associate molecules.

Table 16 Survey of ontology sources, expressivity and reasoning performance in milliseconds

Ontology	Classes	Subcl. Axioms	Eq. Axioms	DL	Classification (h)	CQ1	CQ2	CQ3	CQ4	CQ
x1				$EL++$						
x3				$EL++$						
x10				$EL++$						
x30				$EL++$						
x100				$EL++$						
x300				$EL++$						
x1000				$EL++$						
x3000				$EL++$						
x10000	22,829,141	2,525,829	20,161,132	$EL++$						
Modularized	2,838	4,071	1,203	SRI	7,233,529ms	373ms	381ms	385ms	394ms	397n