Category: Databases and Ontologies

Ontology Patterns for Tabular Representations of Biomedical Knowledge on Neglected Tropical Diseases

Filipe Santana Da Silva[1*], Daniel Schober[2], Zulma Medeiros[3], Fred Freitas[1], Stefan Schulz[4]

[1] Centro de Informática, Federal University of Pernambuco (CIn/UFPE), Recife, Brazil

[2] Institute of Medical Biometry and Medical Informatics, University Medical Center Freiburg, Germany

[3] Aggeu Magalhães Research Center, Oswaldo Cruz Foundation, (CPqAM/Fiocruz), Recife, Brasil

[4] Institute of Medical Informatics, Statistics, and Documentation, Medical University of Graz, Austria

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Ontology-like domain knowledge is frequently published in a tabular format embedded in scientific publications. We explore the re-use of such tabular content in the process of building NTDO, an ontology of neglected tropical diseases, where the representation of the interdependencies between hosts, pathogens, and vectors plays a crucial role.

Results: As a proof of concept we analyzed a tabular compilation of knowledge about the pathogens, vectors and geographic locations involved in the transmission of neglected tropical diseases. After a thorough ontological analysis of the domain of interest we formulated a comprehensive design pattern, rooted in the biomedical domain upper level ontology BioTop. This pattern was implemented in a VBA script which takes cell contents of an Excel spreadsheet and transforms them into OWL-DL. After minor manual postprocessing the correctness and completeness of the ontology was tested using pre-formulated competence questions as DL queries. The expected results could be reproduced by the ontology. The proposed approach is recommended for optimizing the acquisition of domain knowledge from tabular representations.

Availability and Implementation: Domain examples, source code, and ontology are freely available on the web at http://purl.org/steschu/resources/ISMB2011

Keywords: Knowledge Acquisition, Description Logics, Neglected Tropical Diseases

Contact: filipe.santana.silva@gmail.com

1 INTRODUCTION

The high productivity of data in life sciences is reflected by a variety of representational formats. Large amounts of experimental data and research results are disseminated in large databases such as Uniprot[1], Ensembl[2] or ArrayExpress[3], whereas parts of the more aggregated and manually reworked information are published in the scientific literature where there are often systematized in the form of tables. In contradistinction to databases, where entries follow pre-defined database schemas, scientific authors are free in the composition of tables. Aside from tables that mainly contain numeric values, we frequently encounter tabular representations with symbolic entries, i.e. text strings, often from controlled vocabularies. In these tables, terms are displayed in a repetitive form for which interpretations are provided by the row and column headings, the legend, and the reference in the text. The proper interpretation by the reader requires a certain amount of background knowledge, and no semantic standard interpretation can be assumed for this kind of data.

Controlled terms in tabular representations of research results may denote individual entities, such as names of geographic entities, persons, or institutions, but also general terms, which denote classes or types of individual entities such as molecules, organisms, or diseases.

It is this kind of tabular information we will scrutinize under a viewpoint of Formal Ontology. Our hypothesis is that tables at least partly convey ontological content which can be diligently exploited in the construction process of formal ontologies. As both ontology building and maintenance are labor-intensive tasks, the semi-automated knowledge acquisition from tabular representations may constitute an interesting rationalization measure.

We are, however, equally aware that the symbolic content may frequently cross the boundaries of what is expressible by ontologies (Schulz, 2009), thus requiring other knowledge representation formalisms.

This paper is structured as follows. After this introduction, biomedical ontologies and their standards are shortly introduced, followed by the biomedical background of our case study, the field of neglected tropical diseases. In the third section the resources and methods for ontology construction and evaluation are presented; results are given in the fourth section. The paper concludes with a brief review of related work.

[1] http://www.uniprot.org/

[2] http://www.ensembl.org/index.html

[3] http://www.ebi.ac.uk/arrayexpress/

*To whom correspondence should be addressed.

2 BACKGROUND

We here introduce the basic concepts underlying our work, highlighting the syntax and semantics of biomedical ontologies and the details of the application area of neglected tropical diseases.

2.1 Biomedical Ontologies

The information explosion in biology and medicine has stimulated the proliferation of biomedical ontologies. More than 200 biomedical ontologies contained in the BioPortal ontology library (Noy *et al.*, 2009) specify the meaning of over 1.4 million terms. Some of these, i.e. the gene ontology, are used to integrate very large data bodies, illustrating that ontologies have become an indispensable resource in the management of research data. Ontological methods are also increasingly used in the development of medical terminology systems such as SNOMED CT (Donnelly, 2006) and a new generation of WHO classifications [4].

More and more biomedical ontologies today are based on, or at least alternatively disseminated in, description logics (Baader *et al.*, 2007), using the W3C recommended Web Ontology Language (OWL, W3C, 2010). In contrast to terminologies, such *formal* ontologies intend to describe (as much as possible) the consensus on the nature of entities in a given scientific domain, independently of linguistic or conceptual variation. Examples of statements belonging to this consensus core are indisputable true-isms like: all sandflies are arthropods, all cells contain membranes, all portions of saline contain sodium ions, and all Malaria events are caused by plasmodium organisms.

The construction of formal ontologies should obey principled criteria (Spear, 2006), e.g. as enforced by top-level ontologies, like DOLCE (Gangemi *et al.*, 2002), BFO (Grenon *et al.*, 2004), GOL (Heller and Herre, 2004), the OBO RO (Smith *et al.*, 2005) or BioTop (Beisswanger *et al.*, 2008) and good practice guidelines, e.g. as provided by the OBO Foundry (Smith *et al.*, 2007). Upper ontologies roughly coincide in their top level division between foundational disjoint categories such as e.g. material entities, processes, qualities, dispositions, information entities. Orthogonal to this distinction, there is also a coincidence in separating particular entities (e.g., “Brazil”) from the classes they are members of (e.g., Country). This distinction is crucial for properly using the above-mentioned representational formalisms.

The computable OWL DL subset (Horrocks *et al.*, 2003) constitutes a decidable fragment of first-order logic, which is supported by classifiers like Pellet (Sirin *et al.*, 2007) or HermiT (Motik *et al.*, 2009). These reasoners are able to determine whether the ontology contains contradictory assertions, whether classes in the ontology are satisfiable or for the checking of subclass relations. As DL is based on set-theory, a class like *Liver* has all individual livers as members, and a class like *BodilyOrgan* all individual bodily organs. As all individual livers are also members of *BodilyOrgan*, we can infer taxonomic subsumption: The class *Liver* forms a subclass of the class *BodilyOrgan* only if and only if all particular livers are also members of the class *BodilyOrgan*. In Manchester DL syntax (Horridge, 2009), this taxonomic subsumption is expressed by the subClassOf operator, e.g., *Liver* subClassOf *BodilyOrgan*.

Such simple class statements can be combined by different operators and quantifiers, e.g. 'and', 'or', the existential restriction 'some', and the value restriction ‘only’. To give an example, '*InflammatoryDisease* and **hasLocation** some *Liver*' denotes the class all members of which belong to *InflammatoryDisease* and are further related via **hasLocation** to some instance of the class *Liver*. This gives both necessary and sufficient conditions in order to fully define the class *Hepatitis*: *Hepatitis* equivalentTo *InflammatoryDisease* and **hasLocation** some *Liver*. The constructors introduced so far allow for automated classification and the computation of equivalence, but not for satisfiability checking. This is, however, important, wherever the validity of an assertion is to be assured and invalid assertions must be rejected. For instance, *Immaterial Object* subClassOf **hasPart** only *ImmaterialObject* restricts the value of the role **hasPart** by using the universal quantifier ‘only’. It should therefore reject any assertion that states that an immaterial object (e.g. a space) has a material object as part. However, a naïve use of this construct tends to fail. The reason of this is the so-called open world assumption: Unless otherwise stated, everything is possible. The following class *Strange Object* equivalentTo *Immaterial Object* and **hasPart** some *MaterialObject* would remain consistent as long as we do not explicitly state their disjointness, i.e. that there is nothing that can be both a material and an immaterial object: *Immaterial Object* subClassOf not *MaterialObject* .

We will use description logics in order to represent central notions of pathogen transmission for a family of diseases which will be described in the following section.

2.2 Application background

Neglected Tropical Diseases (NTDs) are infectious diseases which affect low-income populations in the developing world (Molyneux *et al.*, 2005; Hotez *et al.*, 2007; WHO, 2010). Although they are of major healthcare impact, NTDs are still seen as a rare event in developed countries (King and Bertino, 2008), compared to Malaria or HIV disease. The burden of the latter is about one order of magnitude higher, measured in DALY (Disability-Adjusted Life Years), a measure gauging the burden of a disease by indicating the time lived with disability and time lost due to premature mortality (Murray, 1994). Nevertheless the NTDs Lymphatic filariasis and Leishmaniasis are responsible for 5.78 million and 2.09 million DALY, respectively (WHO, 2004). Among the NTDs, the diseases transmitted by arthropod vectors (Dengue fever, Leishmaniasis, Chagas disease, American Trypanosomiasis, African Trypanosomiasis, Lymphatic Filariasis, Yellow Fever, among others) persist for a long time and can cause severe disability, disfigurement, and premature death (Beyrer *et al.*, 2007; Hotez *et al.*, 2007, Hotez *et al.*, 2009).

NTDs are increasingly targeted by public policies, and more and more clinical and epidemiological data is collected. In the standardization and management of health care information, ontologies can play an important role. Integrative access to health care data could produce new epidemiological insight and thus help in decision-making processes (Topalis, 2010). The identification of the occurrences of diseases in specific geographic locations is very important, as it comprises further information about the local distributions of the transmitting vectors as well. Consequently,

[4] http://www.who.int/classifications

ontologies should also manage incoming new data in an automatic way, and assist epidemiological data analysis.

In the next section we present materials and methods to construct our NTD ontology.

3 MATERIAL AND METHODS

3.1 Ontology Building

NTDO, the domain ontology for neglected tropical diseases was build and edited via Protégé v.4.1[5] together with the HermIT reasoner. Top level classes and foundational relations are taken from the domain upper-level ontology BioTop (Beisswanger *et al.*, 2007). We followed established ontology construction guidelines, such as normalization according to Rector (2003), suggesting, e.g., the untangling of graphs into disjoint orthogonal axes. The NTDO ontology engineering is done in a middle-out approach, as it was started by general classes that were generalized upward to the BioTop connection level, but also specialized downward to the required leaf node level dictated by the envisioned query granularity. Domain knowledge was harvested from indexed articles (publications of WHO and the Brazilian Health Ministry), as well as domain textbooks.

3.2 Sources for Knowledge Acquisition

The use case for automated knowledge acquisition is to represent the general transmission path of certain vector-borne diseases. The knowledge is extracted from a tabular representation published by (Sharma, 2008), part of which is depicted in Fig. 1.

Table 1: Vector borne disease matrix listing characteristic features: geographic locations (countries, regions) where the transmission takes place, specific manifestations and pathogens and vectors involved.

Geographic Location	Vector	Pathogen	Manifestation
Argentina	*Lutzomyia intermedia*	*Leishmania (V) braziliensis*	Cutaneous Leishmaniasis
Brazil	*Lutzomyia longipalpis*	*Leishmania (L) chagasi*	Visceral Leishmaniasis
South America	*Culex quinquefasciatus*	*Wuchereria bancrofti*	Lymphatic Filariasis
Mexico to Southern South America	*Rhodnius prolixus Triatoma infestans Triatoma dimidiata*	*Trypanosoma cruzi*	Chagas Disease
Africa	*Aedes aegypti*	*Yellow Fever Virus*	*Yellow Fever*

The table exemplifies the players in a typical disease transmission path such as described in the classical epidemiological triad (Fig. 1). Transmission process and disease manifestation are the result of an interaction between the infective **agent** (pathogen) and a susceptible **host** in a given **environment**. The host is any organism capable of being infected by the agent. **Vectors** are defined as merely transmitting the infectious agents, without being the intended host for the parasitic pathogen. Another role that participants of this interaction process can play is the role of pathogen reservoirs, e.g., animals, plant, soil or inanimate matter (Neves *et al.*, 2005).

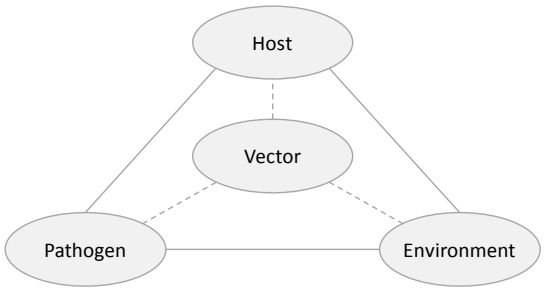

Figure 1: Epidemiological Triad. The main infection components are host, agent, and environment. The vector is frequently related to this path.

Vector-borne diseases may be associated with an ecological landscape profile, where host, vector, pathogen, and reservoir share the same geographic location, the habitat, over some time (Reisen, 2010). Hence, in order to apply effective preventive measures and drive health policy strategies data on the geographic locations, i.e. countries, regions, or micro-environments where the infection takes place, need to be described.

3.3 Knowledge Acquisition

The knowledge transfer from the tabular format to a fully fledged ontology, which supports concrete reasoning tasks, is a multi-step and often iterative process. The acquisition procedure can be described by the following workflow:

1. Ontological analysis of the tabular representation in the context of the text in which the table is embedded. First, it is decided which entities are classes and which are individuals. Then the appropriate classes or upper-level categories are chosen. Implicit references to entities which are not addressed in the table are identified, first of all the ontological category of the table itself. Finally, the relations and dependencies between the entities are identified. Hereby existential dependency needs to be verified (e.g., a disease is existentially dependent on the pathogen, but not vice versa). It should also be investigated whether the information represented is exhaustive, e.g. when we assert that a disease is only caused by the three pathogens cited in the table.
2. Formulation of a general design pattern (ODP, 2010). Using the target representation language, one or more prototypical axiomatic expressions are constructed using variables.
3. Implementation of the design pattern either by a design pattern processor and / or by a set of rules. The patterns

[5] http://protege.stanford.edu/

are interpreted and the desired ontology is constructed with the respective cell contents from the spreadsheet.

4. Manual revision of the automatically expanded ontology. This includes the manual restructuring of the generated ontology, by correcting or enriching it (e.g. reconstructing taxonomies) and finally the integration into the target ontology.

3.4 Evaluation methodology

The ontological scope is specified by gathering a set of competency questions (Gruninger and Fox, 1994) which we want the system to be able to answer and which will later be used to test the ontology for appropriate structure, coverage, expressivity and granularity. If the ontology does not appropriately answer the competency questions, a new iteration of the knowledge iteration cycle is initiated.

4 RESULTS

4.1 Ontological analysis of the table content

As specified in 3.3, the knowledge extraction from our input table (which corresponds to the pattern of Fig. 1) begins with an ontological analysis of the table structure and content. It contains four columns which represent distinct classes of entities: The leftmost column contains names of individual countries which are instances of the BioTop class *Geographical region*. The next column contains terms denoting the vectors, which are subclasses of the BioTop class *Arthropod*. The cells of the third column contain pathogens which are subclasses of the BioTop class *Protist*. The cells of the last column contain disease manifestations which are subclasses of *PathologicalProcess* in BioTop. The following additional observations are noteworthy:

1. There are cells with more than one term, denoting more than one class;
2. Not all cells of a row contain disjoint classes, so do we find *Leishmania sp*, which is a genus term and which consequently denotes a superclass of the species like *Leishmania donovani*;
3. The individuals in the first column are spatially related, e.g. the region *Mexico to Southern South American* spatially includes the region *Brazil*.

We now turn to the rows and analyze what they are describing. Our conclusion was that each row describes a different type of vector borne pathogen transmission pattern. More precisely, each row represents a distinct subclass of the class *Transmission pattern*, which is a subclass of biotop:*BiologicalProcess*. According to how we interpret the overall meaning of the table we can or cannot consider it as an exhaustive description.

We aim to link the classes by BioTop relations (OWL object properties), so we conclude that each instance of this kind of transmission process has a location (biotop:*hasLocus*)(column 1), has an agent (biotop:*hasAgent*), *viz*. the vector (column2), and a passive participant (biotop:*hasPatient*), the pathogen. When the process ends (and only then) the process is instantiated, and in this moment the pathogen is located in the host. The host seems to be the missing link in this table because it is restricted to *homo sapiens*. We therefore need to add the host organism. The relation between the pathogen and the host is, first of all, a locational relationship, simply because the pathogen is located with the host at the end of the process. Therefore we use, again, the relation biotop:*hasLocus*.

The relation between the pathogen (inside the host) and the disease manifestation is not so straightforward, because not every transmission process entails an infection of the host. The later has to happen after a pathogen transmission process from the vector into the host.

We therefore decided to distinguish between disposition and manifestation according to (Schulz, 2011), which are related by biotop:*hasRealization* and biotop: *realizationOf*, with the former as value restrictions and the latter as existential restrictions. What is typical for the diseases under scrutiny is that they only occur in organisms infected by the respective pathogens.

The geographical entities are included in our framework as reifications (Schulz and Hahn, 2001). For example the class *BrazilLocation* extends to all individual geographic places located in Brazil. Using this method we are able to emulate spatial inclusion by taxonomic subsumption (Hahn *et al*. 1999). Whereas Brazil is a part of South America, the class *BrazilLocation* is a subclass of *SouthAmericaLocation*. Thus we have a pure ontology without individuals.

4.2 Design pattern formalization

Table 2: General pattern of a vector borne disease matrix

Geographic Location	Arthropod (Vector)	Vertebrate (Host)	Protist (Pathogen)	Manifestation (Disease)
$G_{a1}\,G_{a2}\ldots G_{ak}$	$V_{a1}\,V_{a2}..V_{aj}$	$H_{a1}\,H_{a2}..H_{am}$	$P_{a1}\,P_{a2}..P_{al}$	$D_{a1}\,D_{a2}..D_{an}$
$G_{b1}\,G_{b2}\ldots G_{bk}$	$V_{b1}\,V_{b2}..V_{bj}$	$H_{b1}\,H_{b2}..H_{bm}$	$P_{b1}\,P_{b2}..P_{bl}$	$D_{b1}\,D_{b2}..D_{bn}$
…	…	…	…	…
$G_{n1}\,G_{n2}\ldots G_{nk}$	$V_{n1}\,V_{n2}..V_{nj}$	$H_{n1}\,H_{n2}..H_{nm}$	$P_{n1}\,P_{n2}..P_{nl}$	$D_{n1}\,D_{n2}..D_{nn}$
…	…	…	…	…
$G_{z1}\,G_{z2}\ldots G_{zk}$	$V_{z1}\,V_{z2}..V_{zj}$	$H_{z1}\,H_{z2}..H_{zm}$	$P_{z1}\,P_{z2}..P_{zl}$	$D_{z1}\,D_{z2}..D_{zn}$

The formalization of the ontology design pattern for the class of disease we are representing, the notation given in Table 2 is used.

Geographic entities

The following array of axioms demonstrates the meaning of the reified geographic locations. In fact we did not include these definitions because the required reasoning can be fully accomplished using the G_loc_i classes. There is no need of instances of the type $\mathbf{G_i}$ in the ontology.

G_loc_i equivalentTo *GeographicLocation* and **hasLocus** value $\mathbf{G_i}$

Pathogen Transfer

Each row *n* in the table is interpreted as a subclass of *PathogenTranferByVector*. We consider the table an exhaustive description of the domain, hence we define the umbrella class *PathogenTranferByVector* as the disjunction of its child classes.

Each child class $PathogenTranferByVector_n$ is fully defined and carries, additionally a set of value constraints.

PathogenTranferByVector equivalentTo
$PathogenTranferByVector_a$ or
$PathogenTranferByVector_b$ or … or
$PathogenTranferByVector_n$ or … or
$PathogenTranferByVector_z$

$PathogenTranferByVector_n$ equivalentTo *Transfer* and
(**hasAgent** some (V_{n1} or V_{n2} or … or V_{nj})) and
(**hasLocus** some (G_loc_{n1} or G_loc_{n2} or … or G_loc_{nk})) and
(**hasPatient** some ((P_{n2} or P_{n2} or … or P_{nl}) and
(**hasLocus** some (H_{n1} or H_{n2} or … or H_{nm}))))

$PathogenTranferByVector_n$ subClassOf *Transfer* and
(**hasAgent** only (V_{n1} or V_{n2} or … or V_{nj})) and
(**hasLocus** only ((not *GeographicLocation*)
or G_loc_{n1} or G_loc_{n2} or … or G_loc_{nk})) and
(**hasPatient** only ((P_{n2} or P_{n2} or … or P_{nl}) and
(**causes** only (D_{n1} or D_{n2} or … or D_{nr}))))

Dispositions and Manifestations

Dispositions express the fact that after the infection there is a different state of the organism, independent of whether the disease eventually breaks out.

D_disp_i equivalentTo *PathologicalDisposition* and
(**hasRealization** only D_i)
D_i equivalentTo *PathologicalProcess* and
(**realizationOf** some D_i)
D_i subClassOf **realizationOf** only D_i

Dependency of Diseases on Pathogens

At least the tropical diseases of interest in our study are, by definition, the consequences of infection with one or more pathogens of the kingdom *Protista*. Causality is complex and there may be other conditioning factors for the outbreak of the disease which may regarded as causal ones. This explains the addition of the expression (not *Protist*).

D_i subClassOf
causedBy only ((not *Protist*) or P_i or P_2 or … or P_l)

General Inclusion Axioms

These axioms state the equivalence of being the host of a pathogen and having the disposition of the respective disease. Of course this assertion may be done thoughtfully because in the case of numerous infectious diseases the host becomes resistant to the pathogen, so that the precense of the pathogen in the host no longer entails the disposition to the disease.

(H_1 or H_2 or … or H_m) and
locusOf some (P_1 or P_2 or … or P_l)
EquivalentTo
((**bearerOf** some D_disp_1) or
(**bearerOf** some D_disp_2) or …
(**bearerOf** some D_disp_i))

Disjointness Axioms:

As a default, all classes are disjoint, as the use cases require a maximally closed T-Box for the checking of constraints.

4.3 Ontology generation

The generation of the ontology was done in the following steps. First, the table content was copied to an Excel spreadsheet. Then the above patterns were implemented in a VBA macro which then generated the output ontology in OWL. We opted for this solution as existing tools and formalisms (e.g. O'Connor, 2010) could not be used or easily adapted due to the presence of multiple values per cell and the need for the attachment of suffixes (for geographic and disposition classes) to the original symbols.

4.4 Ontology postprocessing

The ontology was imported into Protégé and then analyzed for flat lists of siblings which need to be hierarchically restructured. This was necessary in three cases. For instance, the entry "unknown" in the vector table was substituted by the general class Arthropod, and the entry "Leishmania Sp" was put as a parent class to the classes representing the Leishmania species.

4.5 Ontology evaluation

The ontology was tested against a set of competency questions which were formulated as DL queries. For the sake of better understandability we are using a simple fancy ontology (Table 3) with easy understandable name, in contrast to the original names such as *Leishmania Amazonensis, Leishmania Chagasi, Leishmania Braziliensis, Leishmania Yucumansis, Leishmania Llanosmartini* (…) for pathogens and *Lutzomyia Flaviscutellata, Lutzomyia Longipalpis, Lutzomyia Carrerai Carrerai* (…).

The queries were formulated in OWL Manchester syntax and submitted to the DL query interface Protégé 4.1, using the HermiT reasoner.

Table 3: simple ontology for testing of reasoning

Geographic Location	Arthro-pod (Vector)	Vertebrate (Host)	Protist (Pathogen)	Manifesta-tion (Disease)
Brasil	SmallFly FatFly	Human	CrazyBug	GreenFever
Ceara Pernam-buco	FatFly BlackFly	Human	CrazyBug LazyBug	GreenFever BlueFever
Recife Olinda	SmallFly	Human	CrazyBug NastyBug	GreenFever BlueFever
South-America	TinyFly	Human	NastyBug	BlackFever GreenFever BlueFever

Competency Question 1:

"**What pathogen can be transmitted by a given vector in a geographic location?**"

This is formulated as a "can" question, and therefore the query needs to be inverted.

DL Query:

Protist and not (**patientIn** some
(*PathogenTransferByVector* and
hasLocus some (*CearaLocation* and
hasAgent some *BlackFly*)))

Result: *NastyBug*

Second query: *Protist* and not *NastyBug*

Result: *CrazyBug LazyBug*

which is the expected outcome.

Competency Question 2:

"**Can disease X be transmitted by vector Y in a given geographic location**?"

This is formulated as a yes/no question, which corresponds to satisfiability testing.

DL Query:

PathogenTransferByVector and
(**hasLocus** some *RecifeLocation*) and
(**hasPatient** some (*Protist* and
causes some *PinkFever*))

Result: Unsatisfiable, which is the expected result

If the geographic entities are re-arranged after the ontology generation in the sense that *RecifeLocation* subClassOf *BrazilLocation* subClassOf *SouthAmericaLocation*, the same query with *BlackFever* becomes satisfiable.

Competency Question 3:

"**What kind of disease can be transmitted in a given geographic location**?"

Again a "can" question, where a result can only be expected if negated.

DL Query:

PathologicalProcess and not
(**causedBy** some (*Protist* and
(**patientIn** some (*PathogenTransferByVector* and
hasLocus some *CearaLocation*))))

Result: *PinkFever*

Second Query: *PathologicalProcess* and not *PinkFever*

Final Result: *BlackFever BlueFever GreenFever*

which is the expected result

Competency Question 4:

"Could a vector directly cause a certain disease"

Formulated as a yes/no question (Satisfiability test)

DL Query:

Arthropod and **causes**
some (GreenFever or BlueFever or BlackFever)

Result: Unsatisfiable, according to the disease description

Our interpretation of the table is that the diseases can only be caused by protists, not by arthropods.

Competency Question 5:

"Is it possible to acquire disease X by vector transmission on region Y?"

Query Type: yes / no

DL Query:

PathogenTransferByVector and
(**hasPatient** some (*Protist* and (**causes** some *PinkFever*))) and (**hasLocus** some *RecifeLocation*)

Result: Unsatisfiable, according to the disease description

Competency Question 6:

"Which vectors can transmit a certain disease in region Y?"

This is formulated as a "can" question, and therefore the query needs to be inverted

Query Type: subclass

DL Query:

Arthropod and not (**agentIn** some
(*PathogenTransferByVector* and
(**hasLocus** some *OlindaLocation*) and
(**hasPatient** some (*Protist* and **causes** some *BlueFever*))))

Result: *FatFly*, *TinyFly*, *UglyFly*, *BlackFly*

Second query:

Arthropod and not (*FatFly* or *TinyFly* or *UglyFly* or *BlackFly*)

Result: *SmallFly*

which is the expected outcome.

Reasoning performance is a known issue when expressive description logics are used, such as in the case of NTDO with currently 121 classes, 12 object properties, 32 equivalence axioms, 136 subclass axioms, 9 disjoint axioms, and 28 hidden CGIs, using SI expressivity. The satisfiability testing (Competency question 2) took less than one second on an Intel Core i7 Processor 820QM, 8 GB memory and 64-bit Windows.

By artificially increasing the size by a factor of 10, the same test took about 14 minutes. This may be a major obstacle for the use of DL querying in expressive ontologies.

5 DISCUSSION

Our case study has shown that it is possible not only to represent moderately complex biological situations in standard description logics using tools and standards popularized by the Semantic Web community and to reason over them, but also the possibility of using a principled ontological foundation, which imposes strict categorial distinctions and provides a limited repository of object properties with strict domain and range constraints.

Our study supports the use of DL queries instead of SPARQL queries, the latter being much more popular. This is possible because of the absence of particular entities. However, both formalisms are complex and their appropriate use requires considerable training. An advantage of DL queries lies in the absence of variables and the natural way of reasoning over taxonomic hierarchies. Although the latter is of only marginal interest in our use case, it becomes highly relevant as soon as a similar ontology is linked, e.g. to a representation of biological taxa or a gazetteer with geographic names. For the representation of the latter our approach is of reifying expressions which include particulars, such as '**hasLocus** value **Brasil'.**

Tooling support for the construction of complex DL queries could be significantly improved, especially in cases where the underlying ontology provides rich domain and range constraints. Their use as a guidance to the query-builder is still a desideratum for Protégé or other ontology editors and workbenches.

The support of reasoning use cases which include both subclass retrieval and satisfiability testing was the main rationale for the described effort, which makes extensive use of OWL-DL constructors such as disjunction, negation, and value restrictions, as well as numerous fully defined classes. The drastic decrease in reasoning performance was therefore not surprising. Expressivity, together with a strong focus on DL reasoning is certainly one major distinctive criterion when comparing NTDO with OBO ontologies or the ontologies developed by Topalis *et al.* (2010) which represent a similar domain with a much higher coverage, but less expressive axiomatic content, with semantic annotation being their predominating raison d'être.

The automatic population of an ontology from a tabular format was described by O'Connor *et al.* (2010) who proposes a generic solution based on an extension of the OWL Manchester syntax which permits addressing Excel spreadsheet cells in descriptions of ontology design patterns.

Whereas these authors created their method in order to "ontologize" existing tabular information, by enabling the creation of classes and properties among other sophisticated possibilities, Bowers *et al.* (2010) describe a spreadsheet-to-OWL approach in which the spreadsheets are populated with the expressive goal of facilitating ontology construction. It provides also an easier language (than OWL) to describe DL contents. A similar approach is pursued by the quick term templates, described by Peters *et al.*, 2009. These solutions are certainly more amenable to biologists, ecologists, etc, than OWL. However they do not lend themselves to the reuse of legacy data such as extracted from existing tables.

An important limitation of the transfer of tabular information into an ontology is the type of content. Whereas numeric content can represented by OWL data properties (however only rudimentary supported by reasoners), probabilistic associations or default expressions extend the scope of what can be sensibly expressed in a DL ontology, as description logics is not an appropriate means to process this kind of knowledge (Schulz *et al.*, 2009)

6 CONCLUSION

In this study we investigated two questions. Firstly, how canonical domain knowledge about the transmission of rare tropical diseases can be expressed in a way that warrants reliable answers to previously formulated competency questions. Secondly, how legacy information contained in the tables of scientific papers can be transformed into a formal ontology which obeys the principles of philosophically founded ontology design.

We found satisfactory results for both questions, but also encountered serious limitations. Comprehensive domain knowledge can be represented in expressive description logics and can be queried by DL expressions. However, the scalability is limited due to the inherent computational complexity. Furthermore, the construction of such queries is currently not satisfactorily supported by user-friendly tools.

For the second question we successfully developed an export tool based on ontology design patterns which have to be individually crafted for each table. Here the limitation lies in the content of many tables, which do not contain ontological knowledge in a strict sense. In these cases other representational formalisms (e.g. for probabilistic knowledge) need to be employed.

7 ACKNOWLEDGEMENTS

This work was supported by the DFG grant JA 1904/2-1, SCHU 2515/1-1 GoodOD (Good Ontology Design), the BMBF-IB mobility project BRA 09/006.

8 REFERENCES

Baader, F. *et al.* (2007) *The Description Logic Handbook*. Theory, Implementation, and Applications. 2nd edition. Cambridge, U.K. Cambridge University Press.

Beisswanger, E. *et al.* (2008) BioTop: An Upper Domain Ontology for the Life Sciences - A Description of its Current Structure,

Contents, and Interfaces to OBO Ontologies. *Applied Ontology* 3(4), 205-212.

Beyrer, C. *et al.* (2007) Health and Human Rights 3: Neglected diseases , civil conflicts, and the right to health. *The Lancet* 370(9587), 619-27.

Bowers, S. *et al.* (2010) Owlifier: Creating OWL-DL ontologies from simple spreadsheet-based knowledge descriptions. *Ecol Inform* 5(1), 19-25.

Donnelly, K. (2006) SNOMED-CT: The advanced terminology and coding system for eHealth. *Stud Health Technol Inform* 121, 279-290.

Gangemi, A. *et al.* (2002) Sweetening ontologies with DOLCE. Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web. *Lect Notes Comput Sc* 2473/2002, 223-233.

Grenon, P. *et al.* (2004) Biodynamic ontology: applying BFO in the biomedical domain. *Stud Health Technol Inform.* 102, 20-38.

Gruninger, M. and Fox, M. (1994). The role of competency questions in enterprise engineering. In *IFIP WG 5.7, Workshop Benchmarking. Theory and Practice*, Trondheim/Norway.

Hahn, U. *et al.* (1999) Partonomic Reasoning as Taxonomic Reasoning. *In: AAAI'99- Proc. 16th National Conference on Artificial Intelligence*. July 18-22, Orlando, FL., 271-276.

Heller, B. and Herre, H. (2004) Ontological Categories in GOL. *Axiomathes* 14, 57–76.

Horridge, M. and Patel-Schneider, P. F. (2009) OWL 2 Web Ontology Language Manchester Syntax. Available at: http://www.w3.org/TR/owl2-manchester-syntax/. Acessed in 14th Jan, 2011.

Horrocks, I. *et al.* (2003) From SHIQ and RDF to OWL: the making of a Web Ontology Language. *J Web Semant* 1, 7-26.

Hotez, P. J. *et al.* (2009) Rescuing the bottom billion through control of neglected tropical diseases. *The Lancet.* 373(9674), 1570-1575.

Hotez, P. J. *et al.* (2007) Control of neglected tropical diseases. *New Engl J Med* 357(10), 1018-27.

King, C. H. and Bertino, A. (2008) Asymmetries of poverty: why global burden of disease valuations underestimate the burden of neglected tropical diseases. *Plos Neglect Trop D* 2(3), e209.

Molyneux, D. H. *et al.* (2005) "Rapid-impact interventions": how a policy of integrated control for Africa's neglected tropical diseases could benefit the poor. *PLoS Med* 2(11), e336.

Motik, B. *et al.* (2009) Hypertableau Reasoning for Description Logics. *J Artif Intell Res* 36, 165-228.

Murray, C. J. (1994) Quantifying the burden of disease: the technical basis for disability-adjusted life years. *Bull World Health Organ* 72(3), 429-45.

Neves, D. P. *et al.* (2005) *Parasitologia Humana.* 11st ed. São Paulo: Atheneu. 494p.

Noy, N. F. *et al.* (2009) Bioportal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Res* 37(Web Server issue), W170?173.

O'Connor, M. J.; *et al.* (2010) Mapping Master: A Flexible Approach for Mapping Spreadsheets to OWL. 9th International Semantic Web Conference (ISWC 2010), Shanghai, China, Springer-Verlag. Published in 2010.

ODP – Ontology Design Patterns (2010) Available at: http://ontologydesignpatterns.org.
Last accessed Jan 7, 2011.

Peters, B. *et al.* (2009) Overcoming the Ontology Enrichment Bottleneck with Quick Term Templates. *Nature Precedings*. Available at:
http://precedings.nature.com/documents/3970/version/1.
Last accessed Jan 7, 2011.

Rector, A.L. (2003) Modularisation of Domain Ontologies Implemented in Description Logics and related formalisms including OWL. In *Knowledge Capture* 2003, *ACM*, 121-128.

Reisen, W. K. (2010) Landscape epidemiology of vector-borne diseases. *Annu Rev Entomol* 55, 461-83.

Schulz, S. and Hahn, U. (2001) Parts, Locations, and Holes - Formal Reasoning about Anatomical Structures. *Lect Notes Comput Sc* 2101, 293-303

Schulz, S. *et al.* (2009) Strengths and limitations of formal ontologies in the biomedical domain. *RECIIS - Electronic Journal in Communication, Information and Innovation in Health*; 3 (1): 31-45: http://dx.doi.org/10.3395/reciis.v3i1.241en

Schulz, S. *et al.* (2011). Scalable representations of diseases in biomedical ontologies. *Journal of Biomedical Semantics*. Accepted for publication.

Sharma, U. and Singh, S. (2008) Insect vectors of Leishmania: distribution, physiology and their control. *J Vector Borne Dis* 45(4), 255-72.

Sirin, E. *et al.* (2007) Pellet: A practical OWL-DL reasoned. *J Web Semant* 5 (2): 51-53.

Smith, B., *et al.* (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnology* 25, 1251-1255.

Smith, B. *et al.* (2005) Relations in biomedical ontologies**.** *Genome Biology*, 6(5), R46.

Spear, A. D. Ontology for the Twenty First Century : An Introduction with Recommendations. (2006): 1-132. Avaiable at: http://www.ifomis.org/bfo/manual. Last accessed Jan 7, 2011.

Topalis, P. *et al.* (2010) A set of ontologies to drive tools for the control of vector-borne diseases. *J Biomed Inform.* 2010 Apr 2. PMID: 20363364

World Health Organization. (2009) Global programme to eliminate lymphatic filariasis. *Weekly epidemiological record*. 84(42), 437-444.

World Health Organization. (2004) The World Health Report: Changing History. Geneva: World Health Organization, 96p.

W3C. Working Group OWL 2. (2010) Web Ontology Language Document Overview. http://www.w3.org/TR/owl2-overview. Last accessed Jan 7, 2011.