BIOINFORMATICS

Vol. 27 ISMB 2011, pages i349–i356
doi:10.1093/bioinformatics/btr226

# Ontology patterns for tabular representations of biomedical knowledge on neglected tropical diseases

Filipe Santana[1,*], Daniel Schober[2], Zulma Medeiros[3,4], Fred Freitas[1] and Stefan Schulz[5]
[1]Informatics Center, Federal University of Pernambuco (CIn/UFPE), Recife, Brazil, [2]Institute of Medical Biometry and Medical Informatics (IMBI), University Medical Center Freiburg, Freiburg, Germany, [3]Departamento de Parasitologia, Aggeu Magalhães Research Center, Oswaldo Cruz Foundation, (CPqAM/Fiocruz), Recife, Brazil, [4]Pathology Department, Institute of Biological Sciences, University of Pernambuco, Recife, Brazil and [5]Institute of Medical Informatics, Statistics, and Documentation, Medical University of Graz, Graz, Austria

**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 (NTDs), 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 pathogens, vectors and geographic locations involved in the transmission of NTDs. 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 post-processing, the correctness and completeness of the ontology was tested using pre-formulated competence questions as description logics (DL) queries. The expected results could be reproduced by the ontology. The proposed approach is recommended for optimizing the acquisition of ontological domain knowledge from tabular representations.

**Availability and implementation:** Domain examples, source code and ontology are freely available on the web at http://www.cin.ufpe.br/~ntdo.

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

## 1 INTRODUCTION

The results of life sciences research are published in a variety of formats. Large-scale experimental data and research results are disseminated in databases such as Uniprot (http://www.uniprot.org/), Ensembl (http://www.ensembl.org/index.html) or ArrayExpress (http://www.ebi.ac.uk/arrayexpress/), whereas parts of the more aggregated and manually reworked information are published in the scientific literature 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 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 often requires considerable 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 individuals, such as names of geographic entities, persons or institutions, but also general terms, which denote classes or types of individuals such as molecules, organisms or diseases.

It is this kind of tabular information that we will scrutinize under a viewpoint of Formal Ontology. Our hypothesis is that many tables in scientific papers 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, 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 *et al.*, 2009), thus requiring other knowledge representation formalisms.

This article 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 (NTDs). In the third section, the resources and methods for table-guided ontology construction and evaluation are presented; results are given in the fourth section. The article concludes with a brief review of related work.

## 2 BACKGROUND

We here introduce the basic concepts underlying our work, introducing the syntax and semantics of biomedical ontologies and the details of the application area of NTDs.

### 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 (The Gene Ontology Consortium, 2000), 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 (http://www.who.int/classifications).

More and more biomedical ontologies today are based on, or at least alternatively disseminated in description logics (DL) (Baader

*To whom correspondence should be addressed.

*et al.*, 2007), using the World Wide Web Consortium (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 variation. Examples of statements belonging to this consensus core are indisputable axiomatic truisms 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) and good practice guidelines (Smith *et al.*, 2007), e.g. as enforced by top-level ontologies. Examples are the Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE; Gangemi *et al.*, 2002), the Basic Formal Ontology (BFO; Grenon *et al.*, 2004), GOL (Heller and Herre, 2004), the Relation Ontology, the Open Biomedical Ontologies Foundry (OBO RO; Smith *et al.*, 2005) and BioTop (Beisswanger *et al.*, 2008). Upper ontologies roughly coincide in their top-level division between foundational disjoint categories such as material entities, processes, qualities, dispositions and 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 and whether its classes are satisfiable. They also compute subclass relations. As DL is based on set theory, a class like *Appendix* has all individual appendices as members, and a class like *BodyStructure* all individual body structures. As all individual appendices are also members of *BodyStructure*, we can infer taxonomic subsumption: The class *Appendix* forms a subclass of the class *BodyStructure* if and only if all particular appendices are also members of the class *BodyStructure*. In Manchester DL syntax (Horridge and Patel-Schneider, 2009), this taxonomic subsumption is expressed by the subClassOf operator, e.g. *Appendix* subClassOf *BodyStructure*.

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'. For example, '*InflammatoryDisease* and **hasLocation** some *Appendix*' denotes the class all members of which belong to *InflammatoryDisease* and are further related via **hasLocation** to some member of the class *Appendix*. This gives both necessary and sufficient conditions in order to fully define the class '*Appendicitis*: *Appendicitis* equivalentTo *InflammatoryDisease* and **hasLocation** some *Appendix*'. The constructors introduced so far allow for automated classification and the computation of equivalence, but not for satisfiability checking, which is important wherever the validity of an assertion is to be assured and invalid assertions are supposed to 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 for 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 DL 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

NTDs are infectious diseases which affect mainly low-income populations in the developing world (Hotez *et al.*, 2007; Molyneux *et al.*, 2005; WHO, 2010). Although they are of major healthcare impact, NTDs are still seen as rare events in developed countries (King and Bertino, 2008), compared with 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, 2009).

NTDs are increasingly targeted by public policies, which has stimulated the collection of clinical and epidemiological data. In the standardization and management of healthcare information, ontologies can play an important role. Integrative access to healthcare data could produce new epidemiological insight and thus help in decision-making processes (Topalis *et al.*, 2011). 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, which helps to plan counter-measures to fight the disease. Consequently, ontologies should manage incoming new data in an automatic way, and assist epidemiological data analysis.

In the next section, we present materials and methods to construct an ontology for NTDs.

## 3 MATERIAL AND METHODS

### 3.1 Ontology building

NTDO, the domain ontology for NTDs was build and edited via Protégé v.4 (http://protege.stanford.edu/), using the embedded HermIT reasoner (Motik *et al.*, 2009) for auto-classification. The NTDO is an ontological representation for NTD (encompassing diseases, epidemiology and geographic distribution—for additional information see http://www.ntdo.ufpe.br/∼ntdo), for which top-level classes and foundational relations were taken from the domain upper-level ontology BioTop (http://purl.org/biotop) (Beisswanger *et al.*, 2008).

We followed established ontology construction guidelines, such as normalization according to Rector (2003), who suggested the untangling of asserted graphs into disjoint orthogonal axes. NTDO engineering is done in a middle-out approach, as it was started by general classes (taken from universal terms frequently found in table headers and domain database schemas), which were generalized upward to the BioTop connection level,

**Table 1.** Vector borne disease matrix listing characteristic features

| 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 |

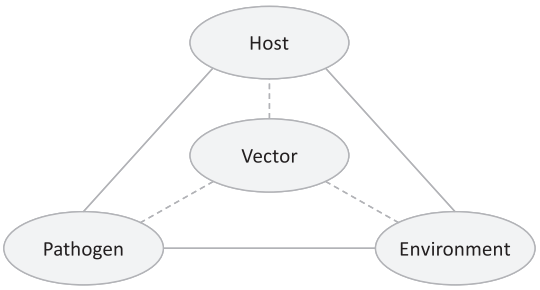


**Fig. 1.** Epidemiological triad. The main infection components are host, agent and environment. The vector is frequently related to all components making it a hub node in the transmission network, and hence a good target for infection control approaches.

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 World Health Organization (WHO) and the Brazilian Ministry of Health], as well as domain textbooks.

### 3.2 Sources for knowledge acquisition

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

The table exemplifies the main 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 organisms 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).

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 to drive health policy strategies relevant geographic locations, such as 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 or 'abstracted' form the individuals. Implicit references to entities which are not addressed in the table are identified, first of all the ontological category represented by 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 three specific pathogens.

(2) Formulation of a general design pattern (ODP, 2011). Using the target representation language, one or more prototypical axiomatic expressions are constructed.

(3) Implementation of the design pattern either manually, or semi-automatically by a design pattern processor and/or by a set of rules, where given patterns are interpreted and the desired ontology is constructed harvesting 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 Section 3.3, the knowledge extraction from an input table (in our case corresponding to the pattern of Fig. 1) begins with an ontological analysis of the four table 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 refer to pathogens, which are subclasses of the BioTop class *Protist*. The cells of the last column contain names of 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 entity.

(2) Not all cells in a column contain disjoint classes; so do we find *Leishmania sp*, which is a genus term denoting a superclass of species classes 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 then link the pattern classes by re-using BioTop relations (OWL object properties): each instance of transmission process has a location (biotop:**hasLocus**) (column 1), has an agent (biotop:**hasAgent**), *viz.* the vector (Column 2), 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 a class to represent the host organism. The relation between the pathogen and the host is, first of all, a locational relationship, simply because the pathogen is located within the host at the end of the disease transmission 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 latter can happen (not an existential relationship) after a pathogen transmission process from the vector into the host.

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

The geographic 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 (Schulz and Hahn, 2005). Whereas the individual 'Brazil' is **part_of** the individual 'South America', the class *BrazilLocation* is a subclass of *SouthAmericaLocation*. Thus, we have modeled all components needed for DL querying on our table derived subject domain on the class rather than the individual level.

## 4.2 Design pattern formalization

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

*4.2.1 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 G$_i$in the ontology.

$G\_loc_i$ equivalentTo
*GeographicLocation* and **hasLocus** value $\mathbf{G}_i$

**Table 2.** General pattern of a vector borne disease matrix

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

*4.2.2 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 additionally carries 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 $H_{nm}$))))

Note that only existential quantifications are used here.

*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 $D_{nr}$))))

Note that only universal quantifications (value restrictions) are used here.

Both definition parts are required to logically define the class in a ways that a DL reasoner can exploit it for question answering.

*4.2.3 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 $D\_disp_i$)
$D_i$ subClassOf **realizationOf** only $D\_disp_i$

*4.2.4 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*.

However, causality is complex and there may be other conditioning factors for the outbreak of the disease which may be regarded as causal ones.

$D_i$ subClassOf
**causedBy** only ($P_i$ or $P_2$ or … or $P_l$)

*4.2.5 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 numerous infectious diseases the host becomes resistant to the pathogen, so that the presence 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 $D\_disp_i$ ))

*4.2.6 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 Script-based ontology generation

The generation of the ontology was done in the following steps. First, the table content was copied to a Microsoft Excel spreadsheet. Then the above patterns were implemented in a Visual Basic for Application (VBA) macro, which takes the spreadsheet content and generates an output ontology in OWL-DL. We opted for this solution as existing tools and formalisms (e.g. O'Connor *et al*., 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 post-processing

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 added to the classes representing the *Leishmania* species, which is not described yet.

## 4.5 Ontology evaluation

The ontology was tested against a set of competency questions which were formulated as DL queries. As an example and for the sake of better understandability, a set of leishmaniasis data (including vector, pathogen and manifestation), as seen in Sharma and Singh (2008), were used (Table 3). To render class names shorter and more understandable, we abbreviate the species full name, e.g. *Lu.Longipalpis* for *Lutzomyia longipalpis*, and *L.chagasi* for *Leishmania chagasi*, and so on. For the diseases, the acronyms VL, ADCL, ML and CL represent visceral leishmaniasis, acute diffuse cutaneous leishmaniasis, mucocutaneous leishmaniasis and cutaneous leishmaniasis, respectively.

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

**Table 3.** Simple ontology for reasoning testing

| Geographic location | Arthropod (Vector) | Vertebrate (Host) | Protist (Pathogen) | Manifestation (Disease) |
|---|---|---|---|---|
| Guadeloupe | *Lu.longipalpis* | Human | *L.chagasi* | VL |
| Mexico | *Lu.longipalpis* | Human | *L.chagasi* | VL |
| | *Lu.olmeca olmeca* | | *L.mexicana* | CL |
| | | | *L.sp* | ADCL |
| Paraguay | *Lu.flaviscutellata* | Human | *L.amazonensis* | CL |
| | *L.Longipalpis* | | *L.chagasi* | ADCL |
| Peru | *Lu.whitmani* | Human | *L.braziliensis* | CL |
| | *Lu.peruensis* | | *L.peruviana* | ML |
| | *Lu.verrucarum* | | | |

***Competency Question 1:***
**What pathogen can be transmitted by a given vector in a geographic location?'**
In order to reflect the 'can' part (a possibility) of the question, we need to 'invert' the query by using negation.
DL Query:
*Protist* and not (**patientIn** some
(*PathogenTransferByVector* and
**hasLocus** some (*GuadeloupeLocation* and
**hasAgent** only *LutzomyiaLongipalpis*)))
Result:
*LeishmaniaAmazonensis*, *LeishmaniaBraziliensis*, *LeishmaniaMexicana*, *LeishmaniaPeruviana*, *LeishmaniaSp*.

Second query:
*Protist* and not (*LeishmaniaAmazonensis* or *LeishmaniaBraziliensis* or *LeishmaniaMexicana* or *LeishmaniaPeruviana* or *LeishmaniaSp*)

Result: *LeishmaniaChagasi*,
which is the expected outcome.

***Competency Question 2:***
'**Can disease X be transmitted by vector Y in a given geographic location Z**?'
Query Type: yes / no (Satisfiability test)
DL Query:
*PathogenTransferByVector* and
(**hasLocus** some *MexicoLocation*) and
(**hasPatient** some (*Protist and*
**causes** some *MucocutaneousLeishmaniasis*))

Result: Unsatisfiable, which is the expected result
According to the table, protists in Mexico do not cause ML. It shows us that the reasoner inferred what the logical definition entails.

***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 *PeruLocation*))))
Result: *AcuteDiffuseCutaneousLeishmaniasis*, *VisceralLeishmaniasis*

Second Query:
*PathologicalProcess* and not
(*AcuteDiffuseCutaneousLeishmaniasis*
*or VisceralLeishmaniasis*)
Final Result: *CutaneousLeishmaniasis*, *MucocutaneousLeishmaniasis*,
which is the expected result

***Competency Question 4:***
**'Is it possible to acquire disease X by vector transmission on region Y?'**
Query Type: yes / no (Satisfiability test)
DL Query:
*PathogenTransferByVector* and
(**hasPatient** some (*Protist* and
(**causes** some *MucocutaneousLeishmaniasis*)))
and (**hasLocus** some *MexicoLocation*)
Result: Unsatisfiable, according to the disease description.

***Competency Question 5:***
**'Which vectors can transmit a certain disease in a 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 *ParaguayLocation*) and
(**hasPatient** some (*Protist* and **causes** some
(*CutaneousLeishmaniasis* or
*AcuteDiffuseCutaneousLeishmaniasis*)))))
Result: *LutzomyiaOlmecaOlmeca*, *LutzomyiaPeruensis*, *LutzomyiaVerrucarum*, *LutzomyiaWhitmani*

Second query:
*Arthropod* and not (*LutzomyiaOlmecaOlmeca* or
*LutzomyiaPeruensis* or *LutzomyiaVerrucarum* or
*LutzomyiaWhitmani*)
Result: *LutzomyiaFlaviscutellata*, *LutzomyiaLongipalpis*
which is the expected outcome.

***Competency Question 6:***
**'Could a vector directly cause a certain disease'**
Query Type: yes/no (Satisfiability test)
DL Query:
*Arthropod* and **causes**
some (*AcuteDiffuseCutaneousLeishmaniasis* or
*VisceralLeishmaniasis* or *CutaneousLeishmaniasis* or
*MucocutaneousLeishmaniasis*)
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.

### 4.6 Performance issues

Reasoning performance is a known issue when expressive DL (using disjoints, inverses and negations) are used, such as in the case of NTDO with currently 154 classes, 28 equivalence axioms, 186 subclass axioms, 22 disjoint axioms and 25 hidden general classes inclusion (GCI), using *SI* (ALC) with transitive and inverse properties) expressivity (NTDO reuses BioTop object properties). The satisfiability testing (e.g. 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 min. This may be a major obstacle for the use of DL querying in expressive larger ontologies.

## 5 DISCUSSION

Our case study has shown that it is possible to represent moderately complex biological situations in standard DL using tools and standards recommended by the Semantic Web community. We proved the ability to reason over such DL models and illustrate the dependency of this approach on principled ontological foundations imposing strict categorial distinctions. Our formal patterns profited in particular from the re-use of a rigid set of object properties found in the upper-level ontology BioTop, which had their domains and ranges strictly constrained.

Our study shows the usefulness and expressive superiority of the use of true DL queries instead of simple resource description framework (RDF)-based SPARQL queries, the latter being much more popular and the former unfortunately still not finalized in a W3C recommendation (with SPARQL-DL being under development). However, both query syntaxes are complex and their appropriate use requires considerable training. An advantage of DL queries lies in the absence of free 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, we recommend reifying expressions which include particulars, such as '**hasLocus** value **Brazil**'. As a result, all reasoning can be performed on a TBox level, i.e. without explicit reference to particulars.

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 large and complex full-class definitions. 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.* (2011) 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 DL is not an appropriate means to process this kind of knowledge (Schulz *et al.*, 2009).

## 6 CONCLUSION

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

We found satisfactory results for both questions, but also encountered serious performance limitations. Comprehensive domain knowledge can be represented in expressive DL and can be queried by DL expressions and a reasoner. 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. Constraining the users by making them choose an optimized OWL 2 profile will either lead to constraints in modeling expressivity (OWL 2 QL) or to performance problems in larger ontologies (OWL 2 EL).

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, which clearly lie outside the realm of ontology and DL.

*Funding*: Deutsche Forschungsgemeinschaft (DFG) grant JA 1904/2-1, SCHU 2515/1-1 GoodOD (Good Ontology Design) and the Bundesministerium für Bildung und Forschung (BMBF)-IB mobility project BRA 09/006.

*Conflicts of Interest*: none declared.

## REFERENCES

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

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. *Appl. Ontol.*, **3**, 205–212.

Beyrer,C. *et al.* (2007) Health and human rights 3: neglected diseases , civil conflicts, and the right to health. *Lancet*, **370**, 619–627.

Bowers,S. *et al.* (2010) Owlifier: creating OWL-DL ontologies from simple spreadsheet-based knowledge descriptions. *Proc. Ecol. Inform.*, **5**, 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. Sci.*, **2473**, 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.

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/ (last accessed date March 28, 2011).

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

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

Hotez,P.J. *et al.* (2007) Control of neglected tropical diseases. *N. Engl. J. Med.*, **357**, 1018–1027.

King,C.H. and Bertino,A. (2008) Asymmetries of poverty: why global burden of disease valuations underestimate the burden of neglected tropical diseases. *Plos Negl. Tropic. Dis.*, **2**, 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**, 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**, 429–445.

Neves,D.P. *et al.* (2005) *Parasitologia Humana,* 11st edn.Atheneu, São Paulo.

Noy,N.F. *et al.* (2009) Bioportal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Res.*, **37**, W170–W173.

O'Connor,M.J. *et al.* (2010) Mapping master: a flexible approach for mapping spreadsheets to OWL. In *Proceedings of International Semantic Web Conference (2)'2010*. Springer, Berlin, Heidelberg, pp. 194–208.

ODP – Ontology Design Patterns (2010) Available at: http://ontologydesignpatterns.org (last accessed date March 28, 2011).

Peters,B. *et al.* (2009) Overcoming the ontology enrichment bottleneck with quick term templates. *Nat. Precedings*. Available at: http://recedings.nature.com/documents/3970/version/1 (last accessed date March 28, 2011).

Rector,A.L. (2003) Modularisation of domain ontologies implemented in description logics and related formalisms including OWL. In *Proceedings of the international conference on Knowledge capture - K-CAP'03*. ACM Press, New York, USA, p. 121.

Reisen,W.K. (2010) Landscape epidemiology of vector-borne diseases. *Ann. Rev. Entomol.*, **55**, 461–483.

Schulz,S. and Hahn,U. (2005) Part-whole representation and reasoning in formal biomedical ontologies. *Artif. Intell. Med.*, **34**, 179–200.

Schulz,S. and Hahn,U. (2001) Parts, locations, and holes - formal reasoning about anatomical structures. *Lect. Notes Comput. Sci.*, **2101**, 293–303.

Schulz,S. *et al.* (2009) Strengths and limitations of formal ontologies in the biomedical domain. *RECIIS – Elect. J. Commun. Inform. Innovat. Health*, 31–45.

Schulz,S. *et al.* (2011) Scalable representations of diseases in biomedical ontologies. *J. Biomed. Seman.*, accepted for publication.

Sharma,U. and Singh,S. (2008) Insect vectors of Leishmania: distribution, physiology and their control. *J. Vector Borne Dis.*, **45**, 255–272.

Sirin,E. *et al.* (2007) Pellet: a practical OWL-DL reasoned. *J. Web Seman.*, **5**, 51–53.

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

Smith,B. *et al.* (2005) Relations in biomedical ontologies. *Genome Biol.*, **6**, R46.
Spear,A.D. (2006) Ontology for the twenty first century: an introduction with recommendations, 1–132. Available at: http://www.ifomis.org/bfo/manual (last accessed date March 28, 2011).
The Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.
Topalis,P. *et al.* (2011) A set of ontologies to drive tools for the control of vector-borne diseases. *J. Biomed. Inform.*, **44**, 42–47.
World Health Organization (2004) *The World Health Report: Changing History*. World Health Organization, Geneva.
World Health Organization (2009) Global programme to eliminate lymphatic filariasis. *Weekly Epidemiol. Record,* **84**, 437–444.
W3C. Working Group OWL 2 (2010) Web Ontology Language Document Overview. Available ta http://www.w3.org/TR/owl2-overview (last accessed date March 28, 2011).