

Formal ontologies in biomedical knowledge representation

Stefan Schulz

Institute for Medical Informatics, Statistics and Documentation,
Medical University of Graz, Austria
Institute of Medical Biometry and Medical Informatics
University Medical Center, Freiburg, Germany
stefan.schulz@medunigraz.at

Ludger Jansen

Institute of Philosophy, University of Rostock,
Rostock, Germany
ludger.jansen@uni-rostock.de

Abstract:

Objectives: Medical decision support and other intelligent applications in the life sciences depend on increasing amounts of digital information. Knowledge bases as well as formal ontologies are being used to organise biomedical knowledge and data. However, these two kinds of artefacts are not always clearly distinguished. Whereas the popular RDF(S) standard provides an intuitive triple-based representation, it is semantically weak. Description logics based ontology languages like OWL-DL carry a clear-cut semantics, but they are computationally expensive, and they are often misinterpreted to encode all kinds of statements, including those which are not ontological.

Method: We distinguish four kinds of statements needed to comprehensively represent domain knowledge: universal statements, terminological statements, statements about particulars and contingent statements. We argue that the task of formal ontologies is solely the representations of universal statements, while the non-ontological kinds of statements can nevertheless be connected with ontological representations. To illustrate these four types of representations, we use a running example from parasitology.

Results: We finally formulate recommendations for semantically adequate ontologies that can efficiently be used as a stable framework for more context-dependent biomedical knowledge representation and reasoning applications like clinical decision support systems.

Keywords: Biomedical Ontology, Description Logic, Formal Ontology, Knowledge Representation, Decision support systems

Content

Formal ontologies in biomedical knowledge representation	1
Introduction.....	3
Medical decision support depends on knowledge representation	3
Ontology: more than a buzzword?	4
Biomedical ontologies gradually make their way	6
Four kinds of statements: A case study	8
Why RDF-triples are too simple to be helpful.....	10
Formal ontologies - Representing what is universally true.....	11
Web Ontology Language (OWL) and Description Logics (DL)	11
Subclasses and class members.....	13
Logical constructors.....	13
The representation of risks and dispositions	14
The representation of non-ontological statements	16
Representing the meanings of natural language terms	16
Representation of individuals.....	17
The really interesting knowledge is non-ontological	19
Default assumptions	20
Frequencies.....	23
Recommendations.....	24
Conclusion.....	27
Acknowledgements	28
References	29

Introduction

Medical decision support depends on knowledge representation

Health care processes are getting more complex. The cognitive abilities of individual caregivers and teams that manage acute and chronic disorders are increasingly challenged. On a population scale, well-founded decisions need to be made to warrant a maximum of health at acceptable costs.

Computerized decision support is, therefore, no longer a convenient add-on, but an urgent necessity. Decision support systems (DSSs) should match evidence-based domain knowledge against clinical data in order to assist the choices humans have to make.

For a long time, decision support has been a major use case for knowledge-based systems in the field of medicine. However, the integration of intelligent systems into the care process has, by and large, not fulfilled the predictions of expert system enthusiasts, not only because of the well-known knowledge acquisition bottleneck, but also as a consequence of the lack of syntactic and semantic integration of DSSs into health information systems: users have not been convinced that redundant data entry pays off. Recent spectacular advances have not been reached by the traditional expert system architectures based upon formal representation and reasoning, but by a massive heuristic re-use of unstructured "big data" [001,002]. An example is the IBM Watson project, which processes texts written by humans for human use, achieving an impressive turn-out in automated question answering. This technique is currently also experimentally applied to medicine [003].

In light of these facts, do we need to revise the traditional approach to represent specific knowledge for machine use? Has not the body of facts and knowledge, constantly produced, maintained and updated by and for humans, accumulated enough to drive knowledge-based systems? Which is an adequate structure of this knowledge? Is human language sufficient for expressing all kinds of facts and knowledge?

Huge efforts are going into controlled biomedical vocabularies

Traditionally, knowledge has been encoded in natural language, be it in writing or in words spoken from teacher to learner. As a means of communications between humans, natural language is easily and naturally used by researchers and health professionals in their daily routine. Major contents of electronic health records are unstructured or semi-structured texts. On the downside, not only is there a plethora of natural languages, but natural language expressions also tend to be ambiguous and are highly dependent on the linguistic and situational context in which they are used. This fact has fuelled costly efforts to precisely define the meaning of domain terms, or better yet, to unambiguously characterize the objects these terms

denote. In biology and medicine, controlled vocabularies have been produced, maintained and enhanced to serve variegated purposes, such as reporting, billing, quality assurance, document retrieval and medical research. The problem is that due to the proliferation of data structures and vocabularies, the integration of data across system and institutional boundaries tends to be even less reliable than the communication of free text. The impressive growth of biomedical vocabularies in number and content can be observed every year by the UMLS Metathesaurus updates [004].

The heterogeneity of these vocabularies reflects the different backgrounds, tasks, needs and languages of user communities and gives rise to serious obstacles in data aggregation and interoperability. Probably the most important use cases are disease classification for health statistics and billing, with the International Classification of Diseases [005] (currently undergoing its eleventh revision) being its most relevant terminological resource. Reimbursement of health services is also a driving force for ICD-9/ICD-10 CM, which also provide codes for medical procedures. A recent shift from a disease point of view to a functional viewpoint (also for a long-term planning of health expenditures) has emphasised the need for specific codes as provided by the International Classification of Functioning, Disability and Health (ICF) [006]. Standardised classification of drugs and their ingredients has motivated the WHO Drug Dictionary [007], the ATC [008], RxNorm [009] and DM+D [010]. Standardisation of lab data is the use case for LOINC [011]. Cross-domain vocabularies are MeSH [012] for medical literature indexing and, as an increasingly important resource on a global scale, SNOMED CT [013] for clinical documentation.

Ontology: more than a buzzword?

For good or bad, more and more biomedical vocabularies are named "ontology". In medical informatics, this term had been occasionally used for semantic networks [014], for the GALEN concept representation system [015] and for domain representations in the context of medical language processing [016,017]. It finally became popular in biomedical research circles with the success of the Gene Ontology [018], and the practical impact of ontology-based terminology research and development has increasingly been recognised [019,020].

The pervasive use of the term "ontology" since the turn of the millennium has been repeatedly criticized. Seven different definitions were proposed in [021] and the continuing ambiguity of this word has been criticized in [022], as well as the fact that "ontology" evolved from a technical term restricted to academic circles to a buzzword [023]. Indeed, in the Artificial Intelligence community it is hardly ever used in its original sense, viz. as a name for the philosophical study of the most general kinds of beings and the relations between these [024], the study of the a priori nature of reality [025], of entities as they exist in reality [026] or simply of *what there is* [027,028]. Therefore, any use of this term must be preceded by an explanation

of its intended meaning. Today, the term "ontology" is used in various ways as a cursory glance at bibliographic databases and web-based resources reveals:

1. Ontology as knowledge representation;
2. Ontology as a representation of terms;
3. Ontology as a representation of concepts;
4. Ontology as a representation of real entities.

Let us have a look at these uses in turn. According to the first group of uses of the term "ontology", an ontology is a part of a knowledge base or a means to structure it [029]. This does, of course, state a use case of ontologies rather than explain what they are. Other authors say that an ontology is a formal knowledge representation model [030,031], or a document or file that contains a set of resources and relations between them, defined as a series of triples [032]. In this group also belongs the idea that by means of an ontology, knowledge can be captured and made available to both machines and humans [033]. This statement, however, is not intended to imply that ontologies and knowledge bases are the same kind of thing.

Few authors use the word "ontology" for a machine-readable representation of a domain's terminology and the relationships among the terms [034] rather follow Gruber's definition and see an ontology as an explicit representation of a conceptualization [035]. According to this point of view, an ontology defines the concepts, relationships, and other distinctions that are relevant for modelling a domain. The meanings (2) and (3) seem to be popular in circles that are mainly interested in thesaurus-like artefacts for query expansions or in controlled terms for semantic annotations, e.g. for texts. However, it appears to us that Gruber's definition of ontologies as conceptualizations is more often parroted than understood.

According to other authors, (formal) ontologies are 'theories that attempt to give precise mathematical formulations of the properties and relations of certain entities' [024]. Ontologies are, thus, 'representational artefacts whose representational units are intended to designate classes or types in reality and to relate them to each other' [036]. They are 'catalogues of the types of things that are assumed to exist' in a domain of interest [037] and consist of 'precise descriptive statements about some part of the world' [038].

There is a strong coalition adhering to (4), though there have been fierce debates between supporters of (3) and (4) [026,039]. With strong initiatives like the OBO Foundry [040] supporting (4), there may be a shift towards (4). This shift can be interpreted as a renaissance of the term's original philosophical use of this term; as a result, ontology as the study of the most general kinds of beings can be seen as such a representation of entities at the most general level. However, a much bigger role in propagating (4) was probably played by the W3C standardization efforts that popularized OWL (and, in consequence, description logics [041] as described in more detail below) as the standard ontology language. As much as the increasing adherence to meaning (4) should be welcome (for reasons that will be set forth in the following), it should not

be forgotten that a sloppy usage of the term "ontology" persists, especially in the commercial world. One has the impression that the term "ontology" is still shiny enough to sell better than the old-fashioned "thesaurus" or "vocabulary".

Biomedical ontologies gradually make their way

While there might be a trend in biomedical ontology towards understanding ontologies as representations of reality, this is not the only trend that can currently be observed. We identify the following trends:

- *Creation of repositories and other infrastructure.* The best starting point to get an up-to-date overview of biomedical ontologies (and terminologies) is BioPortal, maintained by the U.S. National Center of Biomedical Ontologies [042]. It testifies the recent tendency in the field of biomedical ontology to consolidate the ontology infrastructure and to collect open access libraries of biomedical ontologies. BioPortal offers browser and Web service access and enables community participation in content evaluation and mapping [043]. Other repositories include the Open Biological and Biomedical Ontologies (OBO) Foundry initiative [044] and repositories for ontology design patterns (ODPs), like Ontology Design Patterns [045] and the ODPs Public Catalog [046].
- *Development of tools and infrastructure.* Protégé [047] is, by far, the most popular ontology editor. It has evolved from an editor for frame-based knowledge representation systems to a development framework for OWL-based ontologies. It comes with several plugins, e.g. for reasoning and visualization. The OWL API is a JAVA programming interface for manipulating OWL ontologies [048]. By now, a wide variety of tools and web infrastructure for ontology construction has been developed; including OntoFox [049], a tool for importing classes from existing ontologies and evaluation tools like the Ontology Pitfall Scanner OoOPS [050]; and OntoCheck, a tool for checking class labels for compliance with naming conventions [051].
- *Establishments of ontology standards.* The Open Biological and Biomedical Ontologies (OBO) initiative originated as a collection of controlled vocabularies. In the beginning, OBO ontologies consisted of terms which were interconnected by binary relationships, such as 'is a' and 'part of'. Since then, OBO has been supplemented by a formal language, the OBO file format, which has grown in semantic complexity over time. Finally, the use of the description logics-based (DLs) Semantic Web standard OWL [052,038] was encouraged and tools for conversion between the OBO file format and OWL were proposed [053]. An evolving set of shared principles for the coordinated development of non-overlapping ontologies, the so-called OBO Foundry principles, were proposed [040] and the creation of defined classes, so called intra-ontology cross-products, was encouraged.

- *Formally rigid and philosophically founded biomedical ontologies.* While often starting from quite informal hierarchies, ontology development has tended to employ formally rigid languages and axiomatised relations. The OBO Foundry principles also include the rule to subsume any class under one class of the Basic Formal Ontology (BFO) [054] as the common top-level ontology of all OBO ontologies. By now, there are several distinct top-level ontologies, all of which have been developed based on philosophical foundations, including DOLCE [055] and the Sowa diamond [056]. In order to ensure the correctness of taxonomic hierarchies, the evaluation method OntoClean has been developed [057].
- *Collaborative development of modularized biomedical ontologies.* Developing an ontology is not a one-day job. It needs the collaboration of domain experts, computer scientists and often also philosophers. Due to the considerable costs of ontology development, the OBO Foundry has suggested a scheme of combining orthogonal ontology modules, each of which is collaboratively developed by a group of experts. Applied ontologists are now also organised in the International Association on Ontology and its Application (IAOA). The IAOA is publishing a journal, "Applied Ontology", organizing educational events and community efforts. Another very active virtual community is Ontolog [058], which organises, together with the IAOA and others, yearly "Ontology Summits".
- *Ontology-based development of a global terminological standard.* SNOMED CT (Systematized Nomenclature of Medicine - Clinical Terms), developed and maintained by IHTSDO (International Health Terminology Standards Development Organization) is currently the world's most comprehensive clinical healthcare terminology [013]. Its ongoing development is increasingly guided by ontological principles. As such, it can be regarded as a novel and unparalleled large-scale ontology engineering exercise. SNOMED CT grew out of medical nomenclatures and coding systems [059], mainly SNOMED International (via SNOMED RT) and the NHS Clinical Terms Version 3 (the former Read Codes). They had been built to provide semantic descriptors to annotate all aspects of the health care delivery process together with standardised medical terms in different languages. With the advent of SNOMED RT (and later SNOMED CT), logics entered the scene, thus adding a mathematically rigorous layer to the hitherto informal, close-to-human-language representation of medical terms. SNOMED currently includes more than 310,000 representational units called SNOMED CT concepts. SNOMED CT concepts are supplemented by approximately 1,350,000 axioms constituting subclass, equivalence or existential restrictions corresponding to the DL version EL++ [041].
- *Ontology-based development of WHO classifications.* A subset of SNOMED CT is planned to be used as an ontological foundation of the upcoming ICD 11 (International Classification of Diseases) development [060]. A similar approach is planned for other

WHO classifications like the ICF (International Classification of Functioning). These projects mark an important step towards a principle-based convergence of domain vocabularies.

- *The development of domain-specific upper-level ontologies.* The attempt to provide upper-level categories for the biomedical domain started with the UMLS semantic network [061]. The domain-independent upper-level ontology BFO [054] and the domain-specific relation ontology RO [062] were introduced to categorize and normalize OBO Foundry ontologies. BioTop [063] integrated upper-level classes with relations within one representational artefact with constraining axioms in OWL-DL together with mappings to several upper-level ontologies [064]. Ongoing developments in this field are a new version of BFO, which includes both classes and relations [065], a new release of BioTop with mappings to BFO2, GFO-BIO [066] and the Semanticscience Integrated Ontology (SIO) [067].

Four kinds of statements: A case study

Though formal ontology is a discipline neighbouring lexicology, terminology and knowledge representation, it must be clearly delineated from these disciplines: formal ontology is a distinct project. In order to argue for this, we will juxtapose formal ontology with different kinds of knowledge representations.

[075] distinguishes four types of statements that domain modellers may want to represent. The authors emphasised that computable, logic-based domain representation formalisms like DLs are only able to truthfully represent certain aspects of biomedical knowledge, whereas other types of assertions require different means of representation. An inadequate use of undifferentiated representation formalisms will lead to unwanted results. We follow scr2009 in distinguishing between four types of statements that domain modellers may want to represent:

1. Universal statements (which are universally true for all instances of a certain type);
2. Terminological statements (about linguistic items like terms);
3. Assertional statements (about individuals);
4. Contingent statements (ascribing predicates to a class which may or may not be true for all its members).

In this fourfold distinction, ontological statements in a strict sense are limited to (1). They describe commonalities of the fragment of the domain to be represented, whereas (2) describe natural language expressions used for talking about the domain of interest, such as those provided by vocabularies and thesauri. Statements of type (3) are generally outside the ontology, but they are typically the referents of data in databases and their common properties

are described by ontologies. Statements of type (4) correspond to most natural language predicates that refer to non-individuals, but lack the universal validity of (1).¹

In the following, we want to apply this fourfold distinction within a little case study based on information collected from [069,070,071,072]. In a nutshell, the example is about filariasis, an infection by parasitic worms, which affects significant portions of underserved populations in the tropics. The microscopic larvae are transmitted by insects. In their adult stage the worms destroy several body tissues. A severe late complication is elephantiasis in which the limbs swell and, over the years, acquire grotesque dimensions (Fig. 1).

The following assertions will be discussed in the remainder of the article:

- [a01] Nematodes are also known as roundworms.
- [a02] Filariae are nematodes.
- [a03] Filariasis is a parasitic disease.
- [a04] Filariasis is caused by filariae.
- [a05] Filariasis is lymphatic, subcutaneous or serous cavity filariasis.
- [a06] Elephantiasis is normally caused by lymphatic filariasis.
- [a07] The female worms release larvae known as microfilariae into the host's bloodstream.
- [a08] Patient X was diagnosed as being highly suspicious for lymphatic filariasis.
- [a09] Lymphatic filariasis has elephantiasis as typical late complication.
- [a10] Filariasis can be treated by albendazole.
- [a11] Nematode infections may cause eosinophilia.
- [a12] Filariasis is a neglected tropical disease.
- [a13] According to the WHO, 43% of Nigerians are at risk for lymphatic filariasis.

Any of our four classes of statements is exemplified by some sentence in this sample:

- Universal statements: [a02,a03,a04,a05]
- Terminological statements: [a01]
- Statements on individuals: [a08,a13]
- Contingent statements: [a06,a07,a09,a10,a11,a12]

¹ The distinction between (1) and (3) corresponds to the traditional distinction between "TBox" and "ABox" in description logics [068]. As much as the distinction as such is still fundamental for description logics theory and applications, the original naming is rather misleading: The TBox ("terminology box") is not about the terms proper, but about what the (general) terms denote; and the ABox ("assertional box") is not about assertions in general, but only about those restricted to individuals. It should also be noted that a knowledge base is not exhausted by TBox and ABox, as there are important statements (namely those that we call "contingent statements").

We will first introduce RDF triples as a syntactically standardised format in which the above statements can be expressed. This is then contrasted with the representation of these statements within a formal ontology. We will argue that only universal statements make up the proper content of ontologies. In the context of ontologies, terminological statements may be relevant as metadata, but they should, in principle, be dealt with by tools and formats created for thesauri and vocabularies. Finally, individual and contingent statements are best dealt with in separate data repositories which, however, should be structured by interlinking the data represented with appropriate ontologies.

Why RDF-triples are too simple to be helpful

A popular scheme for structuring predications like the above is the so-called Object-Attribute-Value (OAV) triple. As a form of expressing semantic networks, OAVs had already been used in early intelligent systems [073]. Today, they occupy a central role in the formalism of the Resource Description Format (RDF) in the context of the Semantic Web where they are known as Subject-Predicate-Object (SPO) triples [074]. We can easily convert the major part of our sample sentences into this format:

- [r01] <Nematode, synonym-of, Roundworm>
- [r02] <Nematode, rdfs:subClassOf, Worm>
- [r03] <Filariasis, rdfs:subClassOf, Parasitic disease>
- [r04] <Filariasis, caused-by, Filariae>
- [r06] <Elephantiasis, normally caused by, Lymphatic filariasis>
- [r09] <Elephantiasis, late complication of, Lymphatic filariasis>
- [r10] <Filariasis, can be treated by, Albendazole>
- [r11] <Nematode, causes, Eosinophilia>
- [r12] <Filariasis, rdfs:subClassOf, Neglected tropical disease>

An obvious advantage of the SPO-syntax is that simple assertions are represented in a fashion that comes close to simple English sentences. However, there are also some disadvantages. For example, use-mention confusions arise when we try to represent both [a02] and [a01]: In [r02], "Nematode" is a name for a type of worm, while in [r01], "Nematode" is referred to as a noun, i.e. a linguistic entity. This obviously introduces ambiguity. Also, there is no direct translation of the assertions [a05], [a08] and [a13] into RDF triples. The triple format has to resort to nested expressions when it comes to the formulation of more complex statements like [a13], which need to be split into sets of simpler assertions if they are to fit the triple format [075]. Finally, the OAV representation does not prevent different modellers from representing the same proposition in different ways, e.g. [a04] could be rendered in any of the following ways:

- [r04] <Filariasis, caused-by, Filariae>
- [r04'] <Filariasis, has agent, Filariae>

[r04"] <Filariae, pathogen-of, Filariasis>.

As a result, such representations are hardly any more interoperable than natural language statements and they are nearly as ambiguous and flexible. RDF triples share with natural language statements a broad margin of discretion regarding their interpretation. Does <Filariasis, causes, Elephantiasis> mean that this will happen necessarily, usually or frequently? Or does it just state that it is not impossible that it happens at all, as opposed to, e.g. <Filariasis, causes, Malaria>? Does it mean that Elephantiasis can only be caused by Filariasis? Without additional knowledge about how to interpret the relation *causes*, we cannot decide this. Humans, in everyday situations, communicate perfectly well when using such ambiguous statements because they are able to spontaneously adapt to relevant contexts of implicit background assumptions. However, such implicit knowledge is not accessible to computers. Hence, if such statements are to be processed automatically, ambiguity and vagueness have to be avoided by precise naming of representational units, as well as logical axioms attached to them to preclude, or at least constrain, competing interpretations.

Formal ontologies – Representing what is universally true

Web Ontology Language (OWL) and Description Logics (DL)

Ontological representation of entity types presupposes the existence of an objective reality about which truths can be discovered by scientific methods [076]. For instance, science tells us that everything that is of the type *Water molecule* has a part that is of the type *Oxygen atom* without exception. The realist view (which postulates that we have at least partial access to the very nature of things in the domain we want to describe) has raised innumerable controversies (recently, e.g. [077,078,079,080,081]). However, many of these controversies have only limited impact on practical issues in ontology engineering. For instance, the developer of an OWL file does not necessarily need to know whether a class corresponds directly to a type or universal (in the sense of ontological realism) or whether it is only a logical construct. Taking the realist stance has nevertheless some advantages. The most striking one is that, given consensus about the things that exist in a domain of interest, agreement can easily be reached about what is universally true for all members of a class.

Basic ontological assertions that concern the entities in a given domain can be formulated using a logic language that can also be processed by reasoning programmes. Description logics (DLs) have become a standard for ontologies with several language profiles [038] standardised by the World Wide Web Consortium (W3C) [041]; the variants commonly used are decidable segments of first-order logic. In biology and medicine, DLs have increasingly been used, from projects GALEN and GALEN-in-use [082,015], over the SNOMED variants CT and RT [059] to the OBO Foundry ontologies [040]. There are several reasons for this:

- DLs are supported by a variety of software tools, such as the Protégé editor [047] and reasoning engines like Hermit [083], Pellet [084] and FaCT++ [085], which allow checking logical consistency and inferring new assertions.
- DLs commonly used for ontology construction are decidable; it is guaranteed that they will always return some result.
- DLs are user-friendly in a sense that users do not need to handle complicated formula with free variables like in FOL. The Manchester syntax [086] facilitates their intelligibility especially.

To use DLs properly, one has to understand their basic building blocks, represented by the names "class", "object property" and "individual", and also comprehend how their constituent logical symbols and expressions are interpreted. For example, all past, present and future individual humans are members of the class *Human*. "Object properties" are binary relations that can only have pairs of individuals as their extensions [087]. For instance, the pair constituted by the American continent and the planet earth lies within the extension of the relation "part of". Classes in DL are always distinct from individuals; classes of classes are not allowed. There is no straightforward way to represent universals, concepts or types. Roughly, *classes* can be seen as proxies for them, for classes are possible extensions of universals, concepts or types. Relations between classes can, therefore, be interpreted as relations between the possible extensions of universals, concepts or types; this simplification has proved sufficient in most cases. Commonly, one uses the relation 'instance of' as relating individuals to universals, concepts or types, whereas the relationship between an individual and a class is called class membership. For example, a given individual filaria worm is an instance of the universal or type *Filaria* and it is a member of the class *Filaria*. Or the authors of this paper are members of the class *Human*. (Strange enough, the RDF(S) language uses "rdf:type" to speak about class membership - a label that is also used in OWL.)

A limitation of OWL DL is that object properties only express binary relations, whereas it has been argued that the representation of facts in the life sciences often requires ternary relations with time as the third argument, e.g. 'has locus' (a, b, t). Assertions about individuals without time indexes are often incomplete. We may want to express that one individual organism is located within another organism, its mother, before birth, but not after delivery. N-ary description logics have been subject to research [088,089] and ontology design patterns that reify n-ary relations have been proposed [090,091]. Nevertheless, the inclusion of time aspects into the definition of object properties is still in an experimental stage in new versions of BFO and BioTop [092,093].

Subclasses and class members

In the following, we illustrate DL syntax and semantics through a set of increasingly complex examples; wrong statements which are included for clarification are marked with an asterisk.

To start, we take a look at the class *Filaria* (corresponding to the respective disease type or universal). The class *Filaria* extends, at every given time, the (mathematical) set of all individual filarial worms. In the same vein, the class *Nematode* extends to all individual nematode worms at all times. To relate the two classes, we introduce the key concept of subclassing or taxonomic subsumption. OWL expresses class inclusion by the `rdfs:subClassOf` relation; e.g. the class *Filaria* is a subclass of the class *Nematode* [a02]. In Manchester OWL Syntax, this is expressed by:

```
[o02] Filaria subClassOf Nematode
```

Semantically, a subclass statement like this corresponds to a necessary implication, i.e. it claims that in all possible worlds (and at all times, etc.) it is true that if something is a member of *Filaria*, then it is also a member of *Nematode*; or for short: All members of *Filaria* are of necessity also members of *Nematode*.

Logical constructors

More complex statements can be obtained by using operators and quantifiers. In the following example we use the keyword 'and' and add a quantified role, using the existential quantifier 'some' (exists). The expression

```
'Parasitic infection' and  
'has agent' some Filaria
```

denotes the class of all individuals that belong to the class '*Parasitic infection*' and are further related through the relation 'has agent' to some member of the class *Filaria*.

This example actually gives us both the necessary and the sufficient conditions needed in order to fully define the class *Filariasis* [a03], [a04]:

```
Filariasis equivalentTo  
'Parasitic infection' and  
'has agent' some Filaria [o03,o04]
```

It can be shown that this implies

```
'Parasitic infection' and  
not ('has agent' some Filaria)  
subClassOf not Filariasis
```

which can be paraphrased as "no parasitic infection that is not caused by filaria is a filariasis". Such transformations are useful in the ontology engineering process to check the validity of an axiom.

The equivalence statement says that: (i) each particular member of the class *Filariasis* is a member of the class '*Parasitic infection*' that has some filaria as agent and (ii) that everything that is a member of the class *Filaria* and has some filaria as agent is a member of the class *Filariasis*. Hence, in any situation, the term on the left can be replaced by the expression on the right.

All members of the class *Filariasis* belong to one of three subclasses [a05]

```
Filariasis equivalentTo  
'Lymphatic filariasis' or  
'Subcutaneous filariasis' or  
'Serous Cavity filariasis' [o05]
```

In nature, these three classes form a strict partition, i.e. they do not overlap [a05]. It is an important characteristic for DL that classes are assumed to be able to overlap unless otherwise stated (the so-called "open world assumption"). Hence, this knowledge has to be made explicit. In order to express non-overlapping, so-called disjointness axioms have to be added:

```
'Lymphatic filariasis' subclassOf not  
( 'Subcutaneous filariasis' or  
'Serous Cavity filariasis' )  
  
'Subcutaneous filariasis' subclassOf not  
( 'Lymphatic filariasis' or  
'Serous cavity filariasis' )  
  
'Serous cavity filariasis' subclassOf not  
( 'Subcutaneous filariasis' or  
'Lymphatic filariasis' )
```

The representation of risks and dispositions

There are also cases in which we want to define a class *A* in terms of reference to a class *B*, whereby not all members of *A* are related to any member of *B*. Typical cases are plans or risks. For instance, the characterization of the class *Risk for lymphatic filariasis*, a disposition, requires a reference to the class *Filariasis*. Since any non-negated use of existentially quantified roles in a DL formalism corresponds to a statement of the form "for all ... there is some ...", we must resort to so-called value restrictions if we are to bring about the needed effect, as for not every member of *Risk for lymphatic filariasis* [a13] is there an actual member of *Filariasis* which

is its realisation. This means that the quantifier 'only' as used in a quantified role is used to specify the allowed range for a given relation [094]. We could then (correctly) state the following:

```
'Risk for lymphatic filariasis' equivalentTo
Risk and
  'has realization' only 'Filariasis transmission'
```

which is the same as

```
'Risk for lymphatic filariasis' equivalentTo
Risk and not ('has realization' by some
(not 'Filariasis transmission'))
```

In plain words, this expression states that a risk of filariasis transmission is a risk that - *if realised* - can be realised only by some member of the class *Filariasis transmission*. In contrast to the simple existential statements, this does not say that some particular filariasis transmission must exist for each filariasis transmission risk.

It must, however, be emphasised that by using the universal quantification (a.k.a. value restriction) or negation we move away from simple, but scalable DL dialects like *EL++* [095], to DLs with a computational complexity that poses severe problems for large ontologies. For instance, the DL-variant OWL-DL (which is supported by common DL reasoners like *FACT++* and *HermiT*) is NEXPTIME-complete [038].

The admission of dispositions (tendencies, capabilities) in our ontologies significantly broadens the scope of what can be represented in ontologies. For instance, the assertion [a07], "The female worms release larvae known as microfilariae into the host's bloodstream", is clearly dispositional. There are certainly female worms that never release larvae, but all physiologically well-formed female worms have the capability or disposition to do so under certain circumstances.

```
'Physiologically well-formed female filarial worm' subclassOf
'bearer of' (Disposition and
  'has realization' only (Releasing and
    'has patient' some Microfilaria and
    'has target' some ('Portion of blood' and
    'has locus' some (Animal and
      'bearer of' some 'Host role')))) [o07]
```

These examples demonstrate the dilemma of logic-based representations: If the purpose is to logically represent and classify large terminological systems like SNOMED CT [096], then the set of allowed constructors must be limited since value restrictions and negations easily lead to

computational intractability; it has to be paid careful attention to which class a certain predicate is ascribed. Nevertheless, computationally more expensive expressions are important in order not to preclude adequate domain representations [097].

The representation of non-ontological statements

Representing the meanings of natural language terms

We have already seen that ontologies are not always strictly distinguished from lexical representations of words or terms which are also found to be called ontologies, though they would more appropriately be called terminologies, vocabularies or thesauri. These latter artefacts represent the meanings of terms, i.e. the basic building blocks of human language. Such representations have to account both for polysemy (one term has two or more distinct meanings, like "lead"), and (quasi-)synonymy (exchangeability of terms in most discourse contexts, like "nematode" vs. "roundworm" [a01]). Such artefacts are mostly known as thesauri or semantic lexicons. They may further contain semantic relations between the individual lexicon entries such as '*broader than*' or '*narrower than*'. WordNet [098], MeSH [012] and most parts of the UMLS Metathesaurus [004] are typical examples of these sorts of representations, which have a long tradition in library science, with literature retrieval being their most common use case. The hierarchy-building semantic relations are fundamentally different from the subclass relation in formal ontologies. As an example, in MeSH we can find both *Plasma* '*narrower than*' *Blood* and '*Fetal Blood*' '*narrower than*' *Blood*; although, from an ontological point of view, the relations involved here would correspond to parthood and specialization, respectively. This difference may not matter for the applications MeSH was originally intended for since the '*narrower than*' relation fits perfectly well with current needs of literature indexing and retrieval: Articles on blood plasma are as relevant to a query on blood as are articles on foetal blood.

It cannot be denied that ontologies and terminologies have similarities. Both come along as (structured) lists of terms. The crucial difference is whether these terms are intended to represent themselves or something else. In terminologies, terms are *mentioned*, whereas in ontologies they are *used*. If a term is mentioned, it functions as a representation of the linguistic entity that is instantiated by the term used. Terminologies are inventories of linguistic entities, whereas in ontologies, as opposed to terminologies, words are used and not only mentioned. Thus, as a rule, ontologies are inventories of extra-linguistic entities. Here, natural language labels or terms are only instruments for the representation of other things. While both ontologies and terminologies are structures, the meaning of the edges has to change in accordance with the meaning of the nodes. Ontologies require (a) relations not between terms or concepts, but between those things the terms refer to; and (b) they require much more precision, beginning with the strict separation of taxonomy (subClassOf), parthood ('has part',

'part Of') and instantiation (rdf:type), which are often run together as a single hierarchical relationship in thesauri [099].

In sum, any representation of term meanings has to also represent the manifold ambiguities of natural human language, whereas ontologies aim at dissolving ambiguity. Hence, we have to keep *lexico-semantic representations* separate from formal ontologies, even if they are sometimes mixed up in loose language. Problems arise unavoidably when thesauri are automatically translated into ontologies [100,101]. This may be tempting as they have certain features in common with ontologies. As the construction of an ontology out of a thesaurus requires numerous additional assumptions, for example concerning quantification, no automated conversion process can provide anything more than a raw sketch that requires careful manual elaboration and curation before it can be of any serious utility for inference purposes [102,103]. Nevertheless, virtually all formal ontology applications require a link between ontology classes and lexical items in the form of human-understandable labels. As many words change their meaning according to the context in which they are used, this bears a considerable risk when using them as labels. For instance, the first word in the statement "Feet have five toes" denotes feet that do not exhibit any congenital or acquired malformation. This would also exclude feet of (most) non-primates and feet of organisms in early embryonic stages [104,105]. When we say that an ulcer is acute, we use the word "ulcer" in order to denote a process; when we say that an ulcer has a diameter of 15 mm, we use it as denoting a physical object. Unambiguous labels (e.g. "Canonic Human Foot" "Ulceration process") are, therefore, mandatory to prevent incorrect choices.

Optionally, synonyms, language references and information about the intended meaning may be given as metadata, but this information does not belong to the proper ontological content and is not used in automated reasoning. At the bottom line, we advocate that separate issues be treated by the separate artefacts: formal ontologies for the representation of the domain entities on the one hand and lexico-semantic representations for the representation of the domain language on the other.

Representation of individuals

Statements about individuals also do not belong to the content of ontologies proper. Nevertheless, such statements about individuals can be represented in an ontologically adequate way in order to make them interoperable with both ontologies and other repositories of individual statements. An important use case in medical informatics is, of course, the representation of patient histories in electronic health records. Such patient records, whether on paper or electronically stored, contain descriptions of individual entities, such as, e.g. the filariasis infection of patient A, her blood sample, a certain lab test performed at a certain day, a treatment episode, a diagnostic statement, as well as geographical or political entities like

regions, countries, cities, etc. Statements of individual facts can be expressed in a straightforward manner in DL terms as instantiations (membership assertions) of corresponding classes. Consider, for example, the following RDF triple, which asserts that the particular disease, from which a given patient is suffering, is a member of the class *Filariasis*:

```
<Illness_1244426, rdf:type, Filariasis>
```

In practice, the individual/class boundary is often drawn in a loose and idiosyncratic way. For example, UniProt [106] entries are asserted to denote "instances" of the class protein. Some authors claim that it depends on the ontology's use case whether something should be represented as a class or as an individual [033,107]. Such arbitrariness would lead to a forking of representations without need, thus hampering the very interoperability which ontologies are intended to support. The grammar of natural human language does, indeed, allow the use of generic terms and proper names quite similarly, as examples like "Salam is a tiger", "The tiger is a mammal" and "Tiger is a species" show. There are several controversial cases that seem to support such optionality. However, on closer inspection these cases always reveal hidden ambiguities [075]. Hence, in a rigorous formal ontological analysis there is no arbitrariness in the distinction between, e.g. this particular filaria here and now (an individual) and *Filaria* (a class). In case of doubt, a good criterion to keep in mind is that individuals exist in space and time; they cannot subsume each other, they can be referred to by proper names and (in many cases) photographed. Classes, in contrast, do not exist in space and time; they do stand to each other in subclass relations, they can have multiple members and they can be referred to by common nouns. Whether an entity is an individual or a class or type is, thus, not a matter of choice on the part of who builds the ontology.

From an ontological perspective, most of UniProt content describes, in fact, protein types in terms of what is universally true for every single protein molecule of this type and never single protein individuals, i.e. concrete molecules. This kind of representation is, therefore, ontological in nature; in contrast, e.g. to an address book, which is a description of individual entities.

Coding information in health records brings with it another challenge, for this information is not only about individuals; it also comes along with quite different epistemic attitudes. Typically, patients are not described directly, but via the epistemic perspective of some observers, including factual statements about the patient, but also hypotheses, prescriptions and plans. For instance, from the statement [a08], "Patient X was diagnosed as being highly suspicious for lymphatic filariasis," we cannot infer that any member of the class *Lymphatic filariasis* exists in this case, for the diagnosis might well be false. This very common situation is a challenge for an ontologically precise account of clinical statements in, e.g. electronic health records [108]. The

European Network of Excellence *SemanticHealthNet* [109] is currently seeking a general solution for this problem. One suggestion is to represent a clinical statement in the following way [110]:

```
'Statement a08' rdf:type '  
  Diagnostic statement' and  
  'has attribute' some Suspicious and  
  'is about condition' only 'Lymphatic filariasis'           [o08]
```

Here, the use of the value restriction "only" is suggested for making a reference to the condition type without asserting the existence of any disease instance for the individual patient in question. Some ontology developers argue an 'is about' statement is in fact a second-order statement [111], i.e. a statement directly about the type '*Lymphatic filariasis*' as opposed to one or more of its instances, as is required by the semantics of OWL-DL. One problem of this representational pattern used in [o08] is that it cannot deal with diagnostic statements involving two or more diseases. Such a complex diagnosis would have to be split into two single statements. Alternatively, the target of the statement can be a clinical situation, as is being currently discussed for medical terminologies like SNOMED CT and the ICD [112]:

```
'Statement about C1 and C2' rdf:type  
  'Diagnostic statement' and  
  'is about situation' only  
    ('C1 situation' and 'C2 situation')
```

Here, situations are phases (temporal segments) of an organism's life process in which a certain condition is present. A situation with C1 and C2 would, therefore, be both a situation with C1 and a situation with C2. Further research is required to ascertain the feasibility of expressing the linkage between information entities like diagnostic statements and types of medical conditions within a decidable ontological framework.

The really interesting knowledge is non-ontological

Although we pointed at ways how to extend the expressive power of ontology languages, such representation formalisms are only appropriate for the representation of a minor portion of scientific statements. In order to feed intelligent computer applications like decision support systems, much more is required to adequately capture a domain. Most typical textbook knowledge, like the association between diseases, signs and symptoms, is not appropriate content for an ontology; they are quantitative, probabilistic or estimative. As Rector [097] puts it: "There are very few interesting items of knowledge that are truly ontological in this strict sense." The same author introduced the term "background knowledge reference resource" to represent decision-relevant information that is widely known, but which a clinician might need to look up.

Default assumptions

In our example, this could be that the knowledge that elephantiasis is commonly a late complication of lymphatic filariasis [a09], which is usually sensitive to albendazole [a10], or that a foot typically has five toes. It refers to all kinds of statements that are assumed to be typically, but not universally true. Such knowledge is traditionally conveyed through scientific texts in a highly context-dependent fashion, often invoking prototypical assertions, such as about how signs and symptoms are related to diseases or how adverse effects are related to drugs, which are expressed in terms of qualitative and sometimes quantitative probabilities.

We will now discuss some examples that illustrate how formal ontology approaches and logical representation formalisms reach their limits when it comes to representing this kind of knowledge. Default knowledge [113,114] is a typical case. It is knowledge about what can be assumed to be typically true in the absence of contravening evidence. OWL-DL does not give us the means to state what is typically true; e.g. one would like to state elephantiasis is normally caused by filariasis [a06]. A statement² such as

```
* Elephantiasis subclassOf  
  'caused by' some Filariasis
```

would be too strong as it rules out that elephantiasis could be caused by anything but filariasis. While the statement that elephantiasis is caused by filariasis is typically true in one context, it may not be true in another one, e.g. where there is a low prevalence of filariasis. An approximation would be to introduce a subclass such as

```
'Elephantiasis caused by filariasis'  
  equivalentTo  
  Elephantiasis and 'caused by' some Filariasis
```

This would not really add additional knowledge and it does not say anything about how likely this causation is. If one would like to express that elephantiasis *may be* caused [115] by filariasis, it would require adding a disjunction of all possible causes like

```
Elephantiasis subclassOf  
  'caused by' only (Filariasis or Podoconiosis or ...)
```

Similar considerations apply to [o11] "Nematode infections may cause eosinophilia". It may be tempting to represent [a06] as follows:

² Here and in some of the following examples we mark invalid statements by an asterisk (*).

```
*Elephantiasis subClassOf  
  'normally caused by' some 'Lymphatic filariasis'
```

By virtue of the DL interpretation of the existential quantifier, this assertion implies that for every member of the class *Elephantiasis* (without exception) there also exists some member of the class *Filariasis*. Of course, the word *normally* that is part of the relation name could be interpreted by humans, but from the point of view of DL it plays no logical role at all. This leads to unintended entailments. For instance, given

```
'Non-infectious elephantiasis' subClassOf  
  'Elephantiasis' and  
  not ('caused by' some 'Lymphatic filariasis')
```

the following (nonsensical) sentence can be logically inferred.

```
* Non-infectious elephantiasis subClassOf  
  'normally caused by' some  
    'Lymphatic filariasis'  
  and not ('caused by' some  
    'Lymphatic filariasis')
```

Such logical effects are important since errors arise when they are not taken into account by users of DL formalisms. Instances of such errors can be found, e.g. in earlier versions of SNOMED CT [116] and in the NCIT ontology [101]. They illustrate how easy it is to express unintended meanings even when using very simple DLs. The reason such examples are commonplace is that ontology developers are not appropriately trained in formal logic and pay too little attention to the principles of sound ontology development. They encumber ontologies with statements that are not of an ontological nature, e.g. typical signs, symptoms or complications of a disease. The resultant invalid axioms then provide support for invalid inferences when used in automated reasoning, which could have unpredictable consequences if medical decision support or other intelligent systems in health care are based on such axioms.

There are other complicated constructions that result from attempts to encode non-trivial knowledge in formal ontologies. Good examples are potentialities expressed as dispositions. Dispositions can exist without ever being realised and even without being able to tell the precise conditions in which a disposition is realised [117]. Albendazole, for example, is a substance that has a disposition to heal filariasis, but it will realise this disposition only when administered in a certain way to a certain sort of patient, even then the treatment may be unsuccessful. We can represent the class of processes of healing filariasis by:

Healing and **'has patient'**
some *'Person having filariasis'*

We could then represent the class of dispositions realised when filariasis is healed:

Disposition and **'has realization'** only
(*Healing* and **'has patient'** some
'Person having filariasis')

The following definition now declares *'Portion of albendazole'* to be an amount of substance in which this disposition inheres:

'Portion of albendazole' subclassOf
'bearer of' some (*Disposition* and
'has realization' only (*Healing* and
'has patient' some
'Person having filariasis'))

This would, furthermore, require defining the class *'Person having Filariasis'*, as a human that is the locus of a filariasis

'Portion of albendazole' subclassOf
'bearer of' some (*Disposition* and
'has realization' only (*Healing* and
'has patient' some (*Human* and
'locus of' some *Filariasis'*)))

Again, such constructions are complex and tend to negatively affect the scalability of an ontology implementation given current reasoning algorithms. Nevertheless, the additional knowledge introduced by including dispositions may be considered insufficient as the conditions under which the dispositions are realised (e.g. substance concentrations [118]), as well as the probabilities of their realisation, remain unspecified.

Dispositions are also addressed by statements like "Lymphatic filariasis has elephantiasis as typical late complication" [a09]. Here, a DL statement like

'Lymphatic filariasis' subclassOf
'has complication' some *Elephantiasis* [o09*]

would rule out any filariasis without elephantiasis, which can easily be ascertained by a simple syntactic transformation resulting in "there is no member of the class lymphatic filariasis that is not complicated by elephantiasis".

Other statements of background knowledge are meta-statements concerning classes. They are misleading because they look, at a first glance, like simple taxonomic relations.

An example is "Filariasis is a neglected tropical disease" [a12]. An assertion like

```
Filariasis subclassOf
  'Neglected tropical disease' [o12*]
```

is problematic under a DL characterized by the view that all statements about classes *are* statements about the corresponding sets of class members. If a group of well-off patients receive early and state-of-the-art treatment for their filariases, the assertion that their individual disease instances are neglected is not tenable. The issue here is that a simple taxonomy-like statement masks a complex non-universal statement like "The treatment of most filariasis patients is currently neglected". The resultant so-called '*is-a overloading*' has been identified as a typical error that occurs when building ontologies in an unprincipled way [119,120,026] as a consequence of being misled by human language shortcuts.

However, the difficulty of representing all the hidden assumptions underlying background knowledge (and the performance problems that result from using the rich logic needed for this) may suggest that we use, instead, a much simpler triple-based representation and devise special reasoning devices to fit these in. Alternatively, one might resort to a broad range of knowledge representation artefacts, such as default logics [121], frames [122], F-logic [123] and several kinds of computationally expensive DL extensions [041], ch. 6. According to the strict sense we use in this paper, the resultant knowledge representation artefacts themselves should not be considered to be formal ontologies. However, classes used in such databases should be taken from and interlinked with appropriate formal ontologies for the respective domain along the lines pointed out in the above examples.

Frequencies and counts

Statements on frequencies and counts like disease prevalence or incidence, which are characteristic for epidemiology, pose another challenge for knowledge representation. An example is [a13], stating a prevalence of risk for lymphatic filariasis of 43%. Here we have two classes, *viz.* *Inhabitant of Nigeria* and *Inhabitant of Nigeria at risk for Filariasis*. Both classes have a cardinality (integer value) and the risk prevalence is given by their ratio. The prevalence is, therefore, not a characteristic of the disease, but of the population of affected persons. One could extend the DL notation by symbolizing the cardinality of the extension of a class (i.e. the number of class members) by enclosing the class name in "||".

```
| Inhabitant of Nigeria at risk for filariasis | /
| Inhabitant of Nigeria | = 0.43
```

This demonstrates a way that probabilistic background knowledge could be expressed using DL individuals extended by arithmetic operators referring to individuals. This is, therefore, not in

the scope of formal ontologies, just as little as in alternative approaches like probabilistic extensions [124,125].

Recommendations

Knowledge Representation evolved as a subdiscipline of Artificial Intelligence with the purpose of enabling computers to draw new conclusions from existing data and information. When the term "ontologies" became popular in computer science in the nineties, it was thus often regarded as a new catchword for something that already existed, namely knowledge representation artefacts, especially semantic networks [126]. The vision of the Semantic Web nourished the illusion that packaging all kinds of assertions into handy subject-predicate-object triples would automatically create some kind of semantics resulting in some kind of ontology. What usually resulted were, however, just repositories of syntactically standardised statements without semantic standardization. These statements remained as ambiguous and context-dependent as natural language. In medical informatics, triple-based representation formats have been popularized since the first release of the UMLS metathesaurus [127], but they reach back to [128]. Occasionally, even UMLS has been described as an ontology [129,130], but our impression is that - at least in biomedical informatics literature - this terminological confusion has gradually been overcome. It had been a merit of the Semantic Web to help standardise description logics dialects, which were readily taken up by ontology developers which then had interchangeable formats and public-domain tools like editors and reasoners at hand. The formal semantics of the ontology language OWL, together with a large body of work from the emerging Applied Ontology [131] community, has helped to clarify what can be straightforwardly expressed by ontologies and what cannot. The most important message is that ontology axioms can only express what is true for all members of a class. This precludes contingent, probabilistic and default statements, as well as rules and meta-class assertions, to be included into formal ontologies. The advantage of this restriction is that ontologies are, therefore, limited to express the most stable assertions about a domain. This makes them a solid foundation for knowledge representation artefacts proper, including Bayesian networks.

Once consensus exists what ontologies are and are not, good practise principles (similar to the desiderata J. Cimino formulated for terminologies [132]) should guide the development of ontology artefacts. The following recommendations are based on own experiences [133] and published literature:

- *Appropriate Naming*: Even if the representation of domain terms is not the proper goal of an ontology, terms are important as labels for the representational units in an ontology in order to enhance the understanding of human users. Ideally, labels for classes and relations should be unambiguous and self-explaining and conform to naming conventions [134]. This is not trivial, as natural language expressions tend to be ambiguous. Moreover,

a naive literal interpretation of composed terms is often misleading: A complicated pregnancy is a pregnancy, but a prevented pregnancy is not a pregnancy, as little as a planned biopsy is a biopsy or a suspected asthma a kind of asthma. Such idioms should be avoided as they bear the risk of incorrect subclass assertions. To avoid ambiguity, class or relation labels in good ontologies are often somewhat artificial and not commonly used in oral and written communication, yet this is a price worth paying for clarity. Ontology labels are not supposed to provide the lexical basis for text mining or information extraction systems. For these and other use cases, terminologies or vocabularies need to be linked to the ontology as external resources. Homonymy and synonymy are linguistic phenomena that should be addressed by language resources and not by ontologies.

- *Explicit top level categories and relations*: The usefulness of formalized upper-level ontologies (Fig. 2) is not universally acknowledged [135]. However, whoever refrains from explicit upper-levels still keeps their own upper level preferences in mind. These may diverge from others' preferences and obscure important categorial distinctions such as between processes and functions or between domain entities and information entities. Only clear-cut top-level categories are able to support disambiguation of terms and expressions. Upper-level ontologies can enforce type constraints provided that their upper-level categories are mutually disjoint. For instance, if a relation like '**inheres in**' has its range constrained to material objects, any assertion on the inherence of something in a different category, e.g. a process, would render the ontology inconsistent. Current top-level ontologies are increasingly providing for these constraints.
- *Ontological commitment* [116]: No ambiguity can be allowed about what category of things are meant by a class in an ontology, e.g. *Elephantiasis* as the class of pathological processes, the class of pathologically altered body parts or the class of patients with this disorder. One possibility is to leave several interpretations open and allow for disjunctive class definitions in an ontology [136].
- *Keeping ontology apart from epistemology*: Epistemology describes what an agent sees, knows or records about a domain, whereas ontologies, ideally, describe the entities in that domain as they exist. Especially in medicine, the reference to a term or an ontology class often does not mean that a related individual exists. Clinical decisions are often triggered by mere suspicions; e.g. children who exhibit a clinical picture suspicious for meningitis are treated as if they have meningitis. Legacy medical classification systems such as ICD are "infiltrated" by epistemic notions, such as in classes like "Tuberculosis of lung, nodular bacteriological or histological examination not done" in ICD-9-CM [137,138]. It is a repeatedly expressed desideratum that epistemic aspects should be treated in information models; yet there are overlaps between ontologies and information models which give rise to conflicting representations, requiring sophisticated mitigation

strategies [139]. The very same complex information (e.g. a clinician's hypothesis of a stenosis of the left carotid artery) can be represented to different proportions in clinical ontologies and clinical information models and creates interoperability problems [140]. A way out of this dilemma could be to represent epistemic aspects or recorded information in a separate branch of the ontology [108].

- *Avoiding ontology artefacts*: Typical medical classification systems pursue the goal of artificially producing exhaustive partitions for statistical purposes. This leads to sibling classes like the following ones in ICD-10, where I20 (Angina pectoris) has as subtypes: I20.0 (Unstable angina); I20.1 (Angina pectoris with documented spasm); I20.8 (Other forms of angina pectoris); I20.9 (Angina pectoris, unspecified). Whereas I20.8 refers to the set theoretic complement of the union of I20.0 and I20.1 and is, therefore, dependent on the version of the terminology, I20.9 carries epistemic meaning. Such classes should be avoided in ontologies.
- *Educational resources for good ontology training*: There is still a tendency to create ontologies from the scratch by using editing tools like Protégé [141] with idiosyncratic upper-level classes and relations and with limited insight into the semantics of description logics. Training resources are therefore necessary, like the Pizza Ontology [142] and training material based on the GoodOD guideline [143]. Besides basic principles of logic and ontology, these resources should also convey descriptions of ontology construction and maintenance workflow, including quality assurance. The iterative use of a DL classifier to maintain consistency is fundamental. It also seems that a strict adherence to upper-level categories and relations (and, as a consequence, to the constraints attached to them) is a good method to keep developers on the right path and prevents them from generating idiosyncratic artefacts.
- *Appropriate tooling*: As much as we have observed an improvement in ontology editing and reasoning tools [141,083,084,085], we have to admit that they still depend on academic labs and public funding in terms of development and maintenance. It is desirable that ontology as an engineering discipline should be served by industry-standard tools.
- *Evidence-based ontology engineering as an ultimate goal*: Although the concept of Evidence-Based Software Engineering has not found an acceptance as broad as Evidence-Based Medicine [144] and brings up different methodological challenges [145], there is a need to develop and refine approaches to measure the fitness for purpose of representational artefacts. These quality parameters must be empirically validated for their explicit scopes. Research on such parameters and on guidelines to sustain them is only in its beginning and it can hardly be abstracted from research on the efficiency of training such guidelines [146]. One direction of ontology evaluation can be the

deployment of competency questions [147] as a way to formalize requirements and assess their fulfilment. Another option would be to benchmark the performance of an ontology for a certain use case (like annotation or information retrieval) [148]. The very idea of ontologies is, however, to function as a purpose-independent reference artefact. The claim to be empirically substantiated is that such purpose-independent biomedical ontologies can, in a second step, be linked to purpose-focussed computational artefacts like knowledge bases and decision-support systems, yielding both efficient performance and interoperability with other such systems.

Conclusion

In the past, knowledge bases and formal ontologies have not clearly been distinguished. Also, RDF-based representations, as popularized by the Semantic Web, have tended to blur this distinction. It is not necessary for being an ontology to be written in OWL, but neither is this sufficient: OWL files can be created or used for a variety of purposes and in many cases the underlying clear-cut semantics of description logics is being ignored or overridden. This practise is detrimental for the practise of both knowledge representation and formal ontologies. From a medical decision support background, Musen [023] has evaluated the adequacy of ontologies for knowledge representation. He emphasises that ontologies are mostly insufficient for problem solving use cases; although, they provide important basic knowledge on which intelligent systems can be built. Ten years on, this answer is still valid, though we can now build on an impressive body of research in Applied Ontology as well as representational standards like OWL. Both have facilitated drawing a clear boundary of what ontologies should represent and what they should not. In a nutshell, ontologies are limited to express universal statements about types of entities. This excludes statements about the meaning and exchangeability of natural language terms in human discourse, statements about individual entities, as well as probabilistic, default and contingent assertions. We have demonstrated how formal ontologies can be stretched toward this direction, but at the price of additional complexity which is not necessarily compensated by a real benefit. In light of still persisting quality issues, we formulate desiderata for good quality ontologies to be used as a stable framework which more context-dependent biomedical knowledge representation and reasoning applications like clinical decision support system can capitalize on. For this purpose, it can even be combined with probabilistic approaches used in "big data". Future research should examine the question if and how the same probabilistic heuristics working with structured data rooted in formal ontologies can outperform unstructured texts as sources for automated knowledge acquisition.

Acknowledgements

Research for this paper has been supported for both authors by the German Research Agency (DFG) under the auspices of the project Good Ontology Design (GoodOD), JA 1904/2-1, SCHU 2515/1-1 as well as for the first author by the EU commission, which is workpackage leader in the FP7 network of excellence SemanticHealthNet (call ICT-2011-7, grant 288408).

References

(Web resources: last accessed Mar 28, 2013)

[001] Anderson C (2008). The end of theory: the data deluge makes the scientific method obsolete. http://www.wired.com/science/discoveries/magazine/16-07/pb_theory .

[002] Müller H, Hanbury A, Al Shorbaji N. Health information search to deal with the exploding amount of health information produced. *Methods of Information in Medicine*. 2012 Dec 4; 51 (6): 516-518.

[003] Keim B (2012). Paging Dr. Watson: Artificial intelligence as a prescription for health care. <http://www.wired.com/wiredscience/2012/10/watson-for-medicine/all/>.

[004] United States National Library of Medicine (NLM). Unified Medical Language System (UMLS), 2013. <http://www.nlm.nih.gov/research/umls>.

[005] World Health Organization (WHO). International Classification of Diseases (ICD), 2013. <http://www.who.int/classifications/icd>.

[006] World health Organization. International Classification of Functioning, Disability and Health (ICF) <http://www.who.int/classifications/icf/en/> .

[007] UMC (Uppsala Centre for International Drug Monitoring). WHO Drug Dictionary Enhanced, 2013. <http://www.umc-products.com>.

[008] WHO Collaborating Centre for Drug Statistics Methodology (WHOCC). Anatomical Therapeutic Chemical Classification System (ATC), 2013. <http://www.whocc.no/atcddd>.

[009] United States National Library of Medicine (NLM). RxNorm, 2013. <http://www.nlm.nih.gov/research/umls/rxnorm>.

[010] World Health Organization. Dictionary of Medicines and Devices (dm+d), 2013. <http://www.dmd.nhs.uk>.

[011] Regenstrief Institute. Logical Observation Identifiers Names and Codes (LOINC), 2013. <http://loinc.org>.

[012] United States National Library of Medicine (NLM). Medical Subject Headings (MeSH), 2013. <http://www.nlm.nih.gov/mesh>.

[013] Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT), 2008. <http://www.ihtsdo.org/snomed-ct>.

[014] Greenes RA, McClure RC, Pattison-Gordon E, Sato L. The findings-diagnosis continuum: implications for image descriptions and clinical databases. *Proc Annu Symp Comput Appl Med Care*. 1992: 383-387.

[015] Rodrigues JM, Trombert-Pavot B, Baud R, Wagner J, Rusch P, Meusnier F. Galen-In-Use: an EU Project applied to the development of a new national coding system for surgical procedures: NCAM. *Studies of Health Technology and Informatics*. 1997; 43 Pt B: 897-901.

- [016] Schulz S, Romacker M, Hahn U. Part-whole reasoning in medical ontologies revisited -introducing SEP triplets into classification-based description logics. AMIA Symposium Proceedings 1998: 830-834.
- [017] Hahn U, Romacker M, Schulz S. MedSynDiKATe--design considerations for an ontology-based medical text understanding system. Proc AMIA Symp. 2000: 330-334.
- [018] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genetics 2000 May; 25 (1): 25-29.
- [019] Cimino JJ, Zhu X. The practical impact of ontologies on biomedical informatics. Yearb Med Inform. 006: 124-35.
- [020] Bodenreider O. Biomedical ontologies in action: role in knowledge management, data integration and decision support. Yearb Med Inform. 2008: 67-79.
- [021] Guarino N, Giaretta P. Ontologies and Knowledge Bases - Towards a Terminological Clarification. In: Towards Very Large Knowledge Bases, 1995; 25-32. IOS Press, Amsterdam, The Netherlands.
- [022] Kuśnierczyk W. Nontological Engineering. Proceedings of the 2006 conference on Formal Ontology in Information Systems: Proceedings of the Fourth International Conference (FOIS 2006), IOS Press Amsterdam 2006: 39-50.
- [023] Musen M. Ontologies - Necessary-Indeed Essential-but Not Sufficient. In Brewster, C. and O'Hara, K. Knowledge Representation with Ontologies: The Present and Future. IEEE Intelligent Systems 2004; 19 (1): 77-79.
- [024] Hofweber, Thomas, "Logic and Ontology", The Stanford Encyclopedia of Philosophy (Spring 2013 Edition), Edward N. Zalta (ed.), <http://plato.stanford.edu/archives/spr2013/entries/logic-ontology/> .
- [025] Guarino N. Formal Ontology, Conceptual Analysis and Knowledge Representation. International Journal of Human-Computer Studies 1995; 43 (5-6): 625-640.
- [026] Smith B. Beyond Concepts: Ontology as Reality Representation. Proceedings of the International Conference on Formal Ontology in Information Systems, 11, 2004: 39-50.
- [027] Quine WV. On what there is. In: Gibson R. Quintessence-Basic Readings from the Philosophy of W. V. Quine. Cambridge: Belknap Press, Harvard University, 2004.
- [028] Kremer K, Wolf U. Ontologie. In: Ritter J, Gründer K (eds.). Historisches Wörterbuch der Philosophie, vol. 6, Darmstadt: Wissenschaftliche Buchgesellschaft, 1984: 1189-1200.
- [029] Uschold M, Gruninger M. Ontologies: Principles, Methods and Applications. Knowledge Engineering Review, 1996: 11 (2): 39-136.
- [030] Castillo-Barrera FE, Medina Ramírez C, Duran-Limon HA (2011) Knowledge Capitalization in a Component-Based Software Factory: a Semantic Viewpoint. LA-NMR, volume 804 of CEUR Workshop Proceedings: 105-114. .

- [031] Fiannaca A, La Rosa M, Rizzo R, Urso A, Gaglio, S. An ontology design methodology for Knowledge-Based systems with application to bioinformatics. In Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2012 IEEE Symposium on Date of Conference: 9-12 May 2012: 85-91.
- [032] Semy SK, Hetherington-Young, KN, Frey SE. Ontology Engineering: An Application Perspective. Technical Paper. 2004, The MITRE Corporation. http://www.mitre.org/work/tech_papers/tech_papers_04/04_0847/index.html.
- [033] Stevens R, Goble CA, Bechhofer S. Ontology-based knowledge representation for bioinformatics. Briefings in Bioinformatics 2000 Nov; 1 (4): 398-414.
- [034] White S. What the Heck is an Ontology? <http://www.catalysoft.com/articles/WhatIsAnOntology.html>.
- [035] Gruber TR. A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition (Special issue: Current issues in knowledge modeling), 1993; 5 (2): 199-200.
- [036] Schulz S, Johansson I. Continua in Biological Systems. The Monist 2007; 90 (4): 499-522.
- [037] Sowa J. Ontology. <http://www.jfsowa.com/ontology/>.
- [038] OWL2 profiles <http://www.w3.org/TR/owl2-profiles>.
- [039] Klein GO, Smith B. Concept Systems and Ontologies: Recommendations for Basic Terminology. Trans Jpn Soc Artif Intell. 2010 Jan 1; 25 (3): 433-441.
- [040] Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ; OBI Consortium, Leontis N, Rocca-Serra P, Ruttenberg A, Sansone SA, Scheuermann RH, Shah N, Whetzel PL, Lewis S. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature Biotechnology 2007 Nov; 25 (11): 1251-1255.
- [041] Baader F, Calvanese D, McGuinness DL, Nardi D, Patel-Schneider PF. The Description Logic Handbook. Theory, Implementation, and Applications (2nd Edition). Cambridge: Cambridge University Press, 2007.
- [042] Musen MA, Noy NF, Shah NH, Whetzel PL, Chute CG, Story MA, Smith B; NCBO team. The National Center for Biomedical Ontology. JAMIA 2012 Mar-Apr; 19 (2): 190-195.
- [043] Whetzel PL, Noy NF, Shah NH, Alexander PR, Nyulas C, Tudorache T, Musen MA. Nucleic Acids Res. BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. 2011 Jul; 39 .
- [044] The OBO Foundry. <http://www.obofoundry.org>.
- [045] Ontology Design Datterns. <http://ontologydesignpatterns.org>.
- [046] Ontology Design Patterns Public Catalog. <http://www.gong.manchester.ac.uk/odp/html/>.
- [047] BMIR (Stanford Center for Biomedical Informatics Research). The Protégé Ontology Editor and Knowledge Acquisition System, 2013. <http://protege.stanford.edu>.
- [048] The OWL API. <http://owlapi.sourceforge.net/> .

- [049] OntoFox <http://ontofox.hegroup.org/>.
- [050] Poveda-Villalón M, Suárez-Figueroa M, Gómez-Pérez A: Validating Ontologies with OOPS! EKAW 2012: 267-281.
- [051] Schober D, Tudose I, Svatek V, Boeker M. OntoCheck: verifying ontology naming conventions and metadata completeness in Protégé 4. J Biomed Semantics. 2012 Sep 21; 3 Suppl 2: S4.
- [052] Horrocks I, Patel-Schneider PF, van Harmelen F. From SHIQ and RDF to OWL: The Making of a Web Ontology Language. Journal of Web Semantics, 1 (1): 7-26, 2003.
- [053] Golbreich C, Horrocks I: The OBO to OWL mapping, GO to OWL 1.1! OWLED 2007. OWL: Experiences and Directions. Third International Workshop, Innsbruck, Austria 6-7 June 2007.
- [054] Grenon P, Smith B. (2004). SNAP and SPAN: Towards Dynamic Spatial Ontology. Spatial Cognition and Computation 4 (1): 69-103.
- [055] Borgo S, Masolo C. Ontological Foundations of DOLCE. In Staab S, Studer R (eds.), Handbook on Ontologies (Second Edition), Springer Verlag, 2009: 361-382.
- [056] Sowa JF, Knowledge Representation: Logical, Philosophical, and Computational Foundations, Brooks Cole Publishing Co., Pacific Grove, CA, 2000.
- [057] Guarino N. An Overview of OntoClean. Handbook on Ontologies, second edition, Springer, 2009: 201-222.
- [058] ONTOLOG-collaborative work environment. <http://ontolog.cim3.net/>.
- [059] Cornet R, de Keizer N (2008). Forty years of SNOMED: a literature review. BMC Medical Informatics and Decision Making 8 (Suppl. 1), S2.
- [060] Schulz S, Rector A, Rodrigues JM, Chute C, Üstün B, Spackman, K. Ontology-based convergence of medical terminologies: SNOMED CT and ICD 11. eHealth 2012-Health Informatics meets eHealth; May 10-11, 2012; Vienna, Austria.
- [061] McCray AT. An upper-level ontology for the biomedical domain. Comp Funct Genomics. 2003; 4 (1): 80-84 .
- [062] Smith B, Ceusters W, Klagges B, Köhler J, Kumar A, Lomax J, Mungall C, Neuhaus F, Rector AL, Rosse C. Relations in biomedical ontologies. Genome Biology 2005; 6 (5): R46. Epub 2005 Apr 28.
- [063] Beisswanger E, Schulz S, Stenzhorn H, Hahn U: BioTop: An upper domain ontology for the life sciences. A description of its current structure, contents and interfaces to OBO ontologies. Applied Ontology, 2008; 3 (4): 205-212.
- [064] Schulz S, Beisswanger E, van den Hoek L, Bodenreider O, van Mulligen EM. Alignment of the UMLS semantic network with BioTop: methodology and assessment. Bioinformatics. 2009 Jun 15; 25 (12): i69-76.
- [065] Smith B, Almeida M, Bona J, Brochhausen M, Ceusters W, Courtot M, Dipert R, Goldfain A, Grenon P, Hastings J, Hogan W, Jacuzzo L, Johansson I, Mungall C, Natale D, Neuhaus F, Petosa A, Rovetto R,

- Ruttenberg A, Ressler M, Schulz S. Basic Formal Ontology 2.0. Draft specification and user's guide. <http://purl.obolibrary.org/obo/bfo/2012-07-20/Reference>.
- [066] Hoehndorf R, Loebe F, Poli R, Herre H, Kelso. GFO-Bio: A biological core ontology. *Applied Ontology*, 2008, 3 (4): 219-227.
- [067] Dumontier M (2012). The Semanticscience Integrated Ontology (SIO). <http://semanticscience.org>.
- [068] Brachman RJ, Gilbert VP, Levesque HJ: An Essential Hybrid Reasoning System: Knowledge and Symbol Level Accounts of KRYPTON. *IJCAI 1985*: 532-539.
- [069] Filariasis. <http://en.wikipedia.org/wiki/Filariasis> .
- [070] Elephantiasis. <http://en.wikipedia.org/wiki/Elephantiasis> .
- [071] Podoconiosis. <http://en.wikipedia.org/wiki/Podoconiosis> .
- [072] Meeting of the International Task Force for Disease Eradication, April 2011. *Wkly Epidemiol Rec*. 2011 Aug 5; 86 (32): 341-351.
- [073] Shortliffe EH, Davis R, Axline SG, Buchanan BG, Green CC, Cohen SN. Computer-based consultations in clinical therapeutics: explanation and rule acquisition capabilities of the MYCIN system. *Computers and Biomedical Research* 1975; 8 (8): 303-320.
- [074] Klyne G, Carroll GG. Resource Description Framework (RDF): Concepts and Abstract Syntax. W3C Recommendation 10 February 2004 <http://www.w3.org/TR/rdf-concepts/>.
- [075] Schulz S, Stenzhorn H, Boeker M, Smith B. Strengths and limitations of formal ontologies in the biomedical domain. *RECIIS : Revista Eletrônica de Comunicação, Informação & Inovação em Saúde*. 2009 Mar 1; 3 (1): 31-45.
- [076] Boyd R. Scientific Realism, *Stanford Encyclopedia of Philosophy*, 2002. <http://plato.stanford.edu/entries/scientific-realism>.
- [077] Merrill G. Ontological Realism: Methodology or Misdirection. *Applied Ontology* 2010; 5 (2): 79-108.
- [078] Merrill G. Realism and Reference Ontologies: Considerations, Reflections, and Problems. *Applied Ontology* 2010; 5: 189-221.
- [079] Smith B, Ceusters W. Ontological realism: A methodology for coordinated evolution of scientific ontologies. *Applied Ontology* 2010 Nov 15; 5 (3-4): 139-188.
- [080] Maojo V, Crespo J, García-Remesal M, de la Iglesia D, Perez-Rey D, Kulikowski C. Biomedical ontologies: toward scientific debate. *Methods Inf Med*. 2011; 50 (3): 203-216.
- [081] Brochhausen M, Burgun A, Ceusters W, Hasman A, Leong TY, Musen M, Oliveira JL, Peleg M, Rector A, Schulz S. Discussion of "Biomedical ontologies: toward scientific debate". *Methods of Information in Medicine* 2011; 50 (3): 217-236.
- [082] Rector AL, Bechhofer S, Goble CG, Horrocks I, Nowlan WA, and Solomon WD. The GRAIL concept modelling language for medical terminology. *Artificial Intelligence in Medicine* 1997; 9 (2): 139-171.

- [083] Hermit OWL reasoner. <http://www.hermit-reasoner.com/>.
- [084] Sirin E, Parsia B, Cuenca Grau B, Kalyanpur A, Katz Y. Pellet: A Practical OWL DL Reasoner. *Journal of Web Semantics* 2007; 5 (2): 51-53.
- [085] Tsarkov D, Horrocks I. FaCT++ Description Logic Reasoner: System Description. *Proceedings of the Third International Joint Conference on Automated Reasoning*, 8, 2006, Heidelberg: Springer, 2006: 292-297.
- [086] Horridge M, Drummond N, Goodwin J, Rector A, Stevens R, Wang H. The Manchester OWL Syntax. *Proc. of the OWLED Workshop: Experiences and Directions 2006*, 11, 2006. <http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-216>.
- [087] W3C OWL working group. OWL 2 Web Ontology Language, Document Overview. W3C Recommendation 11 December 2012 <http://www.w3.org/TR/owl2-overview/>.
- [088] Krieger HU. Where temporal description logics fail: representing temporally-changing relationships. *KI 2008: Advances in Artificial Intelligence, Lecture Notes in Artificial Intelligence 5243*: 249-257, Springer, 2008.
- [089] Zamborlini V, Guizzardi G: On the Representation of Temporally Changing Information in OWL. *EDOCW 2010*: 283-292.
- [090] Noy N, Rector A. Defining N-ary Relations on the Semantic Web. W3C Working Group Note 12 April 2006. <http://www.w3.org/TR/swbp-n-aryRelations/>.
- [091] Grewe N. A generic reification strategy for n-ary relations in DL. *2nd Workshop of Ontologies in Biomedicine and Life Sciences (OBML)*, Mannheim, 9.-10.9. 2010: <http://www.bioontology.org/node/628>.
- [092] Basic Formal Ontology (BFO). <http://code.google.com/p/bfo/>.
- [093] Schulz S (2012). BioTop-A Top-Domain Ontology for the Life Sciences. <http://purl.org/biotop>.
- [094] Röhl J, Jansen L. Representing Dispositions. *Journal of Biomedical Semantics* 2011; 2 (Suppl. 4), S4 (9. August 2011). <http://www.jbiomedsem.com/content/2/S4/S4>.
- [095] Baader F, Peñalosa R, Suntisrivarporn B. Pinpointing in the Description Logic EL. *Proceedings of the DL 2007 International Workshop on Description Logics (DL2007)*, Edited by Diego Calvanese et al. <http://ceur-ws.org/Vol-250/>.
- [096] Baader F, Litz C, Suntisrivarporn B. CEL - A Polynomial-time Reasoner for Life Science Ontologies. *Proceedings of the International Joint Conference on Automated Reasoning*, 8, 2006, Heidelberg: Springer, 2006, 287-291.
- [097] Rector A. Barriers, approaches and research priorities for integrating biomedical ontologies, 2008. http://www.semantichhealth.org/DELIVERABLES/SemanticHEALTH_D6_1.pdf.
- [098] Fellbaum C (Ed.). *WordNet: An Electronic Lexical Database*. Cambridge: MIT Press, 1998.
- [099] Kless D, Milton SK, Kazmierczak E. Relationships and relata in ontologies and thesauri: Differences and similarities. *Applied Ontology* 7 (4): 401-428 (2012).

- [100] Soualmia LF, Golbreich C, Darmoni S. Representing the MeSH in OWL: Towards a Semi-Automatic Migration. Workshop on Formal Biomedical Knowledge Representation (KR-MED), 7, 2004: 81-87.
- [101] Schulz S, Schober D, Tudose I, Stenzhorn H. The Pitfalls of Thesaurus Ontologization-the Case of the NCI Thesaurus. AMIA Annual Symposium Proceedings 2010 Nov 13; 2010: 727-731.
- [102] Schulz S, Hahn U. Medical knowledge reengineering-converting major portions of the UMLS into a terminological knowledge base. International Journal of Medical Informatics, 2001; 64 (2-3): 207-221, 2001.
- [103] Kless D, Jansen L, Lindenthal J, Wiebensohn J. A method for re-engineering a thesaurus into an ontology. Donnelly M, Guizzardi G (eds.), Formal Ontology and Information Systems (FOIS 2012), Amsterdam: IOS 2012 (Frontiers in Artificial Intelligence and Applications 239), 133-146.
- [104] Neuhaus F, Smith B. Modelling Principles and Methodologies. Relations in anatomical ontologies. In Burger A, Davidson D, Baldock R (eds.): Anatomy Ontologies for Bioinformatics: Principles and Practice, 2007.
- [105] Schulz S, Hahn Schulz S, Hahn U. Towards the ontological foundations of symbolic biological theories. Artificial Intelligence in Medicine 2007 Mar; 39 (3): 237-250 .
- [106] UNIPROT (Universal Protein Resource Consortium). UniProt Protein Knowledgebase, 2013. <http://www.uniprot.org>.
- [107] Noy NF, McGuinness DL. Ontology Development 101: A Guide to Creating Your First Ontology. 2001, Technical Report, <http://ce.sharif.edu/~daneshpajouh/ontology/ontology-tutorial-noy-mcguinness.pdf>.
- [108] Schulz S, Karlsson D. Records and situations. Integrating contextual aspects in clinical ontologies. In: Shah N, Sansone SA, Stephens R; Soldatova L eds. The 14th Annual Bio-Ontologies Meeting. Vienna, Austria: ISCB 2011. 49-52.
- [109] SemanticHealthNet Network of Excellence. <http://www.semantichealthnet.eu/> .
- [110] Martínez-Costa C, Schulz S Ontology-based reengineering of the SNOMED CT context hierarchy. Manuscript under review at AIME 2013.
- [111] Stevens R. Higher Order Knowledge in Ontologies. Ontogenesis Blod, May 2011. <http://ontogenesis.knowledgeblog.org/1080>
- [112] Schulz S, Rector A, Rodrigues JM, Spackman K. Competing interpretations of disorder codes in SNOMED CT and ICD. AMIA Annu Symp Proc. 2012; 2012: 819-827.
- [113] Rector AL. Defaults, Context, and Knowledge: Alternatives for OWL-Indexed Knowledge Bases. Pacific Symposium on Biocomputing 2004: 226-237.
- [114] Hoehndorf R, Loebe F, Kelso J and Herre H. Representing default knowledge in biomedical ontologies: application to the integration of anatomy and phenotype ontologies. BMC Bioinformatics 2007, 8: 377.
- [115] Rector AL, Stevens R, Drummond N. What Causes Pneumonia? The Case for a Standard Semantics for "may" in OWL. OWLED: OWL - Experiences and Directions, 26-27 October 2008, Karlsruhe.

- [116] Schulz, S; Cornet, R; Spackman, K. Consolidating SNOMED CT's ontological commitment. *Applied ontology*, 2011; 6: 1-11.
- [117] Jansen L. "On Ascribing Dispositions", in: Max Kistler, Bruno Gnessounou (Eds.), *Dispositions and Causal Powers*, Aldershot: Ashgate 2007: 161-177.
- [118] Hastings, J; Steinbeck, C, Jansen, L; Schulz, J. Substance concentrations as conditions for the realization of dispositions. *CEUR Workshop Proceedings: KR-MED 2010-Semantic Applications in Life Sciences. Proceedings of the 4th International Workshop on Formal Biomedical Knowledge Representation*, hosted by Bio-Ontologies 2010; July 9-10, 2010; Boston, MA, USA.
- [119] Guarino N. Avoiding IS-A Overloading: The Role of Identity Conditions in Ontology Design. *International Conference on Spatial Information Theory: Cognitive and Computational Foundations of Geographic Information Science, Proceedings*, 1999: 221-234.
- [120] Welty C, Guarino N. Supporting ontological analysis of taxonomic relationships, *Data & Knowledge Engineering* 2001; 39 (1): 51-74.
- [121] Reiter R. A logic for default reasoning. *Artificial Intelligence* 1980; 13: 81-132.
- [122] Minsky M. A Framework for Representing Knowledge. MIT-AI Laboratory Memo 306, June, 1974. <http://web.media.mit.edu/~minsky/papers/Frames/frames.html>.
- [123] Kifer M, Lausen G. F-logic: a higher-order language for reasoning about objects, inheritance, and scheme. *ACM SIGMOD Record*. 2, 1989; 134-146.
- [124] Koller D, Levy A, Pfeffer A. P-classic: A tractable probabilistic description logic. *Proceedings of AAAI-1997*: 390-397.
- [125] Klinov P. Pronto: A Non-monotonic Probabilistic Description Logic Reasoner. *Proc. of the European Semantic Web Conference*, 6, 2008. Heidelberg: Springer, 2008; 822-826.
- [126] Quillian MR. Word concepts: A theory and simulation of some basic semantic capabilities. *Syst. Res.* 1967; 12: 410-430.
- [127] Lindberg C. The Unified Medical Language System (UMLS) of the National Library of Medicine. *J Am Med Rec Assoc.* 1990 May; 61 (5): 40-42.
- [128] Pauker SG, Gorry GA, Kassirer JP, Schwartz WB. Towards the simulation of clinical cognition. Taking a present illness by computer. *Am J Med.* 1976 Jun; 60 (7): 981-996.
- [129] Nardon FB, Moura LA. Knowledge sharing and information integration in healthcare using ontologies and deductive databases. *Stud Health Technol Inform.* 2004; 107 (Pt 1): 62-66.
- [130] Racoceanu D, Lacoste C, Teodorescu R, Vuilleminot N. A Semantic Fusion Approach Between Medical Images and Reports Using UMLS. *3rd Asia Information Retrieval Symposium*. 2006: 460-475.
- [131] Munn K, Smith B, eds. *Applied Ontology. An Introduction*, Frankfurt am Main: Ontos, 2008.
- [132] Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. *Methods Inf Med.* 1998 Nov; 37 (4-5): 394-403.

- [133] Jansen J, Schulz S. The Ten Commandments of Ontological Engineering. Proceedings of the 3rd Workshop of Ontologies in Biomedicine and Life Sciences (OBML), Berlin, 06.-07.10. 2011, www.onto-med.de/obml/ws2011/obml2011report.pdf.
- [134] Schober D, Smith B, Lewis SE, Kusnierczyk W, Lomax J, Mungall C, Taylor CF, Rocca-Serra P, Sansone SA. Survey-based naming conventions for use in OBO Foundry ontology development. *BMC Bioinformatics*. 2009 Apr 27; 10: 125.
- [135] Welty CA. Are Upper-Level Ontologies Worth the Effort? Proceedings of the Eighth International Conference on Principles and Knowledge Representation and Reasoning (KR-02), 2002, 632.
- [136] Schulz S, Spackman K, James A, Cocos C, Boeker M. Scalable representations of diseases in biomedical ontologies. *Journal of Biomedical Semantics*. 2011 May 17; 2 Suppl 2: S6.
- [137] Bodenreider O, Smith B, Burgun A. The Ontology-Epistemology Divide: A Case Study in Medical Terminology. Proceedings of FOIS 2004. Amsterdam: IOS.
- [138] Ingenerf J and Linder R. Assessing applicability of ontological principles to different types of biomedical vocabularies. *Meth Inf Med* 2009; 48: 459-467.
- [139] Cheetham E et al. (2009) Using SNOMED CT in HL7 Version 3; Implementation Guide, Release 1.5, http://www.ihtsdo.org/fileadmin/user_upload/doc/tig/hl7/hl7_terminfo.html.
- [140] Garde S, Knaup P, Hovenga E, Heard S. Towards semantic interoperability for electronic health records. *Methods Inf Med*. 2007; 46 (3): 332-343.
- [141] The National Center for Biomedical Ontology. Protégé. <http://protege.stanford.edu/> .
- [142] Rector A, Drummond N, Horridge M, Rogers J, Knublauch H, Stevens R, Wang H, Wroe C. OWL Pizzas: Practical Experience of Teaching OWL-DL: Common Errors & Common Patterns, in: Motta E, Shadbolt N, Stutt A, Gibbins N (eds.): *Engineering Knowledge in the Age of the Semantic Web*, Lecture Notes in Computer Science 2004. Springer Berlin / Heidelberg: 63-81.
- [143] Schulz S, Seddig-Raufie D, Grewe N, Röhl J, Schober D, Boeker M, Jansen L (2012). Guideline on Developing Good Ontologies in the Biomedical Domain with Description Logics <http://purl.org/goodod/guideline> .
- [144] Sackett DL, Straus SE, Richardson WS, Rosenberg W, Haynes RB. *Evidence-Based Medicine: How to Practice and Teach EBM*. Second edition. 2003, Churchill Livingstone: Edinburgh.
- [145] Kitchenham BA, Dybå T, Jørgensen M (2004). Evidence-Based Software Engineering. ICSE '04 Proceedings of the 26th International Conference on Software Engineering: 273-281.
- [146] Boeker M, Jansen J, Grewe N, Röhl J, Schober D, Seddig-Raufie D, Schulz S. Effects of Guideline-Based Training on the Quality of Formal Ontologies: A Randomized Controlled Trial. *PLOS ONE*. Accepted for publication 2013.
- [147] Gruninger M, Fox MS. The role of competency questions in enterprise engineering. In: Proceedings of the IFIP WG5, 1994; volume 7, p. 212 <http://www.eil.utoronto.ca/enterprise-modelling/papers/benchIFIP94.pdf>.

[148] Hoehndorf R, Dumontier M, Gkoutos GV. Evaluation of research in biomedical ontologies. *Brief Bioinform.* 2012 Sep 8.

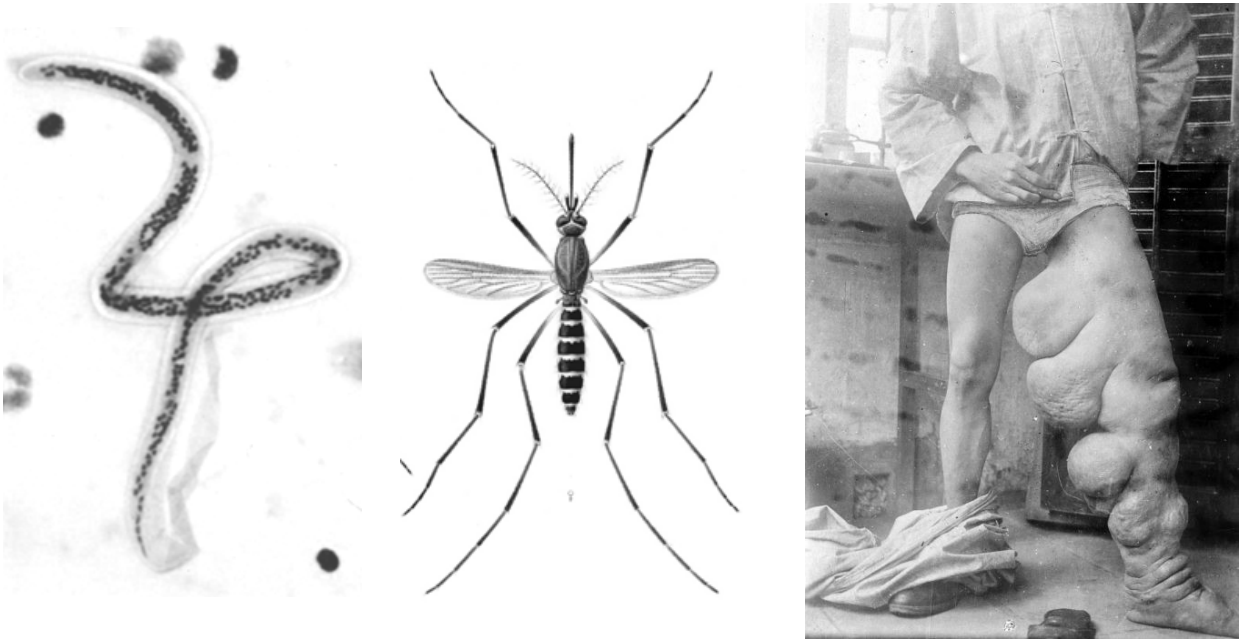


Fig.1 Filariasis: Microfilar (left), transmitting insect (centre), severe case of elephantiasis (right). Source: Wikimedia commons

Fig.2 Classes (left) and Relation (right) of the Upper Level Ontology BioTop

