

Strengths and limitations of formal ontologies in the biomedical domain

DOI: 10.3395/inciis.v3i1.241en



Stefan Schulz

Institute for Medical Biometry and Medical Informatics, University Medical Center Freiburg, Freiburg, Germany
steschulz@uni-freiburg.de



Holger Stenzhorn

Institute for Medical Biometry and Medical Informatics, University Medical Center Freiburg, Freiburg, Germany
holger.stenzhorn@uniklinik-freiburg.de

Martin Boeker

Institute for Medical Biometry and Medical Informatics, University Medical Center Freiburg, Freiburg, Germany
martin.boeker@uniklinik-freiburg.de

Barry Smith

Department of Philosophy and Center of Excellence in Bioinformatics and Life Sciences and National Center for Biomedical Ontology, University at Buffalo, Buffalo, USA
phismith@buffalo.edu

Abstract

We propose a typology of representational artifacts for health care and life sciences domains and associate this typology with different kinds of formal ontology and logic, drawing conclusions as to the strengths and limitations for ontology in a description logics framework.

The four types of domain representation we consider are: (i) lexico-semantic representation, (ii) representation of types of entities, (iii) representation of background knowledge, and (iv) representation of individuals.

We advocate a clear distinction between the four kinds of representation in order to provide a more rational basis for the use of ontologies and related artifacts in order to advance integration of data and enhance interoperability of associated reasoning systems.

We highlight the fact that only a minor portion of scientifically relevant facts in a domain such as biomedicine can be adequately represented by formal ontologies as long as the latter are conceived as representations of entity types. In particular, we show that the attempt to encode default or probabilistic knowledge using ontologies is prone to produce unintended, erroneous models.

Keywords

biomedical ontology, description logic, formal ontology, knowledge representation

Introduction

It is increasingly recognized that the complexity of the health care and life sciences domain demands consensus on the terms and language used in documentation and communication. This need is driven by the exponential growth of data generated by patient care and life science research. At the moment, this data cannot be fully exploited for integration, retrieval, or interoperability, because the underlying terminol-

ogy and classification systems (often subsumed under the heading “biomedical vocabularies”, see Table 1) are inadequate in various ways. Their heterogeneity reflects the different backgrounds, tasks and needs of different communities – including information technology communities – and creates a serious obstacle to consistent data aggregation and interoperability of the sort demanded by biomedical research, health care, and translational medicine.

Table 1 - Examples of biomedical vocabularies. Most of these vocabularies are made available through the Unified Medical Language System (UMLS) Metathesaurus, an umbrella system covering a broad variety of biomedical terminology systems (NLMb 2008, McCray et al. 1995)

Vocabulary	Purpose
ICD-9-CM/ICD-10 (WHO, 2008)	disease classification, in health statistics, hospital billing
WHO Drug Dictionary (UMC, 2008), ATC (WHOCC, 2008), RxNorm (NLMA, 2008) DM+D (NHS, 2008)	drug classification
NCI Thesaurus and Metathesaurus (NCI, 2008)	cancer research
LOINC (REGENSTRIEF INSTITUTE, 2008)	inter-laboratory communication
MedDRA (NORHTROP GRUMMAN, 2008)	medicine-related regulatory activities
DICOM (MITA, 2008)	medical image and imaging process descriptions
MeSH (NLM, 2008)	medical literature indexing
SNOMED CT (IHTSDO, 2008)	clinical documentation

Ontology and biomedical knowledge

What were formerly referred to as “terminology systems” or “vocabularies” are today often called by the name “ontology”. This term first became common in biology circles with the success of the Gene Ontology (GO), and has become increasingly popular in the medical domain as well. The so-called “omics” disciplines are a further major driving force for their development and adoption. Within this context, the Open Biomedical Ontologies (OBO) initiative, with over 60 ontologies at the moment and building on the successes of GO, is becoming a standard resource for the annotation of biomedical research data (SMITH et al., 2007).

But the term “ontology” itself is notoriously ambiguous (KUSNIERCZYK, 2006). Users often tend to have unrealistic expectations as to what ontologies can achieve (STENZHORN et al., 2009). Therefore, any use of this term must be preceded by an explanation of its intended meaning. To illustrate the sorts of problems which can arise, we can point to the stark contrast between sample definitions developed by computer scientists and the ones inspired by philosophers:

– Ontology (Computer Science): An ontology

defines (or specifies) the concepts, relationships, and other distinctions that are relevant for modeling a domain. The specification of an ontology takes the form of the definitions of representational vocabulary (classes, relations, and so forth) that provide meanings for the vocabulary and formal constraints on its coherent use (GRUBER, 1992).

– Ontology (Philosophy): Ontology is the study of what there is (QUINE, 1948). Formal ontologies are theories that attempt to give precise mathematical formulations of the properties and relations of certain entities (HOFWEBER, 2004).

Although these two definitions differ significantly, ontologies are seen in both cases as formal systems that apply fundamental principles and formalisms, drawing on mathematical logic to represent entities of certain kinds, whether on the side of mind and language (“concepts”) or on the side of reality (“properties”, “types”, and “classes”). The major role of ontologies is in both cases to provide a system of domain-independent distinctions to structure domain-specific theories with the goal of integrating and retrieving data and fostering interoperability. We are interested here only in ontologies

in which a formal approach is used to support an aim of this sort. To highlight this feature we shall use the term “formal ontology” in our deliberations henceforth. We believe that the focus on logical formalism most clearly distinguishes the new generation of biomedical ontologies — including SNOMED CT, and recent versions of the Gene Ontology (GO) — from their vocabulary-like predecessors, which still bear traces of their origins in the domain of library science and literature indexing.

In this paper we focus on the role formal ontology can play in resolving some of the problems caused by the heterogeneity of terminologies and classification systems used in the biomedical domain. We want to clarify how the representation of the entities studied by the life sciences can benefit from formal ontologies in a way that will help more adequately to capture domain knowledge. We address two important aspects which are too seldom dealt with explicitly: (i) the representation of meta- or background knowledge, and (ii) the relation of ontologies to human language. We seek to highlight the role played by these factors in developing and using formal ontologies. We also seek to clarify those situations in which domain knowledge cannot be adequately accounted for by formal ontologies, especially for reasons of vagueness and uncertainty. Two questions arise at this point:

– Which criteria can be used to delineate the types of knowledge that can sensibly be expressed by formal ontologies?

– How can the remaining types of knowledge be encoded to satisfy the demands of integration, retrieval, and interoperability?

We try to answer these questions focusing on representation standards developed by the Semantic Web community. We give examples of the use of this formalism for representing biomedical entities. We also point to some associated misconceptions and errors in ontology design and show how they can be rectified.

Informal representations

A simple universal representation scheme that lends itself to a representation of a broad range of entities and relations between them is given by the so-called Object – Attribute – Value (OAV) triples. This encoding scheme was already popular in early expert systems (SHORTLIFFE et al., 1975) and currently plays an important role in the Semantic Web initiative (W3C, 2008) where it is known within the framework of the Resource Description Format (RDF) under the heading of ‘Subject – Predicate – Object (SPO) triples’ (KLYNE et al., 2004). This triple-based representation is also very similar to the way the Unified Medical Language System (UMLS) Metathesaurus and other vocabulary resources link pairs of concepts by means of relations such as *broader_than*, *narrower_than*, *part_of*, *mapped_to*, *is_a*, and so on. Table 2 gives some examples of this kind of representation.

Table 2 - Examples of OAV representations

Concept / Term 1 (Object, Subject)	Relation (Attribute, Predicate)	Concept / Term 2 (Value, Object)
Aspirin	prevents	Myocardial_Infarction
Aspirin	is_a	Sacicylate
Aspirin	has_part	Aromatic_Ring
Blood_Plasma	narrower_than	Blood
Cancer	causes	Weight_Loss
Cell	has_part	Cell_Membrane
Contraceptive_Measure	prevents	Pregnancy
Diabetes_Mellitus	is_a	Frequent_Disease
Diabetes_Mellitus	has_prevalence	2.8%
Diclofenac	has_side_effect	Gastrointestinal_Bleeding
Diphtheria	is_a	Rare_Disease
ELM-2	interacts_with	LMO-2
ELM-2	is_a	Protein
Fever	symptom_of	Malaria_Tropica
Hand	has_part	Thumb
Hepatitis	has_location	Liver
Hepatitis	has_translation	Hepatite
Hypertension	is_a	Cardiovascular_Risk_Factor
Hyperthermia	has_synonym	Fever
Liver	is_a	Bodily_Organ
NaCl_Solution	has_part	Cl-_Ion
Pharyngitis	has_symptom	Hyperthermia

Smoking	causes	Cancer
THC	is_a	Schedule_III_Controlled_Drug
Thumb	has_part	Thumbnail
WHO	located_in	Geneva

One advantage of the triple format becomes evident when looking at this table. Simple assertions are represented in a way which comes close to the expressions used in human language. One disadvantage is that it promotes a confusion of use and mention (because it is asserted that one and the same thing, for example *Fever*, is both a *Synonym of Hyperthermia* and a *Symptom of Inflammation*). The triple format creates difficulties also when it comes to the formulation of more complex assertions such as “In 2008, diabetes mellitus had a prevalence of 18.3% of US citizens age 60 and older”. Such assertions need to be split into sets of simpler assertions if they are to fit the triple format. Table 3 depicts one possible OAV representation of such an assertion, where the successive rows are joined together in a compound conjunctive statement. One drawback here is that many concurring models of this kind can claim to represent the given statement equally well, and this creates forking. Different modelers effectuate the needed encodings in different ways, and as a result their information systems are no longer marked by interoperability. To avoid this silo effect, a single uniform representation model is needed.

Table 3 - OAV triplet representation of the complex statement: “In 2008, diabetes mellitus had a prevalence of 18.3% of US citizens age 60 and older”

Prevalence_1	instance_of	Prevalence
Prevalence_1	has_date	2008
Prevalence_1	has_value	0.183
Prevalence_1	has_population	Population_1
Prevalence_1	has_disease	Diabetes_Mellitus
Population_1	instance_of	Population
Population_1	has_minimum_age	60
Population_1	has_habitat	USA

Another drawback of the OAV representation scheme is that it is not obvious, in any given case, how its assertions are to be interpreted. The assertion that *Smoking causes Cancer*, for example, could be interpreted in such a way that its author believes that smoking *always* (i.e., without exception) causes cancer. But it could also be interpreted to mean that smoking often, usually, or typically causes cancer, or even (as within the UMLS Semantic Network), that the expression

“Smoking causes cancer” is semantically meaningful. Without additional knowledge about how to interpret the relation *causes*, we cannot decide which alternative is meant in any given case. Certainly, in many everyday situations humans communicate perfectly well when using such ambiguous statements. But this is so because humans are able to associate them spontaneously with a relevant context of implicit background assumptions. In the case of machine processing such implicit knowledge is lacking, and it is for this reason that logical definitions and axioms expressed in an appropriate formal language are required to preclude, or at least constrain, competing interpretations. Unfortunately, as will be clear from the examples given below, application of the rigor of logic is not only very demanding of human resources, it is also such that it does not even in principle allow the formal expression of everything we know. We can however still capture an important part of our knowledge in a way that is, we believe, indispensable to computational reasoning and to advance integration, retrieval, and interoperability.

Formal representations

To illustrate how basic ontological assertions concerning the entities in a given domain can be formulated using logical resources, we introduce the family of Description Logics (hereafter DLs) (BAADER et al., 2007). DLs are subsets of first-order logic (hereafter FOL). Although DLs are far from being able to express everything one might desire from a comprehensive logically based ontology (for this one would require at least the full range of FOL), we set them as our focus here for the following reasons:

- DLs have recently become a standard for representing domain knowledge in the context of the Semantic Web, and OWL DL, the DL subtype of the Web Ontology Language (OWL) (PATEL-SCHNEIDER et al., 2004) has been developed and standardized by the World Wide Consortium (W3C).
- DLs have a large user base and are supported by a variety of software tools, such as the Protégé editor (BMIR, 2008). OWL DL also supports the use of reasoning engines like Pellet (SIRIN et al., 2007) and FaCT++ (TSARKOV et al., 2006), which allow algorithmic checking of the consistency of given inputs and the inference of new assertions.
- DLs have certain favorable computational properties. Above all, many of them are decidable, i.e. algorithms exist for which it is guaranteed that they will always return some result. This is the reason why

DLs are preferred for many purposes over FOL, which is considerably more expressive than DLs but falls short of decidability.

- DLs have increasingly been employed in biomedical terminologies. After the GALEN project in the nineties (Rector, 1997), a pioneering effort in the large-scale use of a logic-based formalism for medical domain representation and reasoning, the currently most significant example of the use of DLs is the clinical terminology SNOMED CT (IHTSDO, 2009), which contains more than 300,000 classes. OWL DL is also increasingly used as a representation language for OBO Foundry ontologies (SMITH et al., 2007).

To use DLs properly, one has to understand their basic building blocks, represented by terms like “class”, “relation”, and “individual”, and also understand how their constituent logical symbols and expressions are interpreted. For example, all past, present, and future individual hands in the world are instances of the class *Hand*. Binary relations (“object properties” in OWL DL) have pairs of individuals as their extensions (PATEL-SCHNEIDER et al., 2004), e.g., the pair constituted by the first author’s right thumb and his right hand. Classes in DL are always distinct from individuals; classes of classes are not allowed. OWL DL object properties express binary relations without any direct reference to time. This is a major drawback from an ontological (and biological) point of view¹, since we often need to attach time-indexes to assertions about individuals. For example it is necessary for many purposes to make it explicit that an individual belongs to the class *Embryo* at t_1 and to the class *Fetus* at t_2 .

One also has to be careful to recognize that the same expressions may be interpreted in different ways in different disciplines. For instance, a statement to the effect that all hands have thumbs is limited to the domain of normal (so-called canonical) human anatomy. It clearly does not hold if the domain includes injured or malformed humans, or humans in early embryonic states (SCHULZ et al., 2008; NEUHAUS et al., 2007).

In the following, we illustrate the DL syntax and semantics through a set of increasingly complex examples. To start, we take a look at the class *Liver*. When we introduce this class, we define its extension to be the set of all livers of all organisms at all times. In the same vein, the class *Bodily_Organ* then extends to all individual bodily organs at all times. To relate the two classes, we introduce the key concept of taxonomic subsumption: The class *Liver* is a subclass (subtype) of the class *Bodily_Organ*. In DL notation, this is expressed by the \sqsubseteq operator:

$$Liver \sqsubseteq Bodily_Organ$$

and the relation in question is commonly referred to as the *is_a* relation.

In contrast, the instantiation relation *instance_of* (\in) links individuals to the classes of which they are the instances. For example, each individual liver is an instance of the class *Liver*, so the (individual) liver

of the first author of this paper is one specific *instance_of Liver*. It is noteworthy that DLs do not allow a distinction to be expressed between an individual’s membership in a class on the one hand and an individual instantiation of a universal or type on the other hand. Both are represented by means of the *instance_of* relation (\in).

More complex statements can be obtained by using operators and quantifiers. In the following example we use both the \sqcap (“and”) operator and add a quantified role, using the existential quantifier \exists (“exists”). The expression

$$Inflammatory_Disease \sqcap \exists has_location.Liver$$

then denotes the class of all instances that belong to the class *Inflammatory_Disease* and are further related through the relation *has_location* to some instance of the class *Liver*.

This example actually gives us both the necessary and the sufficient conditions needed in order to fully define the class *Hepatitis*:

$$Hepatitis \equiv Inflammatory_Disease \sqcap \exists has_location.Liver$$

The equivalence operator \equiv in this formula tells that: (i) each particular instance of hepatitis is an instance of inflammatory disease that is located at some liver, and also (ii) that everything that is an instance of inflammatory disease that is located at some liver is an instance of hepatitis. Hence, in any situation, the term on the left can be replaced by the expression on the right without any loss of meaning.

Note that when we express such an equivalence statement, this statement has to hold at all times without exception. Therefore we cannot use statements of this form to express, for instance, that hepatitis has the symptom *fever in most (but not in all) cases*. We could, of course, form the expression

$$Hepatitis \equiv Inflammation \sqcap \exists has_location.Liver \sqcap \exists normally_has_symptom.Fever$$

In virtue of the DL interpretation of the existential quantifier, however, this assertion implies that for every instance of the class *Hepatitis* (without exception) there also exists some instance of *Fever*. The word *normally* in the property name *normally_has_symptom* can be interpreted by humans, but it plays no logical role at all. This is clearly not in accordance with the intended meaning.

Such logical effects are important, since errors arise when they are not taken into account by users of DL formalisms. Abundant instances of such errors can be found in the current version of SNOMED CT. Its concept *Biopsy_Planned* (ID: 183993008), for example, is related to the concept *Biopsy* as follows:

$$Biopsy_Planned \sqsubseteq Situation \sqcap \exists associated_procedure. \\ Biopsy \sqcap \dots$$

This expression states that for each planned biopsy (we assume that this is the meaning of *Biopsy_Planned*) there always exists at least one instance of an actual biopsy, which certainly cannot be what is intended, since not all plans for biopsies are realized. SNOMED CT also has the class *Drug_Abuse_Prevention* (ID: 408941008):

$$Drug_Abuse_Prevention \sqsubseteq Procedure \sqcap \exists has_focus. \\ Drug_Abuse$$

Correctly interpreted, this expression states, quite absurdly, that whenever an act of drug abuse prevention is performed, then it is related to some instance of drug abuse.

These two examples illustrate how easy it is to create statements with unintended meanings when using even very simple DLs. The reason such examples are so common in current biomedical terminologies is that the ontology developers are often domain experts who are not familiar with the complexities of formal logic and pay too little attention to the principles of sound ontology development. They tend to be guided, rather, by the superficial simplicity of such statements, and thus do not realize that their logical interpretation contradicts the intended meaning. The resultant invalid statements then provide support for invalid inferences when used in automated reasoning.

It is clear however that there is a need for classes such as *Biopsy_Plan* or *Drug_Abuse_Prevention*. Because 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. This means that the quantifier \forall is used in such a way as to specify the allowed range for a given relation. We could then (correctly) state the following:

$$Biopsy_Plan \sqsubseteq Plan \sqcap \forall has_realization. Biopsy$$

In plain words, this expression states that a biopsy plan is a plan that – *if realized* – can be realized only by some instance of *Biopsy*. In contrast to the simple existential statements, this does not say that some *Biopsy* must exist for each *Biopsy_Plan*. Similar constructs are needed for other realizable entities, such as functions, roles, and dispositions (GRENON, 2003).

By using the universal quantifier \forall , however, we move away from simple but scalable DL dialects like *EL* (BAADER et al., 2007) to DLs with a computational complexity that poses severe problems for large ontologies like SNOMED CT. It is even more complicated to define classes like *Drug_Abuse_Prevention* with the appropriate logical rigor. Here we need to say that if some abuse prevention procedure is applied, then this causes a state in

the organism that precludes the organism to participate in *Drug_Abuse*. So in order to express this properly, we need to introduce the negation operator \neg as follows:

$$Drug_Abuse_Prevention \sqsubseteq Procedure \sqcap \exists has_participant. \\ Person \sqcap \exists causes. (State \sqcap \exists has_participant. (Person \sqcap \\ \exists participates_in. \neg Drug_Abuse))$$

In this definition the class *Person* is instantiated twice. Unfortunately it is not specified that the two instances are identical, which they would have to be for the assertion to do its job. There is, however, no DL able to express the fact that they are identical, for this would require the full expressive powers of FOL with equality, and thus a move beyond the realm of decidability.

Other cases of medical terms that exceed the expressiveness of decidable description logics include expressions involving “without”, such as “concussion of the brain without loss of consciousness” as discussed in (BODENREIDER et al., 2004; CEUSTERS et al., 2007; SCHULZ et al., 2008). Such expressions are highly relevant and important in medicine. Yet their representation is intricate, not only due to their demand for expressive logical constructors, but also due to the difficulty of gaining univocal agreement upon their meaning, which would involve taking into account tacit assumptions for example relating to time.

The above examples clearly demonstrate the dilemma of logic-based representations: If the purpose is to logically encode and classify large terminological systems like SNOMED CT (BAADER et al., 2006), then the set of allowed constructors must be limited, since value restrictions and negations lead to computational intractability. Some (RECTOR et al., 2008) nonetheless stress that it is important to include even computationally more expensive constructs so that adequate domain representations are not precluded. An alternative strategy is to distinguish the constructs contained within the terminology from their use in specific sentential contexts, where negation and other terms (such as “on examination”) are properly at home.

Categories of domain representation

As should by now be clear, it is often not possible with computable, logic-based domain representation formalisms, like DLs, to truthfully represent important aspects of biomedical knowledge. Many types of assertions require other means of representation. We thus propose to distinguish between different categories of domain representation. We shall show that these call for distinct sorts of treatment even though they have often been treated within formal ontologies as if they were similar. Our interest in keeping these categories apart is to highlight the fact that each representation requires its own formalisms with its own semantics and that inadequate use of undifferentiated representation formalisms leads to unwanted results. As a result of our discussion, we also aim to contribute to a more clear-cut understanding of what formal ontologies can and cannot accomplish in the biomedical domain.

Lexico-semantic representation

We use the term “lexico-semantic representation” to refer to thesauri, semantic lexicons and similar representational artifacts centered on the meanings of the expressions found in natural language. Typically, they address both the fact that one lexical entry may have two or more meanings (as illustrated by the polysemy of terms such as “patient” or “lead”), and the fact that one meaning may be expressed by one or more lexicon entries (for example the synonymy of “hyperthermia” and “fever”). They may also contain word or term translations. Thesauri and semantic lexicons may further contain semantic relations between the individual lexicon entries such as *broader_than* or *narrower_than*. WordNet (FELLBAUM, 1998), MeSH, and most parts of the UMLS Metathesaurus (NLMB, 2008) are examples for such representation systems, which have a long tradition in library science, where literature retrieval as a widely accepted use case.

The question of how lexico-semantic relations such as synonymy should be correctly expressed is not in fact an issue to be addressed by ontologies, as ontologies describe real entities independently of the symbols and formalisms of human language used to denote them. Therefore, relations such as *broader_than* or *narrower_than*, which are semantically arbitrary sub-classification relations (OBRST, 2006) characterizing the MeSH thesaurus or WordNet, are fundamentally different from the subclass (*is_a*) relation that defines the taxonomy backbone of a properly constructed ontology. 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 are fundamentally different. In the first case, we are dealing with a parthood (*part_of*) relation, in the second with a case of the subtype (*is_a*) relation. This difference may not matter in the relevant context since the *narrower_than* relation, even though semantically ill-defined, 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 fetal blood. It may, however, be vitally important in other contexts.

Problems arise already at the present stage of information retrieval when it is proposed to “ontologize” MeSH simply by mapping all *narrower_than* relations to taxonomic subsumption relations (SOUALMIA et al., 2004) such as *Plasma* \sqsubseteq *Blood* and *Fetal_Blood* \sqsubseteq *Blood*. For while the result is a seemingly perfect subsumption graph that can be easily processed by standard DL tools, this exercise once again creates typical cases of unintended models, since it ignores the true meaning of subsumption. Errors such as classifying plasma as a kind of blood are then the result.

While lexico-semantic relations have certain features in common with the ontological relations between entities in reality, the construction of an ontology out of a thesaurus requires numerous additional assumptions, for example concerning quantification. Hence,

an automated conversion process cannot 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 (SCHULZ et al., 2001).

Although we see lexicons or term lists as lying outside the realm of formal ontology, we want to stress that virtually all formal ontology applications require a link between ontology classes and lexical items. However, we advocate that these two issues should be treated by the two separate artifacts of formal ontologies on the one hand, and lexico-semantic representations on the other.

Representation of types of entities

Scientific realism postulates the existence of an objective reality that can be studied by science and about which we can discover truths (BOYD, 2002). A proper scientific theory, and hence a proper ontology, will include assertions to the effect that entities instantiating a given class stand in given relations to entities instantiating some other given class. Such assertions can of course rest on error and thus they must be capable of being revised at every stage. When formalized logically they may in addition rest on metaphysical presuppositions of different sorts, for example theories based on three- and four-dimensionalist approaches. But scientific realism as described here is compatible with a wide range of such theories. While the realist view is still controversial and not shared by all ontology developers (SMITH et al., 2006), it has a number of practical advantages. Thus, for example, it allows a view of ontologies as providing a canon of axiomatic assertions about simple relations between the most scientifically basic types of entities, which can then be taken for granted in further, more complex types of work. Examples of such assertions are “cells have membranes”, “hearts have cavities”, “every case of hepatitis is located in a liver”, “every aspirin tablet contains salicylate”, and so on.

It is useful to produce artifacts that will afford computationally amenable automated reasoning on the basis of such assertions, as demonstrated above. However, this is not identical with the attempt of producing formal theories that aim at characterizing a domain in reality. In practical ontology engineering, these two objectives have to be reconciled. Experience in use of the Gene Ontology supports the thesis that features of reality can often be sufficiently well represented even through a relatively simple logic. However, as will be clear from our discussions of DLs above, we must thereby always bear in mind that such formalisms do not possess the richness necessary to create complete definitions in many cases. The necessary expressiveness conflicts with the need to construct computationally tractable models. It must therefore be accepted that ontologies (like scientific theories) provide only partial representations of reality. They state what is considered to be true of all instances of given classes: “There is no hepatitis outside the liver”; “there is no NaCl solution without chloride ions”; “there

is no cell without a cell membrane”. But it becomes quite clear that such statements constitute only a minor portion of the knowledge that may be required to adequately capture a domain. As Rector (2008) expresses it, “There are very few interesting items of knowledge that are truly ontological in this strict sense.” Yet it is also evident that such items are nonetheless crucially important since they form the basis for all reasoning both by human beings and by computer applications.

Furthermore, it has been largely ignored that domain representation of this kind (statements about what is true of all instances of a class) are also present in numerous artifacts that are seldom identified as ontologies. UniProt, a large, central repository (“database”) of protein data (UNIPROT, 2008), is a typical example. Under ontological scrutiny, most of its content describes protein types (and not individuals) in terms of what is universally true for every single protein molecule of this type. We therefore consider these kinds of representation, too, as essentially ontological in nature.

Representation of background knowledge

The term “background knowledge” as used by Rector (2008) encompasses default knowledge, presumptive knowledge, and probabilistic knowledge, and refers to all kinds of statements that are assumed to be at least typically (but not necessarily universally) true in some domain and in some context. Such knowledge is traditionally conveyed through scientific textbooks in a highly context-dependent fashion, often invoking prototypical assertions, for example, concerning the relationship between diseases, signs and symptoms, or between adverse effects and drugs, which are expressed in terms of qualitative probabilities.

Familiarity with this background knowledge, rather than familiarity with the knowledge that can be conveyed using formal ontologies, distinguishes an expert from a novice, just as it marks the difference in content between a textbook and a dictionary. The examples below highlight how formal ontology approaches and logical representation formalisms reach their limits when it comes to representing this kind of knowledge. Using DL-based formalisms for even simplified accounts of prototypical knowledge tends to lead to flawed results. There exist other logical formalisms capable of expressing this kind of knowledge, but again those formalisms are computationally expensive in ways which go beyond the bounds of decidability.

Default knowledge

One type of background knowledge is default knowledge (Rector, 2004; Hoehndorf et al., 2007). This is knowledge concerned with what can be assumed to be typically true in the absence of contravening evidence. DL does not give us the means to state what is typically true. But especially with regard to canonical anatomy vs. clinical anatomy (SMITH et al., 2005), one would like to state for example that hands normally have thumbs. A statement such as

$$Hand \sqsubseteq \exists has_proper_part.Thumb$$

does not appropriately account for this. It states that every hand has a thumb and rules out the possibility of a hand without a thumb; that is, it rules out non-prototypical hands (e.g. after accidents).

Meta classes

Other statements of background knowledge are meta-statements concerning classes. They hold true when viewed as assertions about classes as wholes, but become false when viewed as assertions about instances. The DL view is that all statements about classes *are* statements about the corresponding sets of instances. When this is ignored, seemingly obvious subsumption statements like:

$$Diabetes_Mellitus \sqsubseteq Frequent_Disease$$

$$Malnutrition_Related_Diabetes_Mellitus \sqsubseteq Diabetes_Mellitus$$

lead to false conclusions, for example that

$$Malnutrition_Related_Diabetes_Mellitus \sqsubseteq Frequent_Disease$$

The problem here is here that a feature concerning the size of the extension of a class (the number of its instances) is erroneously taken as an inheritable property. In the above, the symbol \sqsubseteq (*is_a*) is used in two logically distinct senses, of which only the second is sanctioned by DLs. The resultant so-called *is_a overloading* has been identified as a typical error that occurs when building ontologies in an unprincipled way (GUARINO, 1999; Welty and GUARINO, 2001; SMITH et al., 2004). It involves indiscriminately interpreting natural language expressions containing “is a” as representing a single ontological subclass relation.

Dispositions

Encoding non-trivial facts in formal ontologies may require complicated additional constructs, such as the addition of representations of dispositions to convey information about potentialities. It is important to note that dispositions can exist without ever being realized and even without our being able to specify the precise conditions in which they are realized (JANSEN, 2007). An analgesic drug is a substance that has a disposition to treat pain. But it will realize this disposition only when administered in a certain way to a certain sort of patient. We can represent the class of processes of treating (a patient) for pain with:

$$Treating \sqcap \exists has_participant.Pain$$

We can then represent the class of dispositions realized when pain is treated:

$$Disposition \sqcap \forall has_realization.(Treating \sqcap \exists has_participant.Pain)$$

The following definition now declares an *Analgesic_Drug* to be a substance in which this disposition inheres:

$$\text{Analgesic_Drug} \equiv \text{Substance} \sqcap \exists \text{ bearer_of.}(\text{Disposition} \sqcap \forall \text{ has_realization.}(\text{Treating} \sqcap \exists \text{ has_participant.Pain}))$$

Such constructions can strongly affect the scalability of an ontology implementation, since a larger set of such expressions, e.g., for representing the pharmacodynamics of substances, cannot be handled efficiently by current reasoning algorithms.

Data in context

The body of scientific and clinical assertions is not restricted to the expression of default assumptions and dispositional features. It also includes uncertain assertions, for instance, concerning the effect of a drug in treating a given disease, or concerning the existence of a suspected risk factor for a certain condition. For the aforementioned reasons, the encoding of such assertions in formal ontologies can be very demanding and it is actually questionable whether such assertions should be encoded in a formal ontology in the first place.

As an example, an ontology is being created in the context of the European Union project @neurIST as a basis for the semantic mediation and integration of data in the area of brain aneurysms and *subarachnoidal* bleedings (BOEKER et al., 2007). The data within the project originate from a multitude of sources and show a high degree of fragmentation and heterogeneity both in format and scale. The ontology needs to represent all relevant types of entities and also respect the different views of these entities on the part of those in different disciplines such as medicine or epidemiology. To do justice to all these aspects, the ontology applies dispositional statements in the formulation of class definitions and is split into two parts: (i) an ontology in the proper sense of the word and (ii) a set of representational artifacts capturing context-specific knowledge about certain facts, e.g., risk factors in clinical contexts. (A similar approach is also pursued by the Ontology of Biomedical Investigations (OBI) (OBI, 2008).) In the @neurIST ontology, the class *Hypertensive_Disease* is a subclass of *Biological_Process_or_State* that is associated with *High_Blood_Pressure* and *causes* some *Rupture_Disposition*, i.e., a disposition to the effect that an aneurysm will burst. This disposition is then further connected to the class (and thus identified as a) *Risk_Factor_for_Aneurysm_Rupture* in that this latter class is also defined to be such that its instances cause some instance of *Rupture_Disposition*:

$$\text{Rupture_Disposition} \equiv \text{Predisposition_to_Disease} \sqcap \forall \text{ has_realization.Aneurysm_Rupture}$$

$$\text{Risk_Factor_for_Aneurysm_Rupture} \sqsubseteq \text{Risk_Factor} \sqcap \exists \text{ causes.Rupture_Disposition}$$

The following assertion is crucial to the study of aneurysms but transgresses the limits of formal ontology. It is incomplete in the sense that the constraints that are contextually defined and which make this statement valid are missing:

$$\text{Hypertensive_Disease} \sqsubseteq \text{Risk_Factor_for_Aneurysm_Rupture}$$

The above is an attempt to state that hypertensive disease is in some *generic* sense a risk factor. Hypertensive disease is a risk factor for cerebral aneurysms, but only under certain circumstances. What we want to express is capture the precise nature of the correlation between the two while of course recognizing that there are other risk factors as well.

These examples show the sorts of steps which would have to be taken in order for a DL framework to be extended in such a way as to account for certain kinds of background knowledge, thus gaining the advantage of DL reasoning support without incurring the risk of unintended models.

However, the difficulty of representing all the hidden assumptions underlying background knowledge (and the performance problems that result from using the needed rich logic) may suggest that we use instead a much simpler triple-based representation as mentioned in the introductory section, and devise special reasoning devices to fit. Alternatively one might resort to a broad range of knowledge representation artifacts such as default logics (REITER, 1980), frames (MINSKY, 1974), F-logic (KIFER et al., 1989), and several kinds of computationally expensive DL extensions (BAADER, 2007, ch. 6). The resultant knowledge representation artifacts, however, are not formal ontologies as we use this term. Still, one can and should reuse the classes formally defined in an ontology as symbols in these formalisms, along the lines pointed out in our examples above.

Representation of individuals

Whereas the first three types of representation described above make generalizations about all entities of some given kind, much of medicine involves descriptions of individual entities, such as the tumor of patient A, a certain lab test performed on a certain day in hospital B, a treatment episode as documented in patient record C, or the occurrence of a specific disease in a given patient group D. Disciplines like epidemiology and public health deal similarly with specific political and geographical entities such as *Brazil*, *New Orleans*, *the Southern Pacific Islands*, or *the upper Rio Negro region*.

Statements of individual facts can be expressed in a straightforward manner in DL terms as instantiations of corresponding classes, or in other words, as so-called A-box assertions (with the letter “A” standing for “assertions”), as contrasted with the T-box component of DLs which capture what is called “terminological knowledge” (or perhaps better called “knowledge pertaining to types”). Consider for example,

Hepatitis_162726 \in *Hepatitis*

which asserts that a particular disease (case) is an instance of hepatitis.

A molecular interaction statement such as “Lmo-2 interacts with Elf-2” as typically found in a scientific article typically rests on assertions about certain individuals, namely two portions of *Lmo-2* and *Elf-2* (containing instances of molecules of the corresponding types), that have been observed to exhibit some interaction in some specific experimental assay (SCHULZ et al., 2008).

We can describe a particular interaction event in which the two substance portions under scrutiny participate as follows:

Lmo-2.7760102 \in Portion_of_Lmo-2
Elf-2.776010 \in Portion_of_Elf-2
Interaction.725322 \in Interaction

has_participant (Interaction.725322, Lmo-2.7760102)

has_participant (Interaction.725322, Elf-2.776010)

There are domains like geography, in which individuals, not classes, constitute the primary targets of knowledge. Any detailed description of geographic or political divisions of the sort that would be of interest, for example, for epidemiology or public health, abounds in references to particular entities which instantiate only a small number of classes (SMITH et al., 2005). For instance, a complete political division of the U.S. can be created on the basis of four nested levels (with one instance of countries, 50 instances of states, 3,077 instances of counties, and over 50,000 instances of municipalities) (see also geographic entities in GAZ (GENOMICS STANDARD CONSORTIUM, 2008)). Note the difference in representation compared to anatomical divisions in Table 4.

Table 4 - Example partonomies in geography and anatomy

<i>Orlando</i> \in <i>Municipality</i>	<i>Thumb</i> \sqsubseteq <i>Digit</i>
<i>Orange County</i> \in <i>County</i>	<i>Hand</i> \sqsubseteq <i>Body_Part</i>
<i>Florida</i> \in <i>State</i>	<i>Upper_Extremity</i> \sqsubseteq <i>Limb</i>
<i>USA</i> \in <i>Country</i>	<i>Body</i> \sqsubseteq <i>Anatomical_Structure</i>
\langle <i>Orlando, Orange County</i> \rangle \in <i>proper_part_of</i>	<i>Thumb</i> \sqsubseteq \exists <i>proper_part_of</i> . <i>Hand</i>
\langle <i>Orange County, Florida</i> \rangle \in <i>proper_part_of</i>	<i>Hand</i> \sqsubseteq \exists <i>proper_part_of</i> . <i>Upper_Extremity</i>
\langle <i>Florida, USA</i> \rangle \in <i>proper_part_of</i>	<i>Upper_Extremity</i> \sqsubseteq \exists <i>proper_part_of</i> . <i>Body</i>

This example shows that assertions concerning classes differ formally from assertions about individuals. In DL, however, the employed relations are the same, because DLs do not allow special relations that relate classes. Logically relating classes always requires the use of quantifiers, which are not needed in assertions relating individuals. This explains why, prior to any logic-based representation, it must be made clear whether the entities under scrutiny are classes or individuals. But especially in the field of molecular biology this is not trivial at all. Thus, our assertion example “*Lmo-2 interacts with Elf-2*” can be perfectly well understood as a universal statement concerning the class of *Lmo-2 molecules* and thus as expressing dispositional knowledge along the lines of:

All *Lmo-2* molecules have the disposition to interact with *Elf-2* molecules.

There are good arguments to be made on behalf of either reading, and so disambiguation cannot be affected without first analyzing the context in which the utterance is being made.

In practice, the individual/class boundary is often drawn in an idiosyncratic way. For example, UniProt

entries are asserted to denote “instances” of the class protein. A computer scientist might contend that this choice of terminology is mainly motivated by the view the modeler has of a domain: “Deciding whether a particular concept is a class in an ontology or an individual instance depends on what the potential applications of the ontology are.” (NOY and MCGUINNESS, 2001). France, on this reading, may be conceived either as a class or an instance depending upon the needs of particular ontology developers. We believe, however, that no arbitrariness should be involved in the distinction between this particular cell in this particular test tube here and now (an instance), and *Cell* (a class). Moreover, encouraging the supposition that there is such arbitrariness has the potential to lead to a forking of representations which will hamper the very interoperability of data resources that ontologies are intended to support.

Indeed we contend that a formal ontological analysis can be coherent only on the basis of a view of the distinction between individuals and classes as an unalterable distinction obtaining on the side of the entities themselves. Individuals on the one hand exist in space and time; they do not stand to each other in subsumption relations; they can be referred to by proper names and (in many cases)

photographed. Classes on the other hand do not exist in space and time; they stand in subsumption relations; and they can be referred to by common nouns. Whether an entity is a particular or a class or type is thus not a matter of choice on the part of modelers, and, in our experience, the controversial cases which seem to suggest such optionality always reveal upon closer inspection hidden ambiguities or inadequacies of software frameworks. Some defenders of the view that the human *MPDU-1* gene is an instance of the class *Gene*, refer to genes as instances of information content entities. The same genetic information entity can be encoded in different nucleic acid macromolecules, just as the same text can be disseminated in many hard copies. Others, however, claim that the human *MPDU-1* gene is not an instance but a subclass of the class *gene*; they are then referring to genes as types of macromolecular sequences, the instances of which are the real nucleotide sequences replicated in the cells of our bodies.

As we already saw in the section on background knowledge, an implicit reference to individuals underlies probabilistic statements, which are very characteristic in biomedical discourse. An example is the following statement: “In 2000, worldwide prevalence of diabetes mellitus was 2.8%”. Here we have two classes, *viz.* *Human* and (*case of*) *Human_Diabetes*. Both classes have a cardinality (integer value), and the prevalence is given by their ratio. The prevalence is therefore not a characteristic of the disease but of the population of persons who have a case of the disease. We here extend the DL notation by symbolizing the cardinality of the extension of a class (i.e., the number of instances) by enclosing the class name in “| |”.

$$Human \sqsubseteq Object$$

$$Diabetic_Human \equiv Human \sqcap \exists \text{ bearer_of.Diabetes_Mellitus} \\ |Diabetic_Human|/|Human| = 0.028$$

This demonstrates that probabilistic background knowledge could be expressed by DL A-boxes extended by arithmetic operators (referring to individuals). This is however not in the scope of formal ontologies, just as little as are alternative approaches such as probabilistic T-Box extensions (KOLLER, 1997; KLINOV, 2008). Moreover, the assertions in question cannot be expressed with currently available DL resources.

Discussion and Conclusion

The discipline of knowledge representation evolved in the context of artificial intelligence research 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 artifacts. However, two strands of research have evolved since, demonstrating the need for a more principled methodology.

First, Description Logics (DLs) were developed by delineating and investigating fragments of First Order

Logics (FOL) that are sufficiently expressive to allow the formulation of assertions about classes of individuals as well as their relations in such a way that new theorems could be derived automatically. This required a well-defined semantics calling basically for a bipartition into classes and individuals; it demanded also a formal account of subsumption and of role quantification. While in more primitive, semantic network style representations such as the UMLS Metathesaurus, all statements – such as “aspirin is a salicylate”, “aspirin contains an aromatic ring”, and “aspirin prevents myocardial infarction” – look quite similar, attempts at more formal representation reveal fundamental differences. Using description logics, the first statement is straightforward and does not require any relation beyond that of subclass, the second requires a quantified role expression, whereas the third cannot be adequately represented at all.

Secondly, in parallel to the evolution of representational languages like OWL, philosophers and computer scientists confronted the history-laden discipline of philosophical ontology with the requirements of the modern information society and created the discipline of applied ontology (GUARINO, 1998). Biomedicine became a testbed for the convergence of DLs and applied ontology. The OBO Foundry effort, and increasingly the redesign activities of SNOMED CT, bear witness thereto.

We can now summarize the results of this paper by means of the crude delimitation of four kinds of representation we have introduced above, namely: (i) lexicosemantic representation, (ii) representation of types of entities, (iii) representation of background knowledge, and (iv) representation of individuals.

(i) These are the sorts of statements we find in much of the UMLS, as well as in WordNet and similar artifacts, which strive to represent the terminological component of a domain. They do this by means of relations such as *synonymy*, *polysemy*, *broader than*, *narrower than*, drawn from the realm of thesauri and semantic lexicons. We argue that this approach is useful for information retrieval but not for inference or for knowledge integration.

(ii) At the opposite extreme are the sorts of statements we find in formal ontologies formulated in logics, where formal rigor and inferential power is achieved at the price of constraints on expressiveness along a number of dimensions. These constraints may fall short of meeting the requirements of users who often expect from a domain ontology more than a repository of basic truisms. On the other hand, such truisms are indispensable as a foundation for the more adequate formulation of other sorts of statements, not least in the context of reasoning systems.

(iii) This group of statements constitutes what we call “background knowledge”, a matter of loose associations between classes which cannot be expressed by the “for all ... some” statement scheme typical of DLs. These can to some degree be ontologized by introducing classes of dispositions and other realizable entities. However their introduction occurs at the price of increased complexity. There are other approaches to

the representation of background knowledge, including default logic (REITER, 1980), frames (KIFER et al., 1989), and several kinds of computationally expensive DL extensions (BAADER, 2007, ch. 6). A general recommendation cannot be made as to which of these or other alternatives is appropriate however. This depends heavily on the specific application domain and the specific use case for which reasoning services are needed.

(iv) The final set of statements concerns the representation of individuals. This might be seen as a

minor issue in, for instance, yeast biology, but is of great importance in a domain such as medicine, which is concerned to record information about human beings and populations. We showed, for example, that probabilistic statements concerning disease prevalence are assertions not about classes but rather about individuals.

Table 5 recapitulates the examples given in Table 2 at the beginning of the article and assigns each of them to one of the different categories of knowledge we introduced above.

Table 5 - UMLS Metathesaurus-style (mrrel table) assertions and associated domain representation categories

Concept / Term 1 (Object, Subject)	Relation (Attribute, Predicate)	Concept / Term 2 (Value, Object)	Domain representation Category
Aspirin	prevents	Myocardial_Infarction	BK
Aspirin	is_a	Sacicylate	ONT
Aspirin	has_part	Aromatic_Ring	ONT
Blood_Plasma	narrower_than	Blood	LS
Cancer	causes	Weight_Loss	BK
Cell	has_part	Cell_Membrane	ONT
Contraceptive_Measure	prevents	Pregnancy	BK
Diabetes_Mellitus	is_a	Frequent_Disease	BK
Diabetes_Mellitus	has_prevalence	2.8%	BK
Diclofenac	has_side_effect	Gastrointestinal_Bleeding	BK
Diphtheria	is_a	Rare_Disease	BK
ELM-2	interacts_with	LMO-2	BK, INS
ELM-2	is_a	Protein	ONT
Fever	symptom_of	Malaria_Tropica	BK
Hand	has_part	Thumb	ONT
Hepatitis	has_location	Liver	ONT
Hepatitis	has_translation	Hepatite	LS
Hypertension	is_a	Cardiovascular_Risk_Factor	BK
Hyperthermia	has_synonym	Fever	LS
Liver	is_a	Bodily_Organ	ONT
NaCl_Solution	has_part	Chloride_Ion	ONT
Pharyngitis	has_symptom	Hyperthermia	BK
Smoking	causes	Cancer	BK
THC	is_a	Schedule_III_Controlled_Drug	BK
Thumb	has_part	Thumbnail	ONT
WHO	located_in	Geneva	INS

(BK = background knowledge, INS = instances, LS = lexico-semantic representation, ONT = ontological level)

Our distinctions coincide to some degree with those proposed by OBRST (2006) in the Ontology Spectrum. Our first category corresponds to Obrst's "weak taxonomies and thesauri" and the second to logical theories ("strong ontologies"). The "weak ontologies" category in the Ontology Spectrum integrates aspects from both of these, and is used in data modeling (UML) rather than for domain representation. While Obrst mentions the class vs. instance distinction in his portrayal of strong ontologies, the distinction is not further elaborated.

Our main thesis in the above is that knowledge representation – which might more properly be referred to as the modeling of beliefs among scientists – is not a task of formal ontologies. Nor do formal ontologies describe entities properly belonging to the domain of human language. These two kinds of representational artifact represent different things, serve different purposes and use different formalisms. We postulate that a clearer understanding of these differences will facilitate the definition of more robust and useful interfaces between them, and

thereby reduce the occurrence of unintended models and thus help to create a more rational basis for semantically interoperable systems in biology and medicine.

Acknowledgements

This work was supported by the European Union projects @neurIST and DebugIT, and by the National Institutes of Health through the NIH Roadmap for Medical Research, Grant 1 U 54 HG004028 (National Center for Biomedical Ontology).

Note

1. A “workaround” exists to represent n-ary relations in OWL via reification - see <http://www.w3.org/TR/swbp-n-aryRelations>

Bibliographic references

BAADER F, LUTZ C, SUNTISRIVARAPORN B. CEL – A Polynomial-time Reasoner for Life Science Ontologies. Proceedings of the International Joint Conference on Automated Reasoning, 8, 2006, Heidelberg: Springer, 2006, pages 287-291.

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.

BAADER F, PEÑALOZA R, SUNTISRIVARAPORN B. Pinpointing in the Description Logic EL. Description Logics 2007. <http://ceur-ws.org/Vol-250/>

BEISSWANGER E, STENZHORN H, SCHULZ S, HAHN U. BIOTOP: An Upper Domain Ontology for the Life Sciences. A Description of its Current Structure, Contents, and Interfaces to OBO Ontologies. Accepted for publication in Applied Ontology, 2008.

BMIR (Stanford Center for Biomedical Informatics Research). The Protégé Ontology Editor and Knowledge Acquisition System, 2008. Available from: <http://protege.stanford.edu>. Last accessed: 30 January 2009

BODENREIDER O, SMITH B, KUMAR A, BURGUN A. Investigating Subsumption in DL-Based Terminologies: A Case Study in SNOMED-CT. First International Workshop on Formal Biomedical Knowledge Representation (KR-MED 2004), 2004, pages 12-20.

BOEKER M, STENZHORN H, KUMPF K, BIJLENGA P, SCHULZ S, HANSER S. The @neurIST ontology of intracranial aneurysms: providing terminological services for an integrated IT infrastructure. Proceedings of the 2007 Annual Symposium of the American Medical Informatics Association, Washington: AMIA, 2007, pages 39-50

BOYD R. Scientific Realism, Stanford Encyclopedia of Philosophy, 2002. Available from: <http://plato.stanford.edu/entries/scientific-realism>. Last accessed: 30 January 2009.

CEUSTERS W, SMITH B, FLANAGAN J. Ontology and Medical Terminology: Why Description Logics

Are Not Enough. Towards an Electronic Patient Record Proceedings of TEPR 2003, Boston: Medical Records Institute, 2003.

CEUSTERS W, ELKIN P, SMITH B. Negative Findings in Electronic Health Records and Biomedical Ontologies: A Realist Approach. International Journal of Medical Informatics 2007; 76: 326-333.

FELLBAUM C (Ed.). WordNet: An Electronic Lexical Database. Cambridge: MIT Press, 1998.

GENOMICS STANDARD CONSORTIUM. The GAZ Ontology. http://gensc.org/gc_wiki/index.php/GAZ_Project. Last accessed: 30 January 2009.

GRENON P. BFO in a Nutshell: A Bi-categorical Axiomatization of BFO and Comparison with DOLCE. IFOMIS Technical Report, 06, 2003.

GRUBER TR. A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition (Special issue: Current issues in knowledge modeling), v. 5, n. 2, pages 199-200, 1993.

GUARINO N (Ed.). Formal ontology in information systems. Amsterdam: IOS Press, 1998.

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, pages 221 – 234, 1999.

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.

HOFWEBER T. Logic and Ontology, Stanford Encyclopaedia of Philosophy, 2004. Available from: <http://plato.stanford.edu/entries/logic-ontology>. Last accessed: 30 January 2009.

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. Available from: <http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-216>. Last accessed: 30 January 2009.

IHTSDO (International Health Terminology Standards Development Organisation). Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT), 2008. Available from: <http://www.ihtsdo.org/snomed-ct>. Last accessed: 30 January 2009.

JANSEN L. “On Ascribing Dispositions”, in: Max Kistler, Bruno Gnessounou (Eds.), *Dispositions and Causal Powers*, Aldershot: Ashgate 2007, 161-177.

KIFER M, LAUSEN G. F-logic: a higher-order language for reasoning about objects, inheritance, and scheme. ACM SIGMOD Record. 2, 1989, pages 134-146.

KLINOV P. Pronto: A Non-monotonic Probabilistic Description Logic Reasoner. Proc. of the European Semantic

- Web Conference, 6, 2008. Heidelberg: Springer, 2008, pages 822-826
- KLYNE G, CARROLL J. Resource Description Framework (RDF): Concepts and Abstract Syntax, 2004. <http://www.w3.org/TR/rdf-concepts>. Last accessed: 30 January 2009
- KOLLER D, LEVY A, PFEFFER A. P-classic: A tractable probabilistic description logic. Proc. of AAI-1997, 390 - 397.
- KUSNIERCYK W. Nontological Engineering. Proceedings of the International Conference on Formal Ontology in Information Systems, 11, 2006. Amsterdam: IOS Press, 2006, pages 39-50
- MCCRAY AT, NELSON SJ. The representation of meaning in the UMLS. Methods of Information in Medicine, v. 34, n. 1-2, pages 193-201, 1995.
- MINSKY M. A Framework for Representing Knowledge. MIT-AI Laboratory Memo 306, June, 1974. <http://web.media.mit.edu/~minsky/papers/Frames/frames.html>
- MITA (Medical Imaging and Technology Alliance). Digital Imaging and Communication in Medicine (DICOM), 2008. Available from: <http://medical.nema.org>. Last accessed: 30 January 2009
- NCI (National Cancer Institute). NCI Enterprise Vocabulary Services (EVS), 2008. Available from: <http://www.cancer.gov/cancertopics/terminologyresources>. Last accessed: 30 January 2009.
- 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.
- NHS (World Health Organization). Dictionary of Medicines and Devices (dm+d), 2008. Available from: <http://www.dmd.nhs.uk>. Last accessed: 30 January 2009.
- NLM (United States National Library of Medicine). Medical Subject Headings (MeSH), 2008. Available from: <http://www.nlm.nih.gov/mesh>. Last accessed: 30 January 2009.
- NLMa (United States National Library of Medicine). RxNorm, 2008. Available from: <http://www.nlm.nih.gov/research/umls/rxnorm>. Last accessed: 30 January 2009.
- NLMb (United States National Library of Medicine). Unified Medical Language System (UMLS), 2008. Available from: <http://www.nlm.nih.gov/research/umls>. Last accessed: 30 January 2009.
- NORTHROP GRUMMAN. Medical Dictionary for Regulatory Activities (MedDRA), 2008. Available from: <http://www.meddrasso.com>. Last accessed: 30 January 2009.
- 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>
- OBI (Ontology of Biomedical Investigation Consortium). The Ontology of Biomedical Investigations. <http://purl.obofoundry.org/obo/obi>. Last accessed: 30 January 2009.
- PATEL-SCHNEIDER PF, HAYES P, HORROCKS I. OWL Web Ontology Language Semantics and Abstract Syntax. W3C Recommendation, 2004. Available at <http://www.w3.org/TR/owl-semantics>. Last accessed: 30 January 2009.
- QUINE O. On what there is. In: Gibson R. Quintessence - Basic Readings from the Philosophy of W. V. Quine. Cambridge: Belknap Press, Harvard University, 2004.
- 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, 9(2):139-171, 1997.
- RECTOR AL. Defaults, Context, and Knowledge: Alternatives for OWL-Indexed Knowledge Bases. Pacific Symposium on Biocomputing 2004: 226-237.
- RECTOR A. Barriers, approaches and research priorities for integrating biomedical ontologies, 2008. Available from: www.semantichealth.org/DELIVERABLES/SemanticHEALTH_D6_1.pdf. Last accessed: 30 January 2009.
- REGENSTRIEF INSTITUTE. Logical Observation Identifiers Names and Codes (LOINC), 2008. Available from: <http://loinc.org>. Last accessed: 30 January 2009.
- REITER R. A logic for default reasoning. *Artificial Intelligence*, 13, pages 81-132, 1980.
- SCHULZ S, HAHN U. Medical knowledge reengineering - converting major portions of the UMLS into a terminological knowledge base. International Journal of Medical Informatics, v. 64, n. 2-3, pages 207-221, 2001.
- SCHULZ S, JANSEN L. Molecular Interactions: On the Ambiguity of Ordinary Statements in Biomedical Literature. 2008. Forthcoming in Applied Ontology.
- 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, v.8, n. 8, pages 303-320, 1975.
- SIRIN E, PARSIA B, CUENCA GRAU B, KALYANPUR A, KATZ Y. Pellet: A Practical OWL DL Reasoner. Journal of Web Semantics, v. 5, n. 2, pages 51-53, 2007.
- SMITH B. Beyond Concepts: Ontology as Reality Representation. Proceedings of the International Conference on Formal Ontology in Information Systems, 11, 2004, pages 39-50
- SMITH B, KÖHLER J, KUMAR A. On the Application of Formal Principles to Life Science Data: A Case Study in the Gene Ontology. Proceedings of Data Integration in the Life Sciences (DILS 2004), Berlin: Springer, 2004, pages 79-94.
- SMITH B, MEJINO JL, SCHULZ S, ROSSE C. Anatomical information science. In COSIT 2005: spatial

information theory. *Foundations of Geographic Information Science*, Lecture Notes in Computer Science, Springer; 2005, pp 149-64.

SMITH B, MEJINO JR JLV, SCHULZ S, KUMAR A, ROSSE C. *Anatomical Information Science*, in COHN AG, MARK DM (eds.), *Spatial Information Theory. Proceedings of COSIT 2005*, Heidelberg: Springer, pages 149-164.

SMITH B, KUSNIERCZYK W, SCHOBER D, CEUSTERS W. *Towards a Reference Terminology for Ontology Research and Development in the Biomedical Domain. Proceedings of KR-MED - Biomedical Ontology in Action*, 2006, pages 57-66.

SMITH M, WELTY C, MCGUINNESS DL. *OWL Web Ontology Language Guide*, W3C Recommendation, 2004. Available from: <http://www.w3.org/TR/owl-guide>. Last accessed: 30 January 2009

SOUALMIA LF, GOLBREICH C, DARMONI SJ. (2004). *Representing the MeSH in OWL: Towards a Semi-Automatic Migration. Workshop on Formal Biomedical Knowledge Representation (KR-MED)*, 7, 2004, pages 81-87.

STENZHORN H, SCHULZ S, BOEKER M, SMITH B. *Adapting Clinical Ontologies in Real-World Environments. Journal of Universal Computer Science*, to appear.

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, pages 292-297.


UMC (Uppsala Centre for International Drug Monitoring). *WHO Drug Dictionary Enhanced*, 2008. Available from: <http://www.umd-products.com>. Last accessed: 30 January 2009.

UNIPROT (Universal Protein Resource Consortium). *UniProt Protein Knowledgebase*, 2008. Available from: <http://www.uniprot.org>. Last accessed: 30 January 2009.

W3C (World Wide Web Consortium). *Semantic Web Activity*, 2008. Available from: <http://www.w3.org/2001/sw>. Last accessed: 30 January 2009.

Welty C, Guarino N. *Supporting ontological analysis of taxonomic relationships*, *Data & Knowledge Engineering* 39, Elsevier, 2001

WHO (World Health Organization). *International Classification of Diseases (ICD)*, 2008. Available from: <http://www.who.int/classifications/icd>. Last accessed: 30 January 2009.

WHOCC (WHO Collaborating Centre for Drug Statistics Methodology). *Anatomical Therapeutic Chemical Classification System (ATC)*, 2008. Available from: <http://www.whooc.no/atcddd>. Last accessed: 30 January 2009. 

About the authors

Stefan Schulz

Holds a medical degree (Heidelberg University, Germany) and is senior researcher and professor at the Institute for Medical Biometry and Medical Informatics of the University Medical Center Freiburg, where he leads the Medical Informatics Research Group. His work focuses on biomedical terminologies and ontologies, biomedical knowledge representation, cross-language medical document retrieval, text and data mining in clinical document repositories, eLearning in Medicine, and health informatics in developing countries. After clinical work in surgery and internal medicine he obtained his doctoral degree in the field of tropical hygiene where he carried out a parasitological field study on in São Luís, Brazil. After obtaining a technical qualification in medical computing, he moved to the University of Freiburg, where he participated in clinical and educational software development projects and participated in several research projects in the field of information extraction, biomedical terminologies, medical language engineering, and semantic technologies. He has played a leading role in several EU-funded research projects. Stefan Schulz is author of more than hundred peer reviewed publications and has received several awards. Since 2001, he has repeatedly contributed to Brazilian health informatics research projects as a visiting researcher at the Paraná Catholic University (PUC-PR).

Holger Stenzhorn

is computational linguist (Saarland University, Germany) and research associate at the Institute for Medical Biometry and Medical Informatics of the University Medical Center Freiburg, Germany. His work focuses on the representation and management of information and data, ontologies and Semantic Web technologies, biomedical informatics, natural language processing, multimodal user interface and software design and development. In the past he participated in the development of multilingual document retrieval, information extraction, and natural language generation systems, both in industry and academia. Currently, he is involved in several ontology engineering tasks: an ontology for the research on cerebral aneurysms (EU funded @neurIST project), an ontology for clinical trials on nephroblastoma and breast cancer (EU funded ACGT project), and the BioTop top-domain ontology. Holger is member of the W3C Healthcare and Life Sciences Interest Group.