Towards a Reference Terminology for Ontology Research and Development in the Biomedical Domain

Barry Smith, Ph.D., Werner Ceusters, M.D.
Center of Excellence in Bioinformatics and Life Sciences, and National Center for Biomedical Ontology, University at Buffalo, Buffalo NY/USA
phismith@buffalo.edu, ceusters@buffalo.edu

Ontology is a burgeoning field, involving researchers from the computer science, information technology, data and software engineering, logic, philosophy, terminology and, increasingly, the biomedical domains. Many ontology-related terms with precise meanings in one of these domains have different meanings in others. Our purpose here is to initiate a path towards disambiguation of such terms that is motivated specifically by the needs of biomedical informatics. Our proposal rests on a distinction of three levels that are unfortunately too often run together in biomedical ontology research: 1. the level of reality; 2. the level of cognitive representations of this reality; 3. the level of linguistic, graphical (and digital) artifacts. We propose a reference terminology for ontology research and development that is designed to serve as common hub into which the several competing disciplinary terminologies can be mapped.

PREAMBLE

At least since the invention of the computer, scientists and engineers have been exploring ways of 'modeling' or 'representing' the entities about which machines are expected to reason. But what do 'modeling' and 'representing' mean? What is a 'conceptual model' or an 'information model'? What is 'knowledge representation'?

Two questions here arise: (1) To what do expressions such as 'concept', 'information', 'knowledge', etc. precisely refer? (2) What is it to 'model' or 'represent' such things? If information and knowledge themselves consist in representations, then what could an *information representation* or a *knowledge representation* be? There is, to say the least, some suspicion of redundancy here.

As we have argued elsewhere, the term 'concept' is marked in a peculiar conspicuous manner by problems in this regard.[1] But the problem of multiple conflicting meanings arises also in regard to other terms, such as 'object', 'instance', 'individual', 'class', 'property', 'attribute', 'feature', 'type', 'function', etc., all of which have established, but unfortunately non-uniform, meanings in a range of different disciplines.

Among philosophical ontologists, for example, the term 'instance' means an individual (for example this particular dog Fido), which is an instance of associated universals (*dog*, *mammal*, etc.). In OWL, 'instance' means 'element' or 'member' of a class. And in database theory it means either a specific item in a data table or a local implementation of some system of database software.

Standardization agencies such as ISO, CEN and W3C have been of little help in engendering cross-disciplinary uniformity in the use of such terms, since their standards are themselves directed towards specific disciplinary communities. Standardization efforts based on syntactic regimentation under the auspices of OWL or UML, too, have shown themselves unable to address these problems. For while OWL-DL has a rigorously defined semantics,[2] this does not by any means guarantee that an ontology formulated using OWL-DL is an error-free representation of its intended object-domain, and nor – until the day when the use of OWL or of some successor becomes uniform common practice – will it do anything to resolve the problems of semantic conflict adverted to in the above.

In the domain of biomedical informatics a number of attempts have been made to resolve these problems[3,4,5] in light of an increasing recognition that the wide range of ambitious terminological systems developed in this field are marked by unclarity over what, precisely, they have been designed to achieve, which has contributed to a concomitant shortfall in interoperability. Are they 'concept representations' or 'knowledge models'? And if they are either of these things, how, if at all, do they relate to the diseases or treatments or chemical interactions on the side of the patient?

OBJECTIVES AND METHODS

The purpose of this communication is to initiate a process for resolving some of these problems by drawing on the best practices in ontology research which are now beginning to take root through the efforts of organizations such as the National Center for Biomedical Ontology,[6] the Open Biomedical Ontologies Consortium,[7] the OBO Foundry,[8] and others.[9]

What is needed is a set of terms referring in unambiguous fashion to the different kinds of entities surveyed above, and which can therefore serve as common target for mappings from other discipline-and computational idiom-centric terminologies, thereby mediating efficient pairwise translations between these terminologies themselves.

What follows is a first step towards realization of this need. Our strategy is to advance precision via clear informal definitions rooted in what we assume are commonly accepted intuitions, pointing to associated formal treatments in the footnotes. In selecting terms we have sometimes chosen expressions precisely because they have not been used by others and hence do not have established (and potentially conflicting) meanings In other cases, we have adapted existing terms to our purposes by providing them with more precise definitions.

These proposals are focused primarily on the ontology-related needs of natural science and, by extension, of evidence-based medicine. We believe, however, that the terminology which results can be useful also for those engaged in other types of ontology research, since even non-scientific ontologies typically rely on frameworks which form a proper part of the framework outlined here.

We start out from a distinction of three levels of entities which have a role to play wherever ontologies are used in the field of biomedicine (and, more generally, in the natural sciences):

- Level 1: the reality of objects, processes, qualities, states (for example on the side of the patient);
- Level 2: cognitive representations of this reality on the part of clinicians, researchers, and others;
- Level 3: publicly accessible concretizations of these cognitive representations in textual, graphical and other artifacts.

The importance of Level 1 reflects the fact that even those who see themselves as building for example 'data models' in the domain of the life sciences are attempting to create artifacts which stand ultimately in some representational relation to entities in the real world. Level 2 reflects the fact that a crucial role is played in all ontology and terminology development by the cognitive representations of human subjects. Level 3 reflects the fact that cognitive representations can be shared, and serve scientific ends, only when they are made communicable in a form whereby they can also be subjected to criticism and correction

Note that the three levels overlap; thus the textual and graphical artifacts distinguished in Level 3 are themselves objects on Level 1. Our distinction of levels is thus analogous to distinctions of granularity: if we have apprehended all the liquid in a vessel, then in a sense we have thereby apprehended also all the molecules. Yet for scientific purposes molecules and liquids must be distinguished nonetheless, and the same applies, for the purposes of ontology, to the three levels we have delineated in the above.

FOUNDATIONS

Here we give precise meanings to a number of terms, which will then be used in accordance with these meanings in the remainder of the paper. Really existing ontologies and related artifacts are typically constructed to realize a mixture of different sorts of ends, and so they typically combine the features of artifacts of different basic types. Our terminological proposals here are designed to reflect these basic types. Hence the definitions we propose for terms such as 'ontology' or 'class' do not imply any claim to the effect that everything called an 'ontology' or 'class' in the literature exhibits just the characteristics referred to in the definition..

An **ENTITY** is anything which exists, including objects, processes, qualities and states on all three levels (thus also including beliefs, utterances, documents, software, observations, experiments).

A **REPRESENTATION** is for example an idea, image, record or description which refers to (is *of* or *about*) some entity or entities external to the representation. We are interested here in **COMPOSITE REPRESENTATIONS**, which means: representations built out of constituent sub-representations as their parts, in the way in which paragraphs are built out of sentences and sentences out of words. The smallest constituent sub-representations are called **REPRESENTATIONAL UNITS**; examples are: icons, names, simple word forms, or the sorts of alphanumeric identifiers we might find in patient records.

If we take the graph-theoretic concretization of the Gene Ontology as an example, then the representational units here are the nodes of the graph (taken to include the associated terms and unique IDs), which refer to corresponding entities in reality. But the composite representation refers through its graph structure also to the relations between these types, so that there is reference to entities in reality both at the structural level and at the level of single units.[10]

A **COGNITIVE REPRESENTATION** (Level 2) is a representation whose representational units are ideas and beliefs in the mind of some cognitive subject – for example a clinician engaged in applying theoretical (and practical) knowledge to the task of establishing a diagnosis.

A **REPRESENTATIONAL ARTIFACT** (Level 3) is a representation that is materialized (for example written out and published) in such a way that it can serve to make publicly accessible the cognitive representations existing in the minds of separate subjects. Examples are: a database, a diagram, a cartoon, a map, a map legend, a clinical record, or a vocabulary such as SNOMED CT.

Because each of the latter gives linguistic or graphical (or digital) form to cognitive representations which pre-exist them, some have taken this to mean that they are made up of representations which *refer to* (are *of* or *about*) these cognitive representations, or more precisely to the 'concepts' from out of which the latter are held to be composed. We shall argue below, however, that the constituent units of representational artifacts developed for scientific purposes should more properly be seen as referring to the very same entities in reality – the diseases, patients, body parts, and so forth – to which the underlying cognitive representations of clinicians and others refer. Such artifacts are in this respect no different from scientific textbooks. They are windows on reality, designed to serve as a means by which the knowledge of reality captured by cognitive agents can be made available to other agents, both human and machine.

REALITY

The clinician is concerned first and foremost with **PARTICULARS** in reality (Level 1), i.e. with individual patients, their lesions, diseases, bodily reactions. Some types of particulars, such as human beings, planets, hurricanes, receive proper names; but we can refer to particulars also by means of complex expressions involving general terms (such as *that man*, *this oophorectomy*, *this danger*, and so forth). In some cases the general terms in question ('*cell*', '*fracture*', '*dog*') reflect structures and characteristics in reality which are exemplified – the very same structures or characteristics; over and over again – in an open-ended collection of particulars in arbitrarily disconnected regions of space and time. In others ('danger', 'gift', 'surprise') we are dealing with general terms which draw together entities in reality which do not share common intrinsic structures or characteristics. In yet others ('The Beatles') we are referring to specific collections of particulars tied to specific regions of space and time.

General terms of the first sort refer to **UNIVERSALS** (in the vernacular also called 'types' or 'kinds'). A universal is something that is shared in common by all those particulars which are its **INSTANCES**. The universal itself then exists in Level 1 reality as a result of existing in its instances.

It is universals which are the entities represented in scientific texts. A good *prima facie* indication that a general term, say '*A*', refers to a universal is that '*A*' is used by scientists to make different sorts of law-like assertions about the individual instances of *A* with which they work in the lab or clinic.

Both particulars and universals stand to each other in various **RELATIONS**. Thus particulars stand to the corresponding universals in the relation of instantiation. A series of other binary relations (such as parthood, adjacency, derivation),[17] can be classified as in Table 1. (We use **bold type** for relations involving particulars; *italics* for universals and for relations between universals; and Roman for particulars.)

<universal, universal>	*nose part_of body*
<particular, particular>	Mary's nose **part_of** Mary
<particular, universal>	Mary's nose **instance_of** *nose*

Table 1 – Three Basic Sorts of Binary Relation

A **COLLECTION OF PARTICULARS** (the molecules in John's body, the pieces of equipment in a certain operating theater, the operations performed in this theater over a given period of months) is a Level 1 particular comprehending other particulars as its **MEMBERS**.[12] We note that confusion is spawned by the fact that we use the very same general terms to refer both to universals and to collections of particulars. Consider:

- *HIV* is an infectious retrovirus
- *HIV* is spreading very rapidly through Asia

A **CLASS** is a collection of all and only the particulars falling under a given general term ('cell', 'oophorectomy' 'VA Hospital'). Where the general term in question refers to a universal, then the corresponding class, called the **EXTENSION** of the universal, comprehends all, and only, those particulars which as a matter of fact instantiate the corresponding universal.

The totality of classes is wider than the totality of extensions of universals, since it includes also **DEFINED CLASSES**, designated by terms like 'employee of Swedish bank', 'daughter of Finnish spy'. Set-theory-based languages like OWL are of course ideally suited for the formal treatment of such defined classes, and the usage of such languages has encouraged the view that general terms designate defined classes (or the 'concepts' which are their close cousins). Indeed, since OWL identifies the relation of instantiation with that of membership,[13] this means that it identifies universals with their extensions, so that universals themselves become, by default, invisible.

Universals are organized into trees of greater and lesser generality, whereby the instances of more specific universals are *ipso facto* also instances of the corresponding more general universals. Classes, too, including extensions of universals, are organized into analogous trees on the basis of the subset relation.

There is an issue as concerns the drawing of the line between those terms which refer to universals (or their extensions) on the one hand and those referring to defined classes on the other. Certainly there are clear cases on either side: 'electron' or 'cell', on the one hand, and 'fall on stairs or ladders in water transport NOS, occupant of small unpowered boat injured' (Read Codes) on the other. But what of borderline cases such as 'alcoholic non-smoker with diabetes', or 'age-dependent yeast cell size increase'? We need to point out, first of all, that science is of course subject to constant revision, and a term taken to refer to a universal by one generation of scientists may be demoted to the level of non-designating term ('unicorn', 'phlogiston') by the next. But then such changes are not brought about through research at the level of terminology. Rather, they are a concomitant part of the advance of science proper, and this means that representational artifacts themselves, to the extent that they are conceived as part of the practice of science, should be subject to continual update in light of such advance.

We use the term **PORTION OF REALITY** to comprehend both universals and particulars and their simple and complex combinations. Some portions of reality – for example single organisms, planets – reflect autonomous joints of reality (that is, they would exist as separate entities even in a world denuded of cognitive subjects). Other portions of reality are products of fiat demarcations of one or other type,[14] as when we delineate a portion of reality by focusing on some specific granular level (of molecules, or molecular processes), or on some specific family of universals (for example when we view the human beings living in a given county in light of their patterns of alcohol consumption).

A **DOMAIN** is a portion of reality that forms the subject-matter of a single science or technology or mode of study; for example the domain of proteomics, of radiology, of viral infections in mouse. Representational artifacts will standardly focus on domains of reality delineated in terms of some level of granularity, so that entities smaller than a given threshold value will be excluded from the domain of the ontology because they are not salient to a specific scientific or clinical purpose.[15]

REPRESENTATIONAL ARTIFACTS

In developing *theories*, biomedical researchers seek representations of the universals existing in their respective domain of reality. They first develop cognitive representations, which they transform incrementally into representational artifacts of various forms for use by humans or machines.

In developing *diagnoses*, and in compiling such diagnoses into clinical records, clinicians seek a representation of salient particulars (diseases, disease processes, drug effects) on the side of their patients. Drawing on their theoretical understanding of the universals which these particulars instantiate, they first develop a cognitive representation of what is taking place within a given collection of particulars in reality, which they then transform into representational artifacts in the form of clinical documents, entries in databases, and so forth.

The mentioned representations are typically built up in modular fashion out of sub-representations each of which, in the best case, mirrors a corresponding salient portion of reality. The most simple representations mirror universals or particulars taken singly; more complex representations – such as therapeutic schemas, diagnostic protocols, scientific texts, pathway diagrams – mirror more complex portions of reality.

In the ideal case a representation would be such that all portions of reality salient to the purposes for which it was constructed would have exactly one corresponding unit in the representation, and every unit in the representation would correspond to exactly one salient portion of reality.[29] Unfortunately, of course, in a domain such as biomedicine, which is both highly complex and subject to rapid scientific advance, this ideal case is difficult to achieve. Researchers working on the level of universals may fall short by creating representations which either (i) fail to include general terms for universals which are salient to their domain, or (ii) include general terms which do not in fact denote any universals at all. Similarly, clinicians working on the level of particulars may fall short of the best case by creating misdiagnoses, either (i) by failing to acknowledge particulars which do exist and which are salient to the health of a given patient, or (ii) by using representational units assumed to refer to particulars where no such particulars exist.

A cognitive representation needs to be made accessible to third parties, including software, through some sort of representational artifact.

A **TAXONOMY** is a tree-form graph-theoretic representational artifact with nodes representing universals or classes and edges representing *is_a* or subset relations.

An **ONTOLOGY** is a representational artifact, comprising a taxonomy as proper part, whose representational units are intended to designate some combination of universals, defined classes and concepts, and certain relations between them, in addition to the *is_a* relation.[17]

Different types of ontology can be distinguished. A **REALISM-BASED ONTOLOGY** is built out of terms which are intended to refer exclusively to universals, a **CONCEPT-BASED ONTOLOGY** out of terms intended to refer exclusively to what are called 'concepts', and so on.

We can define a **TERMINOLOGY** as a representational artifact consisting of representational units which are domain-specific general terms of some natural language. An **INVENTORY** is a representational artifact built out of singular referring terms such as proper names or alphanumeric identifiers, which are intended to designate particulars in reality (objects, processes, qualities, and so on)**.** Electronic Health Records (EHRs) incorporate inventories in this sense, though in addition to terms denoting particulars ('patient #347', 'lung #420') they will include also more complex expressions incorporating also terms designating universals and defined classes ('the history of cancer in patient #347's family').[18]

In the best case, again, each of the representational artifacts listed above (ontologies, taxonomies, inventories) will be such that its representational units stand in a one-to-one correspondence with the salient entities in its domain. In practice, however, such artifacts can be classified on the basis of the various ways in which they fall short of this best case, in terms of properties such as correctness, degree of structural fit, degree of completeness and degree of redundancy.[15,19] By exploiting such classifications we can measure the quality improvements made in successive versions, and even use such measures as a basis for further improvement.[19]

To make a representation accessible to software, it must be published in a formal language and so converted into a **FORMALIZED REPRESENTATION.** The choice of language will depend on the complexity of what one needs to express and on the types of reasoning one needs to perform. While OWL, for example, can cope well with defined classes, it may not have sufficient expressive power to meet the needs of ontologies in the life sciences domain.[20,21]

Most inventories in the biomedical field (including most EHRs) have still exploited hardly at all the powers of formal reasoning. The paradigm of Referent Tracking represents an exception to this rule,[18] since it involves precisely the embedding of a highly structured representation of particulars in a formalized representation of the corresponding universals.

THE CONCEPT ORIENTATION

To see how the terms defined above can be used to enhance clarity in discussions of ontologies and related artifacts, we focus on the term 'concept', extending our earlier treatment[1] through consideration of the ways this term is used in the field of medical terminologies.

We believe that all ontologies, terminologies, inventories and related artifacts should consist exclusively of representational units which are intended to designate entities in Level 1 reality. What are the arguments which have been given for making an exception to this rule in the specific case of medical terminologies, arguments to the effect that the latter should include also (or exclusively) representational units referring to what are called 'concepts'?[22]

First, is what we can call the *argument from intellectual modesty*, which asserts that it is up to domain experts, and not to terminology developers, to answer for the truth of whatever theories the terminology is intended to mirror. Since domain experts themselves disagree, a terminology should embrace no claims as to what the world is like, but reflect, rather, the coagulate formed out of the concepts used by different experts.

Against this, however, it can be pointed out that communities working on common domains in the medical as in other scientific fields accept a massive and ever-growing body of consensus truths about the entities in these domains. Many of these truths are, admittedly, of a trivial sort (that mammals have lungs, that organisms are made of cells), but it is precisely with such truths that science-based ontology development must begin. Where conflicts do arise, these are highly localized, and pertain to specific mechanisms, for example of drug action or disease development, which can serve as the targets of conflicting beliefs only because researchers share a huge body of presuppositions.

We can think of no scenario under which it would make sense to postule special entities called 'concepts' to which terms subject to scientific dispute would refer. For either, for any such term, the dispute is resolved in its favor, and then a level 1 entity is available to serve as its referent; or it is established that the term is non-designating, and then it is no longer a candidate for inclusion in a terminology.

Second, is the *argument from creativity*. Designer drugs are conceived, modeled, and described long before they are successfully synthesized, and the plans of pharmaceutical companies may contain putative references to the corresponding chemical universals long before there are instances in reality. But again: such descriptions and plans can be perfectly well apprehended even within terminologies and ontologies conceived as relating exclusively to what is real. On the other hand it would be an error to include in an ontology of drugs terms referring to pharmaceutical products which do not yet (and may never) exist, solely on the basis of plans and descriptions. Rather, such terms should be included precisely at the point where the corresponding instances do indeed exist in reality, exactly in accordance with our proposals.

Third, is what we might call the *argument from unicorns*. Some of the terms needed in medical terminologies refer, it is held, to what does not exist. Some patients do, after all, believe that they are James Bond, or that they see unicorns during hallucinations. The realist approach is however perfectly well able to comprehend also phenomena such as these, even though it is restricted to the representation of what is real. For the beliefs and hallucinatory episodes in question are of course as real as are the persons who suffer (or enjoy) them.

Fourth, is the *argument from medical history*. The history of medicine is a scientific pursuit; yet it involves use of terms such as 'diabolic possession' which, according to the best current science, do not refer to universals in reality. But again: the history of medicine has as its subject-domain precisely the beliefs, both true and false, of former generations (together with the practices, institutions, etc. associated therewith). Thus a term like 'diabolic possession' should be included in the ontology of this discipline in the first place as component part of terms designating corresponding types of beliefs. In addition it may appear also as part of a term designating some fiat collection of those diseases from which the patients diagnosed as being possessed were in fact suffering. The evolution of our thinking about disease can then be understood in the same way that we deal with theory change in every other part of science, as a reordering of our beliefs about the ontological validity and salience of specific families of terms, and once again concepts themselves play no role as referents.[19,23]

Fifth, is the *argument from syndromes*. The subject-matters of biology and medicine are, it is held, replete with entities which do not exist in reality but are rather convenient abstractions. A syndrome such as congestive heart failure, for example, is nothing more than a convenient abstraction, used for the convenience of physicians to collect together many disparate and unrelated diseases which have common final manifestations. Such abstractions are, it is held, mere concepts.

According to the considerations on fiat demarcations advanced above, however, syndromes and related phenomena are indeed fully real – though their reality is that of defined (fiat) classes rather than of universals. A similar response can be given also in regard to the many human-dependent delineations used in expressions like 'obesity' or 'hypertension' or 'abnormal curvature of spine'. These terms, too, refer to entities in reality, namely to defined classes which rest on fiat thresholds established by consensus among physicians.

Sixth is the *argument from findings*. When erroneous entries are entered into the clinical record, for whatever reason, then if these entries are interpreted as being about level 1 entities then logical conflicts can arise. For Rector *et al.*, this implies that the use of a meta-language should be made compulsory for all statements in the EHR, which should be, not about entities in reality, but rather about what are called 'findings'.[24] Instead of *p* and *not p*, the record would contain entries like: *McX observed p* and *O'W observed not p*, and logical contradiction is thereby avoided. This move to the meta-level comes at a price, however. For it means that the terms in terminologies devised to serve such EHRs refer not to diseases themselves, but rather to mere 'concepts' on this meta-level. It also blurs the distinction between entities in reality and associated findings, and opens the door to the inclusion in a terminology of problematic findings-related expressions such as SNOMED's 'absent nipple', 'absent leg', etc.

In the domain of scientific research we do not embargo entirely the making of object-language assertions simply because there might be, among the totality of such assertions, some which are erroneous. Rather, we rely on the normal workings of science as a collective, empirical endeavor to weed out error over time, providing facilities to quarantine erroneous entries and to resolve logical conflicts as they are identified, and we have argued elsewhere that these same devices can be applied also in the medical context.[23]

The argument for the move to the meta-level is sometimes buttressed further by appeal to special medico-legal considerations, which are seen as requiring that the EHR be a record not of what exists but of clinicians' beliefs and actions. Yet the forensic purposes of an audit trail can equally well be served by an object-language record if we ensure that meta-data are associated with each entry identifying by whom the pertinent data were entered, at what time, and so forth.

On the other side, moreover, even the move to meta-level assertions would not in fact solve the problems of error, logical contradiction and legal liability. For the very same problems arise not only when human beings are describing, on the object-level, fractures, or pulse rates, or symptoms of coughing or swelling, but also on the meta-level when they are describing what clinicians have heard, seen, thought and done. The latter, too, are subject to error, fraud, and disagreement in interpretation.

THE SEMIOTIC TRIANGLE

Finally is what we might call the *argument from multiple perspectives*. Different patients, clinicians and biologists have their own perspectives on one and the same reality. To do justice to these differences, it is argued, we must hold that their respective representations point not to this common reality, but rather to their different 'concepts' thereof.

This argument has its roots in the work of Ogden and Richards, and specifically in their discussion of the so-called 'semiotic triangle' (Figure 1.), which is of importance not least because it embodies a view of meaning and reference that still plays a fateful role in the terminology standardization work of ISO.[25]

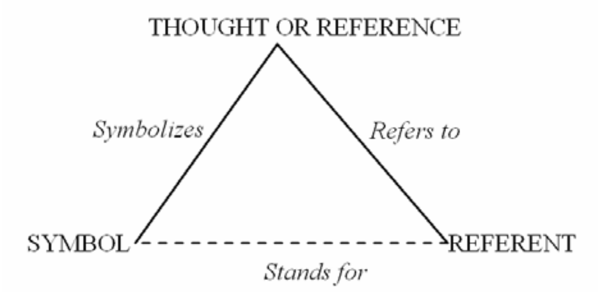

Figure 1 – Ogden and Richards' Semiotic Triangle

As the Figure makes clear, the triangle refers not to 'concepts', but rather to 'thought or reference'.[26] To understand what this means, we note that Ogden and Richards' account is rooted in a theory of psychological causality. When we experience a certain object in association with a certain sign, then memory traces are laid down in our brains in virtue of which the mere appearance of the same sign in the future will, they hold, 'evoke' a 'thought or reference' directed towards this object through the reactivation of impressions stored in memory.

The two solid edges of the triangle are intended to represent what are held to be *causal* relations of 'symbolization' (roughly: evocation), and 'reference' (roughly: perception or memory) on the part of a symbol-using subject. The dashed edge, in contrast, signifies that the relation between term and referent – the relation that is most important for the discussion of terminology – is merely 'imputed'.

The background assumption here is that the meanings words have for you or me depend on our past experiences of uses of these words in different kinds of contexts. Ambiguity is all-pervasive, and is resolved only locally and transiently in each successive context of use. From this, Ogden and Richards infer that a symbolic representation can never refer to an object *directly*, but rather only *through a* 'thought or reference' within the mind.

It is a depsychologized version of this latter thesis which forms the basis of the concept orientation in contemporary terminology research. The terms in terminologies do not refer to entities in reality, it is held; rather, they refer to 'concepts'. The latter are not (as on our view) *transparent mediators* of reference; rather they are its *targets*, and the job of the terminologist is to callibrate his list of terms in relation not to reality but to the 'realm of concepts'.[25]

The relation between terms in a terminology and the reality beyond becomes hereby obscured. Reality exists, if at all, only behind a conceptual veil – and hence there arise familiar confusions according to which for example *the concept of bacteria* would *cause* an *experimental model of disease,* or *the concept of vitamin* would be 'essential in the diet of man'.[28]

'CONCEPTS' AND 'MODELS'

How, then, should 'concept' be propertly treated in the terminology literature henceforth? There are of course sensible uses of this term, for example in the literature of psychology. In the terminology literature, however, 'concept' has been used in such a bewildering variety of confused and confusing ways that we recommend that it be avoided altogether.

It is tempting to suppose that, when considered extensionally, all of the mentioned alternative readings come down to one and the same thing, namely to an identification of 'concept' with what we have earlier called 'defined class'. If 'concept' could be used systematically in this way in terminological circles then this would, indeed, constitute progress of sorts, though the question would then arise why 'defined class' itself should not be used instead. Unfortunately, however, the proposal in question stands in conflict with the fact that 'concept' is used by its adherents to comprehend also putative referents even for terms – such as 'prevented abortion', 'surgical procedure not carried out because of patient's decision', 'prevented abortion', etc. – which do not designate defined classes because they designate *nothing at all.* Here again, we believe, a proper treatment would involve appeal to appropriate defined (fiat) classes. If we have findings currently annotated as 'prevented abortion', then this is because there are really existing plans, expectations, etc. on the part of the subject involved.

What, now is to be said of terms such as 'concept model', 'knowledge representation', 'information model', and so forth referred to in our premble above? To the extent that concept-based terminological artifacts consist in representations not of the reality on the side of the patient but rather of the entities in some putative 'realm of concepts', the term 'concept model' may be justified. This term is indeed used by SNOMED CT in its own self-descriptions, though given SNOMED's scientific goals, we believe that, on the basis of the arguments given above, it should be abandoned. Still more problematic is the term 'knowledge model' or 'knowledge representation' (GALEN). For in the absence of a reference to reality to serve as benchmark, what could motivate a distinction between *knowledge* and mere *belief*.[23] And what, in the absence of a reference to reality, could motivate adding or deleting terms in successive versions of a terminology, if every term is in any case guaranteed a reference to its own specially tailored 'concept'.

As to 'information model', here one standard uncertainty concerns the relation between an entity in reality and the body of information used to 'represent' this entity in some information system. Is it information which is being 'modeled', or the reality which this information is about? The documentation of the Reference Information Model (RIM) proposed by the HL7 organization[30] sees its principal formulas as referring to the *acts* in which entities are observed for example in a clinical context. Simultaneously, however, these formulas are held to refer also to the *documentation* of such acts for example in an information system. The apparent contradiction is resolved by the RIM on the basis of the assertion that there is in any case 'no distinction between an activity and its documentation'.[31]

CONCLUSION

On the basis of the distinction of the three levels of *reality*, *cognition* and *representational artifact* we have sought to formulate an unambiguous terminology for describing ontologies and related artifacts. The proposed terminology allows us to characterize more precisely the sorts of things which go wrong when the distinction between these levels is ignored, or when one or other level is denied, so that the approach may also help in improving such artifacts in the future.

Acknowledgements

This work was supported by the Wolfgang Paul Program of the Humboldt Foundation, the Volkswagen Foundation, and by the Roadmap Grant U54 HG004028 from the National Institutes of Health. Thanks are due also to Jim Cimino, Chris Chute, Gunnar Klein, Waclaw Kusnierczyk, Alan Rector, Stefan Schulz, Kent Spackman and anonymous referees for fruitful discussions.

References (URLs last accessed July 1, 2006)

1. Smith B. Beyond concepts, or: Ontology as reality representation, Formal Ontology in Information Systems (FOIS 2004), p. 73-84.
2. Patel-Schneider PF, Hayes P, Horrocks I. OWL Web Ontology Language: Semantics and abstract syntax. W3C Recommendation 2004. http://www.w3.org/TR/owl-semantics.
3. Spackman KA, Reynoso G. Examining SNOMED from the perspective of formal ontological principles. Workshop on Formal Biomedical Knowledge Representation (KR-MED 2004), p. 72-80.
4. Johansson I. Bioinformatics and biological reality. J Biomed Inform. 2006;39(3):274-87.
5 Klein GO, Smith B. Concept systems and ontologies. Recommendations based on discussions between realist philosophers and ISO/CEN experts concerning the standards addressing "concepts" and related terms. http://ontology.buffalo.edu/concepts/ConceptsandOntologies.pdf.
6. http://ncbo.us/.
7. http://obo.sourceforge.net/.
8. http://obofoundry.org/.
9. Rosse C, Mejino JL, Jr. A reference ontology for biomedical informatics: the Foundational Model of Anatomy. J Biomed Inform 2003;36:478-500.
10. Wittgenstein L. 1921 Tractatus Logico-Philosophicus, London: Routledge, 1961.
12. Bittner T, Donnelly M, Smith B. Individuals, universals, collections: On the foundational relations of ontology. Formal Ontology in Information Systems (FOIS 2004), p. 37-48.
13. Smith B. Against fantology, in Marek JC and Reicher ME (eds.), Experience and Analysis, Vienna, 2005, p. 153-70.
14. Smith B. Fiat objects. Topoi, 2001;20(2):131-48.
15. Bittner, T. and Smith, B. (2003). A theory of granular partitions. In: Foundations of Geographic Information Science, Duckham M, et al (eds.) London: Taylor & Francis, 117-51
18. Ceusters W, Smith B. Strategies for referent tracking in Electronic Health Records. J Biomed Inform. 2006 Jun;39(3):362-78.
19. Ceusters W, Smith B. A realism-based approach to the versioning and evolution of biomedical ontologies. Proc AMIA Symp 2006.
20. Bera P, Wand Y. Analyzing OWL using a philosophy-based ontology. Formal Ontology in Information Systems (FOIS 2004), p. 353-62.
21. Kazic T. Putting semantics into the semantic web: How well can it capture biology? Pac Symp Biocomputing 2006;11:140-51.
22 Cimino JJ. Related Articles, In defense of the desiderata. J Biomed Inform. 2006;39:299-306.
23. Smith B. From concepts to clinical reality: an essay on the benchmarking of biomedical terminologies, J Biomed Inform. 2006 Jun;39(3):288-98.
24. Rector A, Nolan W, Kay S. Foundations for an electronic medical record. Methods Inf Med, 1991;30:179-86.
25. Smith B, Ceusters W, Temmerman R. Wüsteria, Stud Health Technol Inform. 2005;116:647-652.
26. Ogden CK, Richards IA. The Meaning of Meaning. 3rd ed. New York, 1930.
28. The UMLS Semantic Network. http://semanticnetwork.nlm.nih.gov/.
30. HL7 V3 Reference Information Model: Version V 01-20. Normative Ballot 11/22/2005.
31. Smith B, Ceusters W. HL7 RIM: An incoherent standard, Stud Health Technol Inform, 2006, in press.